

Design and Analysis of Experiments

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Harvey Goldstein, Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, Joseph B. Kadane, Jozef L. Teugels*

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

Design and Analysis of Experiments

Volume 3

Special Designs and Applications

Edited by

KLAUS HINKELMANN

Virginia Polytechnic Institute and State University
Blacksburg, Virginia



A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2012 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Hinkelmann, Klaus, 1932–

Design and analysis of experiments / Klaus Hinkelmann.

p. cm. – (Wiley series in probability and statistics)

Includes index.

Contents: v. III. Special designs and applications

ISBN 978-0470-53068-9 (cloth)

1. Experimental design. I. Title.

QA279.K45 2008

519.5'7-dc22

2007017347

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

In Memoriam
Oscar Kempthorne

Contents

Preface	xvii
Contributors	xxi
1 Genetic Crosses Experiments	1
<i>Murari Singh, Sudhir Gupta, and Rajender Parsad</i>	
1.1 Introduction, 1	
1.2 Basic Objectives and Models, 2	
1.2.1 Generation Mean Analysis, 4	
1.2.2 Generation Variance Analysis, 5	
1.2.3 Covariance between Relatives, 6	
1.2.4 Mating (<i>M</i>) and Environmental (<i>E</i>) Designs, 6	
1.2.5 Fixed Effects and Random Effects Models, 7	
1.3 Diallel Mating Design of Type I, 8	
1.3.1 North Carolina Design I (<i>NCI</i>), 9	
1.3.2 North Carolina Design II (<i>NCII</i>), 10	
1.3.3 Sets of North Carolina Design II, 11	
1.3.4 North Carolina Design III (<i>NCIII</i>), 11	
1.3.5 Line \times Tester Approach, 12	
1.3.6 A Modified Line \times Tester Approach, 13	
1.4 Diallel Crosses: Type II Designs, 14	
1.4.1 Hayman Approach for Diallel Analysis, 14	
1.4.2 Griffing's Method, 21	
1.5 Partial Diallel Crosses: No Blocking or Complete Blocks, 25	
1.6 Partial Diallel Crosses in Incomplete Blocks, 32	
1.6.1 Construction of Mating–Environment Designs, 33	
1.6.2 Analysis of <i>M–E</i> Design, 36	

1.6.3	An Example of PDC in Incomplete Blocks,	41
1.6.4	Other $M-E$ Designs,	41
1.7	Optimality,	44
1.7.1	Optimal <i>CDC</i> Designs for Estimation of <i>gca</i> ,	45
1.7.2	Optimal Partial Diallel Crosses,	53
1.7.3	Estimation of Heritability,	57
1.8	Robustness,	59
1.9	Three- or Higher-Way Crosses,	61
1.9.1	Triallel or Three-Way Crosses,	61
1.9.2	Double- or Four-Way Crosses,	63
1.10	Computation,	65
	Acknowledgments,	65
	References,	66
2	Design of Gene Expression Microarray Experiments	73
	<i>Dan Nettleton</i>	
2.1	Introduction,	73
2.2	Gene Expression Microarray Technology,	74
2.2.1	Introduction,	74
2.2.2	Definition of a Microarray,	74
2.2.3	Using Microarrays to Measure Gene Expression,	74
2.2.4	Types of Gene Expression in Microarrays,	75
2.3	Preprocessing of Microarray Fluorescence Intensities,	76
2.3.1	Introduction,	76
2.3.2	Background Correction,	76
2.3.3	Normalization,	77
2.3.4	Summarization,	78
2.4	Introduction to Gene Expression Microarray Experimental Design,	80
2.5	Two-Treatment Experiments Using Two-Color Microarrays,	81
2.6	Two-Color Microarray Experiments Involving More Than Two Treatments,	86
2.7	Multifactor Two-Color Microarray Experiments,	89
2.7.1	Introduction,	89
2.7.2	Admissible Designs,	89
2.7.3	w -Optimal Designs,	92
2.7.4	e -Efficiency,	93
2.8	Phase 2 Designs for Complex Phase 1 Designs,	94
	References,	106

3 Spatial Analysis of Agricultural Field Experiments	109
<i>Joanne K. Stringer, Alison B. Smith, and Brian R. Cullis</i>	
3.1 Introduction, 109	
3.2 Methods to Account for Spatial Variation, 110	
3.2.1 Design of Experiments, 110	
3.2.2 Spatial Analysis Methods, 112	
3.3 A Spatial Linear Mixed Model, 116	
3.3.1 Estimation, Prediction and Testing, 118	
3.3.2 The Spatial Modeling Process, 119	
3.4 Analysis of Examples, 122	
3.4.1 Herbicide Tolerance Trial, 122	
3.4.2 Variety Trial, 126	
References, 132	
4 Optimal Designs for Generalized Linear Models	137
<i>John Stufken and Min Yang</i>	
4.1 Introduction, 137	
4.2 Notation and Basic Concepts, 141	
4.2.1 Binary Data, 142	
4.2.2 Count Data, 142	
4.2.3 Optimality Criteria, 143	
4.3 Tools for Finding Locally Optimal Designs, 145	
4.3.1 Traditional Approaches, 145	
4.3.2 An Analytical Approach, 146	
4.4 GLMs with Two Parameters, 149	
4.5 GLMs with Multiple Parameters, 155	
4.5.1 GLMs with Multiple Covariates, 155	
4.5.2 GLMs with Group Effects, 158	
4.6 Summary and Concluding Comments, 161	
Acknowledgments, 162	
References, 162	
5 Design and Analysis of Randomized Clinical Trials	165
<i>Janet Wittes and Zi-Fan Yu</i>	
5.1 Overview, 165	
5.2 Components of a Randomized Clinical Trial, 168	
5.2.1 Target, or Reference, Population, 168	
5.2.2 Study Population, 170	
5.2.3 Outcomes, 171	

5.2.4	Projected Timeline, 173
5.2.5	Choice of Control Group, 174
5.3	Bias, 175
5.3.1	Unbiased Entry Criteria and Recruitment, 175
5.3.2	Outcome Measures—Unbiased Assessment, 177
5.3.3	Once Randomized, Always Analyzed (Intent-to-Treat), 177
5.3.4	Masking Participants, Investigators, and Others, 178
5.3.5	Noncompliance and Study Dropout, 179
5.4	Statistical Analysis of Randomized Clinical Trials, 182
5.5	Failure Time Studies, 184
5.5.1	Basic Theory, 184
5.5.2	Actuarial and Product-Limit Survival Curves, 185
5.5.3	Exponential Survival, Hazard Rates, and Ratios and Proportional Hazard Ratios, 192
5.5.4	The Logrank Family of Tests, 192
5.5.5	The Cox Proportional Hazards Model, 194
5.5.6	Some Sample SAS Code, 195
5.5.7	Some Sample Splus Code, 201
5.5.8	Calculations of Number of Replications, or Sample Size, 203
5.5.9	Group Sequential Analysis, 206
5.6	Other Topics, 206
5.6.1	Multiplicity, 206
5.6.2	Subgroups, 208
5.6.3	Large, Simple Trials, 209
5.6.4	Equivalence and Noninferiority Trials, 209
	References, 210

6	Monitoring Randomized Clinical Trials	213
	<i>Eric S. Leifer and Nancy L. Geller</i>	
6.1	Introduction, 213	
6.2	Normally Distributed Outcomes, 215	
6.3	Brownian Motion Properties, 217	
6.4	Brief Historical Overview of Group Sequential Methods, 219	
6.5	Dichotomous Outcomes, 223	
6.6	Time-to-Event Outcomes, 225	
6.7	Unconditional Power, 227	

6.8	Conditional Power, 229	
6.9	Spending Functions, 232	
6.10	Flexibility and Properties of Spending Functions, 233	
6.11	Modifying the Trial's Sample Size Based on a Nuisance Parameter, 235	
6.11.1	Sample Size Modification for a Continuous Outcome Based on an Interim Variance Estimate, 236	
6.11.2	Sample Size Modification for a Dichotomous Outcome Based on an Interim Estimate of the Pooled Event Rate, 239	
6.12	Sample Size Modification Based on the Interim Treatment Effect, 240	
6.13	Concluding Remarks, 246	
	References, 246	
7	Adaptive Randomization in Clinical Trials	251
	<i>Lanju Zhang and William F. Rosenberger</i>	
7.1	Introduction, 251	
7.2	Adaptive Randomization Procedures, 252	
7.2.1	Restricted Randomization Procedures, 253	
7.2.2	Covariate-Adaptive Randomization, 256	
7.2.3	Response-Adaptive Randomization, 258	
7.3	Likelihood-Based Inference, 264	
7.3.1	Restricted Randomization, 266	
7.3.2	Covariate-Adaptive Randomization, 266	
7.3.3	Response-Adaptive Randomization, 266	
7.3.4	Asymptotically “Best” Procedures, 268	
7.4	Randomization-Based Inference, 269	
7.4.1	Randomization Tests, 269	
7.4.2	Monte Carlo Unconditional Tests, 271	
7.4.3	Monte Carlo Conditional Tests, 271	
7.4.4	Expanding the Reference Set, 273	
7.4.5	Stratified Tests, 273	
7.4.6	Regression Modeling, 274	
7.4.7	Covariate-Adaptive Randomization, 275	
7.4.8	Power, 275	
7.5	Conclusions and Practical Considerations, 276	
	Acknowledgment, 278	
	References, 279	

8	Search Linear Model for Identification and Discrimination	283
	<i>Subir Ghosh</i>	
8.1	Introduction, 283	
8.2	General Linear Model with Fixed Effects, 284	
8.3	Search Linear Model, 285	
8.4	Applications, 288	
8.4.1	2^m Factorial Designs, 288	
8.4.2	3^m Factorial Experiments, 291	
8.5	Effects of Noise in Performance Comparison, 293	
	References, 297	
9	Minimum Aberration and Related Criteria for Fractional Factorial Designs	299
	<i>Hegang H. Chen and Ching-Shui Cheng</i>	
9.1	Introduction, 299	
9.2	Projections of Fractional Factorial Designs, 302	
9.3	Estimation Capacity, 304	
9.4	Clear Two-Factor Interactions, 307	
9.5	Estimation Index, 310	
9.6	Estimation Index, Minimum Aberration, and Maximum Estimation Capacity, 314	
9.7	Complementary Design Theory for Minimum Aberration Designs, 315	
9.8	Nonregular Designs and Orthogonal Arrays, 317	
9.9	Generalized Minimum Aberration, 320	
9.10	Optimal Fractional Factorial Block Designs, 322	
	References, 326	
10	Designs for Choice Experiments for the Multinomial Logit Model	331
	<i>Deborah J. Street and Leonie Burgess</i>	
10.1	Introduction, 331	
10.2	Definitions, 332	
10.2.1	Standard Designs, 334	
10.3	The MNL Model, 335	
10.4	Design Comparisons, 338	
10.4.1	Optimality, 338	
10.4.2	Structural Properties, 339	

10.5	Optimal Designs for DCEs, 340
10.5.1	Generic Forced Choice DCEs, 340
10.5.2	Extensions, 351
10.5.3	Alternative-Specific Attributes, 362
10.6	Using Combinatorial Designs to Construct DCEs, 364
10.6.1	OAs and BIBDs, 364
10.6.2	A Recursive Construction using DCEs and BIBDs, 365
10.6.3	Using the OA Symbols as Ordered Pairs, 365
10.6.4	Using Hadamard Matrices to Construct DCEs, 366
10.6.5	Partial Profiles, 367
10.7	Bayesian Work, 368
10.8	Best–Worst Experiments, 368
10.8.1	Multiatribute Best–Worst Experiments, 369
10.8.2	Attribute-Level Best–Worst Experiments, 369
10.9	Miscellaneous Topics, 370
10.9.1	Other Models, 370
10.9.2	Complete Determination of Optimal Designs, 370
10.9.3	Analyzing Results from a DCE, 370
	References, 374
11	Computer Experiments
	<i>Max D. Morris</i>
	379
11.1	Introduction, 379
11.1.1	Models, 379
11.1.2	Some Notation, 381
11.1.3	Computer Experiments, 381
11.2	Sensitivity/Uncertainty Analysis, 382
11.2.1	Descriptive Methods for Local Analysis, 382
11.2.2	Methods Based on Input Sampling and Conditional Variance, 383
11.2.3	Fourier Amplitude Sensitivity Test, 385
11.3	Gaussian Stochastic Process Models, 385
11.3.1	Model Structure, 386
11.3.2	Accommodating Random Noise, 388
11.4	Inference, 389
11.4.1	Maximum Likelihood Parameter Estimation, 389
11.4.2	Numerical Issues, 390
11.4.3	Bayesian Approach, 392

11.5	Experimental Designs, 398	
11.5.1	Model-Based Designs, 398	
11.5.2	Distance-Based Designs, 399	
11.5.3	Latin Hypercube Designs, 400	
11.5.4	Uniform Designs, 402	
11.6	Multivariate Output, 403	
11.6.1	Extending the Univariate GaSP Model, 403	
11.6.2	Principal Components, 405	
11.6.3	Derivatives, 405	
11.7	Multiple Data Sources, 406	
11.7.1	Multiple Models, 406	
11.7.2	Model and Reality, 407	
11.8	Conclusion, 409	
	References, 409	
12	Designs for Large-Scale Simulation Experiments, with Applications to Defense and Homeland Security	413
	<i>Susan M. Sanchez, Thomas W. Lucas, Paul J. Sanchez, Christopher J. Nannini, and Hong Wan</i>	
12.1	Introduction, 413	
12.2	Philosophy: Evolution of Computational Experiments, 414	
12.2.1	Context, 414	
12.2.2	Why Simulation?, 415	
12.2.3	Why DOE?, 416	
12.2.4	Which DOE?, 417	
12.2.5	Implementing Large-Scale DOE, 422	
12.3	Application: U.S. Army Unmanned Aerial Vehicle (UAV) Mix Study, 422	
12.3.1	Study Overview, 423	
12.3.2	Study Goals, 424	
12.3.3	Experimental Setup, 424	
12.3.4	Results, 425	
12.3.5	Descriptive Statistics, 426	
12.3.6	Interactive Regression Modeling, 427	
12.3.7	Regression Trees, 430	
12.3.8	Other Useful Plots, 434	
12.3.9	Summary, 437	
12.4	Parting Thoughts, 437	
	References, 438	

13 Robust Parameter Designs	443
<i>Timothy J. Robinson and Christine M. Anderson-Cook</i>	
13.1 Introduction, 443	
13.2 Taguchi Signal-to-Noise Ratio Approach, 445	
13.3 Dual Model Response Surface Methodology, 448	
13.3.1 Overview, 448	
13.3.2 Designs for Dual Response Modeling , 448	
13.3.3 Analysis with Dual Response Modeling, 449	
13.4 Single Model Response Surface Methods Using Combined Arrays, 451	
13.4.1 Overview, 451	
13.4.2 Combined Array Designs, 453	
13.4.3 Analysis of Combined Array RPD Experiments, 456	
13.4.4 Analysis of Combined Arrays with Multiple Responses, 460	
13.5 Computer Generated Combined Arrays, 461	
13.6 RPD Involving Quantitative and Qualitative Factors, 465	
13.7 Conclusions, 466	
References, 467	
14 Split-Plot Response Surface Designs	471
<i>G. Geoffrey Vining</i>	
14.1 Introduction, 471	
14.2 Differences between Agricultural and Industrial Experimentation, 472	
14.2.1 Basic Differences, 472	
14.2.2 Classical Agricultural Split-Plot Design and Analysis, 474	
14.2.3 First-Order Industrial Split-Plot Design and Analysis, 477	
14.2.4 Issues for Second-Order Industrial Split-Plot Designs, 477	
14.3 OLS–GLS Equivalent Second-Order Split-Plot Designs and Analysis, 482	
14.3.1 Balanced Equivalent Designs, 482	
14.3.2 Non-VKM Balanced OLS–GLS Equivalent Designs, 485	
14.3.3 Unbalanced OLS–GLS Equivalent Designs, 486	
14.4 Exact Tests for the Coefficients, 488	
14.5 Proper Residuals for Checking Assumptions, 493	

14.6	“Optimal” Second-Order Split-Plot Designs, 496	
	References, 499	
15	Design and Analysis of Experiments for Directional Data	501
	<i>Sango B. Otieno and Christine M. Anderson-Cook</i>	
15.1	Summary, 501	
15.2	Introduction and Historical Background, 501	
15.2.1	Overview of Directional Data, 502	
15.2.2	Existing Designs for Directional Data, 508	
15.3	ANOVA for Circular Data, 509	
15.3.1	One-Way ANOVA, 510	
15.3.2	Multiway ANOVA, 518	
15.4	ANOVA for Cylindrical Data, 521	
15.5	ANOVA for Spherical Data, 524	
15.5.1	One-Way ANOVA for Spherical Data, 525	
15.5.2	One-Way ANOVA for Axial Data, 526	
15.5.3	Multiway ANOVA for Axial Data, 528	
15.6	Conclusions, 530	
	References, 531	
Author Index		533
Subject Index		545

Preface

Recently, the Department of Statistics at Virginia Tech offered a course on the history of statistics, and I was asked to present a lecture on the history of experimental design. There is no question that this history goes back many centuries, with one of the first recorded experiments being a controlled experiment on scurvy. It was carried out by the surgeon James Lind on board of HM Bark *Salisbury* in 1747. Twelve equally sick sailors were divided into pairs, and each pair was given a different treatment. Although this experiment showed already some indication of what we now consider to be the principles of experimental design, such as essentially uniform experimental units and replication of treatments, no statistical analysis was used or needed to come to the conclusion that the treatment consisting of two oranges and one lemon every day showed the best results for getting the men back to work. This may very well be the first recorded rudimentary form of what we now call a clinical trial.

The analogue to this experiment in an agricultural setting is the Broadbalk experiment set up by John Lawes in 1843 at Rothamsted, England. The experiment was set up to test the effects of various forms of inorganic fertilizer and farmyard manure on the yield of winter wheat. The experiment was not laid out according to modern principles of experimental design, but it is an elementary form of a factorial experiment, where one factor was changed at a time. Thus, even though not suited for statistical analysis, the Broadbalk experiment resulted in important numerical comparisons between certain forms of fertilizer and management treatments. Even today, it is considered to be one of the longest lasting and most successful agronomic experiments. It stimulated, of course, much of the groundbreaking work in experimental design done by R.A. Fisher, Frank Yates, and others in the 1920s and 1930s at Rothamsted Experiment Station.

From beginnings such as these, the importance of experimental design has spread throughout the worlds of scientific and industrial experimentation. It is difficult to imagine that today any scientific or industrial empirical research can be done successfully without using principles of experimental design and

statistical analysis. In addition to Fisher and Yates, many people have contributed to the advances of experimental design, among them J. Neyman, R.C. Bose, C.R. Rao, W.G. Cochran, Gertrude Cox, D.J. Finney, O. Kempthorne, J. Kiefer, G.E.P. Box, D.R. Cox, and J.N. Srivastava. Much of their work we have described in Hinkelmann and Kempthorne's *Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design, Second Edition* (2008) and *Volume 2: Advanced Experimental Design* (2005). (In the chapters of this volume, these will be referred to as HK1 and HK2, respectively). This includes the most commonly used error-control and treatment designs for comparative experiments, with the analysis based on randomization theory and the general linear model.

The principles and notions exposed in HK1 and HK2 have, over the years, been expanded for and adapted to special situations and applications by many researchers. The stimulus has come very often from scientists and practitioners working in applied fields, such as genetics, medicine, marketing, manufacturing, industrial production, agriculture, forestry, pharmacy, engineering, defense, national security, and others. These areas may require special adaptations or implementations of existing designs and/or special methodologies for analyzing data from such experiments. The reason for writing this *Volume 3: Special Designs and Applications* is to acquaint readers with these types of problems. Each of the 15 chapters gives an introduction, often with a historical background, to the topic under consideration and then discusses solutions to the particular problems, with references to the most recent results.

We begin in Chapter 1 with the discussion of designs for genetic crosses, the topic that fostered my interest in experimental design. Incomplete block designs are not only used to generate appropriate mating designs for the evaluation of the combining abilities of inbred lines for the purpose of producing well performing hybrids, but they are also used to grow the offspring from the mating designs in environmental designs. Another topic of genetic interest is discussed in Chapter 2, where the design aspects of two-phase microarray experiments are considered. Whereas the usual error-control designs are used in the first phase for the purpose of obtaining tissue samples for measurement with microarrays, the second phase of measuring the mRNA content of the tissue imposes certain new and special conditions on the construction of designs. Special aspects of data analysis are discussed in Chapter 3 in the context of agricultural field experiments. Because the growing conditions over a large experimental site can be quite variable, in addition to planning the experiment carefully, it may be useful and necessary to model spatial heterogeneity using some form of mixed linear model when analyzing the data from such field trials. Such a model is quite different from and much more complicated than the linear model used in HK1 and HK2 for analyzing data for the commonly used error-control and treatment designs. Still other models, namely generalized linear models, are considered in Chapter 4 for experiments with, for example, binary or count responses. These types of responses may occur in noncomparative experiments, where the main objective is to study

the relationship between treatment and response. This leads to questions about methods of constructing optimal designs for purposes of modeling such relations.

Entirely different considerations for designing and conducting an experiment arise in the context of clinical trials, the subject of Chapters 5, 6, and 7. The major difference from most other experiments is the fact that the experimental units are humans who have consented to be part of the trial. This imposes logistical constraints and ethical considerations within a regulatory framework. Careful monitoring of the subjects and the outcome of the trial are of paramount importance, which often leads to the use of sequential trials combined with adaptive randomization of subjects to treatments and ensuing interim analyses of the data. Thus, aspects of both design and analysis present special challenges, and they are discussed in these chapters.

Factorial experiments have long played an important role in scientific and industrial experimentation. A common feature and major concern is the fact that these experiments can become rather large and hence difficult to carry out. This has led to the introduction of fractional factorial experiments, and it has become important to find methods of constructing designs for this purpose. These methods are predicated on the assumption, and, indeed, existence of the sparsity of factor or factor combination effects. One way then to construct fractional factorial designs is to assume an approximate linear model for the data and then construct designs that will be able to estimate the effects contained in that model, as is done in Chapter 8. Another approach, as exposited in Chapter 9, is to provide various characterizations of fractional factorial designs and establish the estimating capabilities of such designs associated with certain properties. An interesting use of factorial and fractional factorial designs is described in Chapter 10 in the context of choice experiments, where people are asked to state a preference for, say, a particular type of service. The service is described by a number of attributes, each of which has a number of different alternatives. The attributes correspond to factors and the alternatives to the levels of those factors. The notion of fractional factorial designs is then used to construct and compare different designs for choice experiments, that is, designs consisting of different choice sets.

In the physical world, experiments are often limited by the number of factors that can be accommodated and by the degree of complexity involving the interaction among factors. These limitations can be overcome by empirical computer experiments through the simulation of complex systems and subsequent analysis of “data” in the form of a computer or simulation model. A general discussion and various alternatives of such an approach are described in Chapter 11. A special problem involving defense and homeland security is discussed in Chapter 12, where the study goal, the experimental setup, and regression and graphical analysis are provided in detail.

A different situation involving factorial designs arises in the context of robust parameter designs, where the factors are divided into control and noise factors. The basic problems, as discussed in Chapter 13, are concerned with

understanding the influence of noise factors and with mitigating their impact on the response through judicious choice of the control factor settings. It is shown how this can be achieved by using either the dual model or single model response surface approach for modeling the mean response, as well as the variance of the response. Another problem arising often in practice when factorial designs are used in the context of response surface designs is addressed in Chapter 14. The distinction between hard-to-change and easy-to-change factors leads to special design considerations and to split-plot type designs, referred to as equivalent estimation designs. Special emphasis is being paid to OLS–GLS equivalent second-order response surface designs.

Chapter 15 is concerned with questions of analysis for a different type of responses or observations, namely directional data. The essential characteristics of circular, cylindrical, and spherical data are described as well as the distributions for such data, for example, the von Mises distribution. Even though familiar techniques are used for analyzing the data, the nature of the data and the different distributions lead to forms of test statistics different from the usual ones for linear models.

These chapters thus span a wide range of theory and application for the design and analysis of experiments. In conjunction with HK1 and HK2, the reader should get a very good impression of how much the field has developed in its scope and sophistication since the beginnings of James Lind and John Lawes, as many of the latest results are discussed in this volume. This suggests that this volume is not intended as a self-contained textbook, but should be seen rather as an extension of HK1 and HK2 (or books of a similar nature) with particular interest for teachers, graduate students, and researchers in the field of experimental design as a supplemental text and reference book. Some of the more applications-oriented chapters also should be accessible and useful to practitioners. To help with the understanding of the material, most chapters provide a number of illustrative, and, where appropriate, numerical examples, in the latter case using available software, such as SAS, JMP, R, and specially written software programs, which will be available as indicated in the particular chapter or at the FTP (ftp://ftp.wiley.com/public/sci_tech_med/special_designs) for this book, maintained by wiley.com.

This volume is the work of many excellent researchers in the field of experimental design, and I would like to express my gratitude and thanks to them for agreeing to contribute to this volume, for providing insightful and challenging chapters, and for cooperating with me on my wishes for revisions.

*Blacksburg, Virginia
February 2011*

KLAUS HINKELMANN

Contributors

Christine M. Anderson-Cook is a Research Scientist in the Statistical Sciences Group at Los Alamos National Laboratory, Los Alamos, New Mexico. Before this, she held a position as Associate Professor of Statistics at Virginia Polytechnic Institute and State University. She is a Fellow of the American Statistical Association and a Senior Member of the American Society for Quality. In 2009, she received the National Laboratory STAR Award. Dr. Anderson-Cook is the author/coauthor of more than 80 publications in statistical and applications journals and coauthor (with R.H. Myers and D.C. Montgomery) of *Response Surface Methodology: Process and Product Optimization Using Designed Experiments, Third Edition*. She is serving as the chair of the American Society for Quality Statistics Division, as associate editor for the *Journal of Statistics Education*, and on several editorial boards and as guest editor of statistics and quality engineering journals.

Leonie Burgess is a Visiting Fellow in the School of Mathematical Sciences at the University of Technology, Sydney, Australia. She is the author/coauthor of more than 20 papers on the design of experiments and co-author (with D.J. Street) of *The Construction of Optimal Stated Choice Experiments: Theory and Practice*. Dr. Burgess is a consultant on experimental design for academic researchers and industry clients.

Hegang H. Chen is an Associate Professor in the Division of Biostatistics and Bioinformatics at the University of Maryland School of Medicine, with prior service as Assistant Professor of Statistics at Virginia Polytechnic Institute and State University. Dr. Chen is the author/coauthor of more than 60 publications in professional journals. He is serving as associate editor of the *International Journal of Statistics and Management Systems*.

Ching-Shui Cheng is a Professor of Statistics at the University of California at Berkeley. He served previously as director of the Institute of Statistical Science at the Academia Sinica. Dr. Cheng is a Fellow of the Institute of

Mathematical Statistics and of the American Statistical Association. He is the author/coauthor of more than 80 publications in statistical journals and serves as chair editor of *Statistica Sinica* and as associate editor for the *Annals of Statistics*, *Biometrika*, *Technometrics*, and *Journal of Statistical Planning and Inference*.

Brian R. Cullis is a Professor of Biometry in the Faculty of Informatics at the University of Wollongong, Australia. Dr. Cullis has more than 30 years experience in statistics with applications in the biological and agricultural sciences, with particular emphasis on experimental design, plant improvement, and genomics. His work has been widely adopted within the Australian grains industry and overseas. He is the author/coauthor of more than 150 articles in professional journals. He has served as coeditor of *Biometrics*.

Nancy L. Geller is the Director of the Office of Biostatistics Research at the National Heart, Lung, and Blood Institute in Bethesda, Maryland. She is a Fellow and the 2011 president of the American Statistical Association. Dr. Geller is the author/coauthor of more than 200 publications in professional journals and the editor of *Advances in Clinical Trial Biostatistics* (2005). She is a former president of the International Society for Clinical Biostatistics. Her editorial service includes associate editor of *Biometrics* and member of the Editorial Board of *Clinical Trials*.

Subir Ghosh is a Professor of Statistics at the University of California, Riverside, where he received the Academic Senate Distinguished Teaching Award and the Graduate Council Dissertation Advisor/Mentoring Award. Dr. Ghosh is a Fellow of the American Statistical Association, the American Association for the Advancement of Science, and an elected member of the International Statistical Institute. He is the author/coauthor of about 90 publications in professional journals and books. He served as executive editor of the *Journal of Statistical Planning and Inference* and as president of the International Indian Statistical Association.

Sudhir Gupta is a Professor of Statistics at Northern Illinois University, where he served as director of the Division of Statistics (2000–2004). He is a Fellow of the American Statistical Association. Dr. Gupta has authored/coauthored about 85 research papers and one Springer-Verlag research monograph. He is a member of the editorial boards of the *Journal of Statistical Planning and Inference* and *Communications in Statistics* and is a founding editor of the *Journal of Statistics and Applications* of the Forum for Interdisciplinary Mathematics. He has edited several issues of the *Journal of Statistical Planning and Inference*.

Eric S. Leifer is a Mathematical Statistician at the National Heart, Lung, and Blood Institute in Bethesda, Maryland. He is the author/coauthor of several

publications in statistical journals and has performed editorial services for *Biometrics*, *Biometrical Journal*, *Journal of Biopharmaceutical Statistics*, *Statistics in Medicine*, and the *Encyclopedia of Clinical Trials*.

Thomas W. Lucas is an Associate Professor and Co-Director of the Simulation Experiments and Efficient Designs (SEED) Center for Data Farming in the Operations Research Department at the Naval Postgraduate School, Monterey, California. He is the author/coauthor of about 20 publications in professional journals.

Max D. Morris is a Professor of Statistics and Professor of Industrial and Manufacturing Systems Engineering at Iowa State University. He was previously scientific staff member at Oak Ridge National Laboratory, Oak Ridge, Tennessee. Dr. Morris is a Fellow of the American Statistical Association and a recipient of the Jack Youden Prize (2001), the Jerome Sacks Award for Cross-Disciplinary Research (2002), and the Frank Wilcoxon Prize (2010). He is the author/coauthor of about 65 publications in statistical and other professional journals and the author of *Design of Experiments: An Introduction Based on Linear Models* (2010). He has served as associate editor and editor of *Technometrics* and as associate editor for the *Journal of Statistical Computation and Simulation* and the *Journal of Quality Technology*.

Christopher J. Nannini, Lieutenant Colonel, U.S. Army, is a Military Instructor in the Department of Operations Research at the Naval Postgraduate School, Monterey, California, where he is also a PhD student in the Modeling, Virtual Environment and Simulations (MOVES) Program. He has worked on a simulation tool used for the allocation of unmanned aerial vehicles (UAVs) to mission areas. He is an Army Chemical Officer and Operations Research Systems Analyst.

Dan Nettleton is a Professor of Statistics and holds the Laurence H. Baker Endowed Chair in Biological Statistics at Iowa State University. He is a Fellow of the American Statistical Association. Dr. Nettleton is the author/coauthor of about 100 publications in statistical and biological journals. He served as president of the Iowa Chapter of the American Statistical Association and serves as associate editor of the *Journal of the American Statistical Association*, *Journal of Agricultural, Biological, and Environmental Statistics*, and *Biometrics*.

Sango B. Otieno is an Associate Professor of Statistics at Grand Valley State University, Michigan, where he also serves as director of the Statistical Consulting Center. He is author/coauthor of numerous publications in professional journals.

Rajender Parsad is the Head of the Division of Design of Experiments at the Indian Agricultural Statistics Research Institute, New Delhi, India. He had a

previous appointment as a National Fellow at the International Center for Agricultural Research, India. Among his honors and awards are: recipient of the Young Scientist Award for Social Sciences from the Indian National Academy of Agricultural Sciences, recipient of the P.V. Sukhatme Gold Medal award, elected member of the International Statistical Institute, and Fellow of the National Academy of Agricultural Sciences. Dr. Parsad is the author/coauthor of more than 100 publications in professional journals and coauthor of two IASRI monographs. He is the joint secretary of the Indian Society of Agricultural Statistics and serves on the Executive Council, Forum for Interdisciplinary Mathematics. He is the coordinating editor of the *Journal of the Indian Society of Agricultural Statistics*.

Timothy J. Robinson is an Associate Professor of Statistics at the University of Wyoming. He is the author/coauthor of about 40 articles in professional journals and is coauthor (with R.H. Myers, D.C. Montgomery, and G.G. Vining) of *Generalized Linear Models with Applications in Engineering and the Sciences, Second Edition* (2010). Dr. Robinson has served as program chair for the Quality and Productivity Section of the American Statistical Association and serves as student grant chair for the Statistics Section of the American Society for Quality. His editorial service includes associate review editor for the *Journal of the American Statistical Association* and associate editor for *Quality Engineering*. He has served as statistical consultant for industry and government.

William F. Rosenberger is a Professor and Chairman of the Department of Statistics, George Mason University, Fairfax, Virginia. He is a Fellow of the American Statistical Association and of the Institute of Mathematical Statistics. Dr. Rosenberger is the author/coauthor of about 70 articles in professional journals and the coauthor (with J.L. Lachin) of *Randomization in Clinical Trials: Theory and Practice* (2002) and (with F. Hu) of *The Theory of Response-Adaptive Randomization in Clinical Trials*. He is the editor of two IMS monographs and serves as associate editor for several journals.

Paul J. Sanchez is a faculty member in the Operations Research Department at the Naval Postgraduate School, Monterey, California. His research focuses on the intersection between computer modeling and statistics. Dr. Sanchez has been an active member of the simulation community for more than 25 years and has served as referee, session chair, and proceedings editor for the Winter Simulation Conference.

Susan M. Sanchez is a Professor of the Department of Operations Research and Graduate School of Business and Public Policy, and Co-Director of the Simulation Experiments and Efficient Designs (SEED) Center for Data Farming, Naval Postgraduate School, Monterey, California. Her previous appointment was at the University of Missouri-St. Louis. Dr. Sanchez was a

NRC Senior Postdoctoral Research Fellow, received the Kelleher U.S. Army TRADOC Analysis Center-Monterey Director's Award for Research Excellence, and Outstanding Service Recognition from the INFORMS Simulation Society. She has published approximately 75 articles/chapters in professional journals/books. Her service to the profession includes: chair, Winter Simulation Conference Board of Directors; president, INFORMS College on Simulation; president, Forum on Women in OR/MS; and member, NATO Modeling and Simulation Group. She served as simulation area editor for *INFORMS Journal on Computing*, associate and deputy editor for *Naval Research Logistics*, and associate editor for *Operations Research*.

Murari Singh is a Senior Biometrician at the International Center for Agricultural Research in the Dry Areas (ICARDA) in Aleppo, Syria since 1989, with a limited term appointment in the Department of Mathematics and Statistics at Concordia University, Montreal, Canada from 2008 to 2010. From 1982 to 1989, he was statistician at the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) in Patancheru, India. Dr. Singh is a Fellow of the Royal Statistical Society, a Fellow of the Indian Society of Genetics and Plant Breeding, and an elected member of the International Statistical Institute. He is the author/coauthor of about 130 publications in statistical and subject matter journals and has served as associate editor and guest editor for the *Journal of the Indian Society of Agricultural Statistics*.

Alison B. Smith is a Senior Research Scientist at the Wagga Wagga Agricultural Research Institute, New South Wales Department of Industry and Investment, Australia. Dr. Smith has served more than 25 years as a consultant biometrician with emphasis on statistical methods for improving the efficiency of plant breeding and evaluation programs. She is the author/coauthor of 45 publications in professional journals.

Deborah J. Street is a Professor of Statistics in the School of Mathematical Sciences at the University of Technology, Sydney, Australia. She is a foundation Fellow of the Institute of Combinatorics and Its Applications. Dr. Street is the author/co-author of more 60 publications in statistical and mathematical journals and has co-authored (with A.P. Street) *Combinatorics of Experimental Design* and (with L. Burgess) *The Construction of Optimal Stated Choice Experiments: Theory and Practice*. She serves on the editorial board of *Ars Conjectandi* and has served on the editorial boards of the *Australasian Journal of Combinatorics*, *Biometrics*, *Journal of Combinatorial Designs*, and *Utilitas Mathematica*.

Joanne K. Stringer is a Senior Biometrician with BSES Limited in Queensland, Australia. Dr. Stringer has more than 25 years experience in the design, analysis and interpretation of data for the Australian Sugar Industry, with particular emphasis on new statistical approaches for the Australian sugarcane plant

improvement program. She is the author/coauthor of 30 publications in professional journals.

John Stufken is a Professor and the Head of the Department of Statistics, University of Georgia. Previous appointments have been as program director for Statistics, Division of Mathematical Sciences of the National Sciences Foundation, Washington, DC, and professor of Statistics at Iowa State University. Dr. Stufken is a Fellow of the Institute of Mathematical Statistics, a Fellow of the American Statistical Association, and an elected member of the International Statistical Institute. He is a 2011 Rothschild Distinguished Visiting Fellow, Isaac Newton Institute for Mathematical Sciences, Cambridge, United Kingdom. He is the author/coauthor of about 70 articles in statistical and other professional journals and is coauthor (with A.S. Hedayat and N.J.A. Sloane) of *Orthogonal Arrays: Theory and Applications*. His editorial service includes editor of *The American Statistician*, associate editor of *Journal of the American Statistical Association*, *Statistical Methodology*, *Journal of Statistical Theory and Practice*, *Journal of Statistical Planning and Inference*, and *Communications in Statistics*.

G. Geoffrey Vining is a Professor of Statistics at Virginia Polytechnic Institute and State University, where he also served as department head (1999–2006). He is a Fellow of the American Statistical Association and of the American Society for Quality. Dr. Vining is the author/co-author of more than 50 publications in professional journals and the recipient of the 1990 Brumbaugh Award for the paper that has made the greatest contribution to the development of industrial applications of quality control, as well as the 2011 recipient of the Shewhart Medal. He is the author of *Statistical Methods for Engineers* and coauthor (with D.C. Montgomery and E. A. Peck) of *Introduction to Linear Regression Analysis* and (with R.H. Myers and D.C. Montgomery) of *Generalized Linear Models*. He served as editor of the *Journal of Quality Technology* and as vice chair of the Publications Management Board of the American Society of Quality.

Hong Wan is an Assistant Professor in the School of Industrial Engineering at Purdue University, Indiana. She is the author/coauthor of about 15 publications in professional journals. Dr. Wan served as the 2009 Winter Simulation Conference Advanced Tutorial Track Coordinator. She is an associate editor of the *ACM Transactions on Modeling and Computer Simulation* and has served as referee for several operations research and computing journals. She is a member of the Institute of Operations Research and the Management Sciences.

Janet Wittes is the Founder and President of Statistics Collaborative, Inc., Washington, D.C. She was previously chief of the Biostatistics Research Branch, National Heart, Lung, and Blood Institute, Bethesda, Maryland.

Dr. Wittes is a Fellow of the American Statistical Association, the American Association for the Advancement of Science, and the Society of Clinical Trials. In 2006, she received the Janet L. Norwood Award for Outstanding Achievement by a Woman in the Statistical Sciences. She is the author/coauthor of approximately 200 articles in professional journals. She is the coauthor (with M.A. Proschan and K.K.G. Lan) of *Statistical Monitoring of Clinical Trials: A Unified Approach*, (with A. Turk and J. Turk) of *Ecology, Pollution, Environment*, and (with A. Turk, J. Turk, and R. Wittes) of *Environmental Science*. She is former president of The Biometric Society, ENAR, and the Society for Clinical Trials. Her editorial service includes editor-in-chief of *Controlled Clinical Trials* (1994–1998) and member of the Editorial Board of *Clinical Trials* and *Trials*.

Min Yang is an Associate Professor of Statistics at the University of Missouri. He is the recipient of a National Science Foundation CAREER Award. Dr. Yang is the author/coauthor of 15 articles in statistical journals.

Zi-Fan Yu is a Consultant, Statistics Collaborative, Inc., Washington, DC. Dr. Yu is the author/coauthor of 10 publications in professional journals.

Lanju Zhang is a Senior Principal Statistician in the Biostatistics Department of MedImmune LLC. He is the author/coauthor of more than 20 papers in statistical and other professional journals and books. He has performed editorial service for *Biometrics*, *Biometrika*, *International Journal of Biostatistics*, *Journal of Statistical Planning and Inference*, *Metrika*, *Journal of Biopharmaceutical Statistics*, and *Biometrical Journal*.

CHAPTER 1

Genetic Crosses Experiments

Murari Singh, Sudhir Gupta, and Rajender Parsad

1.1 INTRODUCTION

A major objective of biometrical genetics is to explore the nature of gene action in determining quantitative traits. This also includes determination of the number of major genetic factors or genes responsible for the traits. The history of genetic experiments can be traced back to Mendel's famous experiments on peas, the results of which he published in 1864. His work remained obscure until it was rediscovered independently by three scientists Hugo de Vries, Carl Correns, and Erich von Tschermak-Seysenegg, and published in 1900 (Monaghan and Corcos 1986, 1987); see <http://www.eucarpia.org/secretariate/honorary/tschermak.html>. Further genetic experimentation quickly followed these discoveries, and the subject of experimental genetics was thus founded.

This chapter deals with the type of genetic experiments that help assess variability in observed quantitative traits arising from genetic factors, environmental factors, and their interactions. To generate information on the variability, genetic entities, such as individual plants, animals, lines, clones, strains, and populations, are involved. Experimental design plays a twofold role in these experiments: a design to form genetic crosses and a design to evaluate the crosses in chosen environments. These two designs are called the mating design M and the environment design E , respectively. Some of the key resources in this area include standard texts and expository papers by Kempthorne (1956), Mather and Jinks (1982), Hayman (1954a, 1954b), Hinkelmann (1975),

Singh and Chaudhary (1979), Falconer and MacKay (1996), Kearsey and Pooni (1996), and Lynch and Walsh (1998). There is also a wealth of research published in scholarly journals and special issues of symposia on the topic. Obviously it is not practically feasible to cover all the important themes and methodologies of genetical experiments here. This chapter, therefore, makes a subjective selection of the topics with the aim of providing a moderate account of the concepts necessary for achieving some of the major objectives of genetical experiments, and designs and analyses thereof.

Section 1.2 discusses some more specific objectives, basic generations raised for estimation of parameters and covariances between relatives. Various types of M and E designs are discussed in later sections. Specifically, M designs for diallel experiments of type I and type II are discussed in Sections 1.3 and 1.4, respectively. These designs are generally complete crosses evaluated in experimental material without any blocking system or with complete blocks. Designs based on partial diallel crosses and their analyses are presented in Section 1.5 for complete blocks and in Section 1.6 for incomplete blocks. Sections 1.7 and 1.8 are devoted to incomplete block designs with desirable properties, such as optimality and robustness. A number of variance and covariance parameters cannot be estimated from diallel crosses in the presence of epistatic effect. For this purpose, three- and higher way crosses are required. A short review of M and E designs involving three- or higher-way crosses and their analyses are given in Section 1.9. A real data set has been used to illustrate the analyses. Some references where one could obtain SAS[®] codes for carrying out the analysis are given in Section 1.10, while codes in R language are provided on the John Wiley website (URL: ftp://ftp.wiley.com/public/sci_tech_med/special_designs).

1.2 BASIC OBJECTIVES AND MODELS

For a desired quantitative trait, the more specific objectives of genetic experiments are to separate genetic variability from environmental variability, and partition genetic variability into its components, such as additive effects, dominance effect, interactions for a given gene, and interaction between genes. Another objective is to assess the contribution of inheritance in terms of heritability in determining phenotypes. Heritability in the broad sense measures the degree to which individuals' genotypes determine their phenotypes, while heritability in the narrow sense measures the extent to which genes transmitted from parents determine phenotypes (Falconer and MacKay 1996). Estimating the number of major genes or gene-factors that control a quantitative trait is also of interest.

The variation in traits in progenies of crosses formed from diverse genetic materials called parents, from immediate and future generations, provide the basis for studying inheritance of genes. The set of six generations, known as basic generations, provide information on a single gene, that is, locus, control-

ling a desired trait. With P_1 and P_2 denoting the two parents' generations (or inbred lines exhibiting spectacular differences in a trait), the other four generations developed through crossing (or mating in animals) are F_1 : a cross between P_1 and P_2 or simply $P_1 \times P_2$, BC_1 (*backcross 1*): $F_1 \times P_1$, BC_2 (*back-cross 2*): $F_1 \times P_2$ and F_2 : $F_1 \times F_1$.

Qualitative traits are observed in a few discrete categories, called genotype classes. These genotype classes can be explained using models based on only a few genes and are expected to remain uninfluenced by the environment. An underlying hypothesis leading to genotype classes is tested by fitting genetic ratios. The distribution of the individuals in genotype classes is generally studied using remarkably segregating generations F_2 , BC_1 , and BC_2 (Mather and Jinks 1982).

Throughout, quantitative traits modeled using a linear model are discussed here; the parameters of the model being functions of genetic parameters. Thus, the statistical parameters estimated from the model provide estimates of the genetic parameters for interpretation and conclusions regarding the nature of genes. We will discuss specific cases in the sections to follow.

Biometrical genetics investigates various models for phenotypic observations obtained from multiple generations. With the advent of DNA markers, however, it has naturally become possible to develop genetic maps, identify loci responsible for a given trait and evaluate marker trait association at various levels of precision. For this area, known as genetic mapping and quantitative trait loci (*QTL*) estimation, the most commonly exploited crosses include F_2 , doubled-haploid, backcross BC_1 , and recombinant inbred lines (RIL_n). A doubled-haploid genotype is formed by doubling haploid cell (ovule or pollen) occurring naturally in gametophytic phases of higher plants either spontaneously or using the chemical colchicides to double the chromosome number. Although haploids could be produced *in vivo* following delayed pollination, irradiation of pollen, temperature shocks, colchicine treatment and distant hybridization, the most important methods currently being used are based on *in vitro* culture. These methods include *in vitro* culture of anther or pollen or ovules and chromosome elimination following interspecific hybridization. With doubled-haploid technology, the complete homozygosity is achieved in one generation while with conventional breeding in plants, it may take six or more generations (Dunwell 2010; Forster and Thomas 2010). RIL_n genotypes are formed by crossing two inbred lines and are followed by repeated selfing or sibling mating up to generation n to produce recombinant inbred lines whose genomes are mosaic of the parental genomes (Brownman 2005). The area of *QTL* detection has been systematically addressed in Lander and Botstein (1989), Jensen (1989), Knapp (1991), Stam (1993), Zeng (1994), Doerge, Weir and Zeng (1997), and Sen and Churchill (2001), among others. Detection of *QTL* from the families derived from the diallel crosses (see Section 1.3), has been discussed by Verhoeven, Jannink, and McIntyre (2006) and Maria-Joao et al. (2008). An advantage of *QTL* estimation from diallel crosses is that they also provide information on possible variations in *QTLs*.

from cross to cross, whereas *QTL* technology is generally based on the results of a single cross.

1.2.1 Generation Mean Analysis

This analysis models generation means in terms of genetic components such as additive and dominance effects, and also provides statistical tests for deviation from the model. To begin with, consider a single locus with two alleles, say *A* and *a*, influencing a trait. Let P_1 and P_2 be two homozygous inbred lines that are identical at all loci except for one gene, and let the genotypes (at that locus) of the parents P_1 and P_2 be *AA* and *aa*, respectively. Without loss of generality, let the phenotype of P_1 have a higher value of the trait under consideration in comparison to the phenotype of P_2 . Then, the expected values of the phenotypes are simply expressed as $\bar{P}_1 = m + d$, $\bar{P}_2 = m - d$ and $\bar{F}_1 = m + h$, where the mid-parent value $m = (\bar{P}_1 + \bar{P}_2)/2$, $+d$ and $-d$ are the effects of genotypes *AA* and *aa* measured as deviations from the mean m , and h is the effect of the heterozygous genotype F_1 . The parameters d and h are called the additive genetic and dominance genetic (or nonadditive genetic) components of the means, respectively. The values of h characterize the degree of dominance of allele *A* or *a* over the other allele. The parameters m , d , and h are estimated using the three genotypic means. The hypothesis of the adequacy of this additive-dominance model is tested using the F_2 generation. The F_2 generation individuals produced by selfing or intercrossing the F_1 genotypes have *AA*, *Aa* and *aa* genotypes in proportions 1/4, 1/2, and 1/4, respectively. Using the above additive-dominance effect model, the expected value of F_2 mean equals

$$\frac{1}{4}(m+d) + \frac{1}{2}(m+h) + \frac{1}{4}(m-d) = m + \frac{1}{2}h.$$

Similarly, the expected values of the means of BC_1 and BC_2 are given by

$$m + \frac{1}{2}d + \frac{1}{2}h \text{ and } m - \frac{1}{2}d + \frac{1}{2}h,$$

respectively. For several noninteracting genes, the above model can be extended in terms of the additive/dominance genetic effects as the sum of such effects over all the genes/loci. When deviation from the additive-dominance is significant, that is, when the genes interact, various interaction parameters are also included in the model. For instance, for a two-gene model, interactions such as additive \times additive, additive \times dominance, and dominance \times dominance effects are introduced in the model to take care of the deviation from the additive-dominance model.

Various tests are available for testing these components of generation means in Mather and Jinks (1982) and Kearsey and Pooni (1996). A number of test statistics, called scaling tests, have been developed to test the departure from the additive-dominance model using data from basic generations.

1.2.2 Generation Variance Analysis

No genetic variation is expected between the individuals within each of the three generations P_1 , P_2 and F_1 . Any observed variation within the three generations is then attributed to the environmental contribution, denoted by V_E say. The variation within the other three (segregating) generations, BC_1 , BC_2 and F_2 is also due to the additive and dominance components, and their interactions if present. The estimates of variations due to additive and dominance components relative to the environmental variation provide estimates of heritability. The estimates of variations obtained from the basic generations can be improved by assessing variation from individuals of higher generations. For instance, for the case of a single gene, the expected genetic variance of the F_2 generation genotypes is given by $\frac{1}{4}(m+d)^2 + \frac{1}{2}(m+h)^2 + \frac{1}{4}(m-d)^2 - (m+\frac{1}{2}h)^2 = \frac{1}{2}d^2$ (the additive genetic component of variance) + $\frac{1}{4}h^2$ (the dominance or nonadditive genetic component of variance). In case of several noninteracting and nonlinked genes, the total genetic variance in F_2 , say V_G , can be written as $V_G = \sigma_A^2 + \sigma_D^2$ where σ_A^2 and σ_D^2 , respectively, are the sums of additive genetic variances ($\frac{1}{2}d^2$), and dominance genetic variances ($\frac{1}{2}h^2$), over all the genes. These components of variation are used to measure heritability of the trait. The heritability in the broad sense is defined as the ratio of genotypic variance V_G to the phenotypic variance $V_G + V_E$,

$$h_b^2 = \frac{V_G}{V_G + V_E} = \frac{\sigma_A^2 + \sigma_D^2}{\sigma_A^2 + \sigma_D^2 + V_E},$$

while heritability in the narrow sense is defined as

$$h_n^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_D^2 + V_E},$$

the additive genetic variance as a proportion of the phenotypic variance. The various components of total genetic variance are obtained in other generations also. The F_3 generation individuals are produced by selfing F_2 individuals, for example, in self-pollinated plant species, and selfing F_3 generation individuals produces F_4 generations. One normally obtains the analysis of variance (ANOVA) using data on various families (i.e., progenies of a cross or self). The estimates of variance components and heritability are obtained using the ANOVA mean squares and their expected values. We will summarize these estimates in subsequent sections for some selected families derived from designed crosses. Estimation of various genetic components also requires expressions for variances and covariances between relatives or individuals from families or generations raised through crosses. We provide these expressions in the next section for the general case as well as for crosses involved in some specific experiments.

1.2.3 Covariance between Relatives

We shall briefly present here some basic details on obtaining genetic parameters from a mating system when the environmental design is orthogonal, for example, a completely randomized design (*CRD*), that is, no-blocks, or a randomized complete block design (*RCBD*) that allows estimation of the crosses effects without any loss of information due to blocking or environmental factors. General expressions for covariances between relatives needed for statistical modeling are given in Matzinger and Kempthorne (1956), Kempthorne (1957), Willham (1963), and Hinkelmann (1975). The relatives arising in diallel crosses are half-sibs (*HS*), that is, progenies that have only one parent in common, and full sibs (*FS*), that is, progenies that have both parents in common. The covariances between *HS* and *FS* obtained from parents generated in a population through inbreeding of degree *F* are given by

$$\begin{aligned}\text{Cov}(\text{HS}) &= \left(\frac{1+F}{4}\right)\sigma_A^2 + \left(\frac{1+F}{4}\right)^2\sigma_{AA}^2 + \dots \\ \text{Cov}(\text{FS}) &= \left(\frac{1+F}{2}\right)\sigma_A^2 + \left(\frac{1+F}{2}\right)^2\sigma_D^2 + \left(\frac{1+F}{2}\right)\sigma_{AA}^2 + \dots\end{aligned}$$

where σ_A^2 , σ_D^2 , and σ_{AA}^2 are the components of variance due to additive, dominance, and additive \times additive effects, respectively. These general expressions and their simplified versions for random mating population having $F = 0$ and for completely inbred parents having $F = 1$ are often used in practice. Willham (1963) gave the following simplified version for individuals that are *FS*, paternal half-sibs (*PHS*) or maternal half-sibs (*MHS*), with random mating population having $F = 0$, under the assumption that the trait is controlled by only additive and dominant effects,

$$\text{Cov}(\text{PHS}) = \frac{1}{4}\sigma_A^2 \quad (1.1)$$

$$\text{Cov}(\text{MHS}) = \frac{1}{4}\sigma_A^2 + \sigma_{A_m}^2 \quad (1.2)$$

$$\text{Cov}(\text{FS}) = \frac{1}{4}\sigma_A^2 + \frac{1}{4}\sigma_D^2 + \sigma_{A_m}^2, \quad (1.3)$$

where $\sigma_{A_m}^2$ is the variance component due to the additive maternal effect.

1.2.4 Mating (*M*) and Environmental (*E*) Designs

Once genetic crosses have been selected through a mating design *M*, they can be observed only under certain environmental conditions. The environmental conditions must permit a precise evaluation of the genetic parameters.

Examples of environmental designs E include *CRD*, *RCBD*, and incomplete block design (*IBD*). Thus, the phenotypic data are generated through an embedding of a mating design in an environment design. Data analysis adjusts for the variability due to any environment factors before statistical assessment of genetic parameters. Normally, a linear model is fitted to the data in terms of statistical parameters, and the resulting estimates are used to evaluate the genetic parameters (Hinkelmann 1975). Specific M and E designs and the relationship between covariance parameters of the model and genetic parameters will be discussed in subsequent sections.

1.2.5 Fixed Effects and Random Effects Models

If a set of genetic material has been purposely selected or developed with a view to evaluate them specifically, then their effects in the model are treated as fixed parameters and the statistical inference is drawn on the performance of these materials per se. The experiments of this type have been termed *comparative experiments*. On the other hand, if a set of genetic material is randomly selected from a population or a process of development and is included in the experiment, then the effects of these materials are treated as random, as these effects are a random sample from a population of effects. In this case, one infers on the population or the process through which the lines have been generated rather than the lines or the *materials* per se. Such experiments have been termed *exploratory experiments* (Hinkelmann 1975). In comparative experiments, interest lies in estimating contrasts of the effects, while exploratory experiments deal with estimation of variance components and predicted values of the random effects reflecting the nature of the population or the process.

Once the observations have been generated through a suitable embedding of a mating design and an environmental design, they are used to fit a statistical model whose parameters are estimated. For further explanation, consider the observed performance (i.e., the phenotypic value) y_{ijk} of the k th progeny of a cross between individuals i (a male) and j (a female), which comprises a genetic component g_{ij} and an environment component e_{ijk} , assuming the absence of genotype \times environment interaction. The genetic component g_{ij} can be expressed in terms of the contributions: s_i due to male i , d_j due to female j , $(sd)_{ij}$ due to male–female interaction, and ε_{ijk} , any genotype specific effect of the k th offspring.

The model for the data is given by

$$y_{ijk} = g_{ij} + e_{ijk} = \mu + s_i + d_j + (sd)_{ij} + \varepsilon_{ijk} + e_{ijk}. \quad (1.4)$$

Sprague and Tatum (1942) defined the general combining ability (*gca*) of a line as its average performance in hybrid combination with other lines. Thus, the contributions s_i and d_j measure *gca* effects of male i and female j , respectively, while $(sd)_{ij}$ measures the combining effect specific to lines i and j , called

Table 1.1 Covariances between FS, HS, and Relationships between Statistical Variances and Genetic Variances

Covariance	Statistical Variance	Genetic Variance
$\text{Cov}(PHS) = \text{Cov}(y_{ijk}, y_{ij'k'})$	σ_s^2	$\frac{1}{4}\sigma_A^2$
$\text{Cov}(MHS) = \text{Cov}(y_{ijk}, y_{ijk'})$	σ_d^2	$\frac{1}{4}\sigma_A^2 + \sigma_{A_m}^2$
$\text{Cov}(FS) = \text{Cov}(y_{ijk}, y_{ijk'})$	$\sigma_s^2 + \sigma_d^2 + \sigma_{sd}^2$	$\frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2 + \sigma_{A_m}^2$

specific combining ability (*sca*) of the pair. In the fixed effects model, one seeks *M-E* designs that estimate *gca* comparisons $s_i - s_{i'}$ and $d_j - d_{j'}$, and *sca* comparisons $(sd)_{ij} - (sd)_{ij'} - (sd)_{i'j} + (sd)_{i'j'}$ with desired precisions.

In the random effects model, $s_i, d_j, (sd)_{ij}, \epsilon_{ijk}$ and e_{ijk} are assumed to be independently and normally distributed with means zero and variances $\sigma_s^2, \sigma_d^2, \sigma_{sd}^2, \sigma_e^2$, and σ_e^2 , respectively. Here these statistical variance components are first estimated. Genetic variance parameters are then estimated using the expressions such as those given above in Equations (1.1)–(1.3). For illustration, suppose the *M* design consists of m males, each crossed with f females (the same females being used each time), there being n offsprings produced from each cross. Suppose the *E* design is a *CRD* in which the *M* design is embedded. Table 1.1 provides the details for linking the statistical variance components of the model with genetic variance parameters (Kempthorne 1957; Willham 1963); all other genetic variance components are assumed to be absent.

The estimates of genetic variances in terms of the ANOVA estimates of the statistical variances are then given by $\hat{\sigma}_A^2 = 4\hat{\sigma}_s^2, \hat{\sigma}_D^2 = 4\hat{\sigma}_{sd}^2$ and $\hat{\sigma}_{A_m}^2 = \hat{\sigma}_d^2 - \hat{\sigma}_s^2$, where $\hat{}$ indicates estimate.

When additional genetic parameters are to be estimated, the *M* design will have to be chosen appropriately so that it permits their estimation. Again, the estimation is done by considering covariances among various relatives. There are several ways of choosing the *M* design to yield information on the genetic parameters of interest, after taking into account the factors associated with *E* designs. Cockerham (1963) gave a listing of mating designs and their usefulness in estimation of the variance components. The *M* designs are mainly factorials, diallels, and higher-way crosses, while the main *E* designs include complete blocks, incomplete blocks, or simply *CRD* when no environmental factors are involved. These will be discussed in subsequent sections along with few illustrative examples.

1.3 DIALLEL MATING DESIGN OF TYPE I

Diallel mating designs were suggested by Schmidt (1919), where a set of f females is crossed with a set of m males in all possible combinations. The *gca*'s can be used to rank the lines for their potential as parents in a breeding

program (Sprague and Tatum 1942). This design is also known as design II of Comstock and Robinson (1952), type I design of Hinkelmann and Stern (1960), or as factorial mating design of Cockerham (1963). There are various subtypes of this design as discussed below. The crosses of such a mating design can be evaluated in various environmental designs.

1.3.1 North Carolina Design I (NCI)

In this type I design due to Comstock and Robinson (1952), a number (s) of males are generally selected from an F_2 generation or an advanced generation by random mating, and each of the selected males are crossed with two or more (f) different females. This gives sf full sib (FS) families. With r plants per full sib family in a CRD, the response y_{ijk} on the k th progeny from the cross of male i and female j within the male i ($i = 1, \dots, s$, $j = 1, \dots, f$, $k = 1, \dots, r$) can be modeled using:

$$y_{ijk} = \mu + s_i + d_{j(i)} + e_{ijk}, \quad (1.5)$$

where s_i , $d_{j(i)}$, and e_{ijk} are assumed to be independently and normally distributed random variables with means zero and variances σ_s^2 , $\sigma_{d(s)}^2$, and σ_e^2 , respectively.

The total (phenotypic) variation in y_{ijk} can be partitioned into variations: (1) due to males, that is, between male half-sib (HS) family groups on $(s - 1)$ degrees of freedom (df), (2) due to females within male, that is, between full sib (FS) family within a male HS family groups, on $s(f - 1)$ df, and (3) within FS families on $sf(r - 1)$ df. Let the mean sum of squares of these sources of variation from the observed data be denoted by MS_s , $MS_{d(s)}$, and MS_w , respectively. Then equating the expressions of ANOVA mean squares MS_s , $MS_{d(s)}$, and MS_w due to males, females within males, and within full sibs, we have the following estimating relations:

$$\hat{\sigma}_s^2 = (MS_s - MS_{d(s)})/(rf), \hat{\sigma}_{d(s)}^2 = (MS_{d(s)} - MS_w)/r \text{ and } \hat{\sigma}_e^2 = MS_w.$$

Using these estimates of statistical variances, the genetic variance components of the trait arising from multiple independent genes with no epistasis can be estimated by:

$$\hat{\sigma}_A^2 = 4\hat{\sigma}_s^2 \text{ and } \hat{\sigma}_D^2 = 4(\hat{\sigma}_{d(s)}^2 - \hat{\sigma}_s^2).$$

The environmental variance component σ_E^2 can be obtained from the equations for average variance within FS families:

$$MS_w = \hat{\sigma}_e^2 = \frac{1}{2}\hat{\sigma}_A^2 + \frac{3}{4}\hat{\sigma}_D^2 + \hat{\sigma}_E^2.$$

Hence

$$\hat{\sigma}_E^2 = MS_w - \left(\frac{1}{2} \hat{\sigma}_A^2 + \frac{3}{4} \hat{\sigma}_D^2 \right).$$

The above estimates of components of variation $\hat{\sigma}_A^2$, $\hat{\sigma}_D^2$, and $\hat{\sigma}_E^2$ are also used to estimate heritability in the broad and narrow senses. The total phenotypic variance σ_P^2 of the progeny response y_{ijk} can be expressed either in terms of statistical variance components or in terms of genetic and environmental variance components as

$$\sigma_P^2 = \sigma_s^2 + \sigma_{d(s)}^2 + \sigma_e^2 = \sigma_A^2 + \sigma_D^2 + \sigma_E^2.$$

The above computations can be done similarly for the *RCBD*, where we have one more source of variation due to blocks, and the variation within *FS* families will have $(r - 1)(sf - 1)$ df.

1.3.2 North Carolina Design II (*NCII*)

In this design, a set of males (s) and a set of females (f) are selected, say from an F_2 population. Each of the males is crossed with every female in the set, and thus sf *FS* families are generated, say each with r progenies. Note that f *FS* families form s male half-sib (*HS*) groups and s *FS* families form f female half-sib (*HS*) groups. Here, the total phenotypic variation can be partitioned into variation (1) between *FS* families—males differences (i.e., between male *HS* family groups) on $(s - 1)$ df, females differences (i.e., between female *HS* family groups) on $(f - 1)$ df, and male \times female interaction on $(s - 1)(f - 1)$ df—and (2) within *FS* families on $(r - 1)sf$ df. The response on the individual k of the *FS* family from the cross of male i and female j can be modeled as:

$$y_{ijk} = \mu + s_i + d_j + (sd)_{ij} + e_{ijk}, \quad (1.6)$$

where the s_i , d_j , $(sd)_{ij}$ and e_{ijk} are independently and normally distributed with means zero and variances σ_s^2 , σ_d^2 , σ_{sd}^2 , and σ_e^2 , respectively. Let the mean sum of squares of these sources of variation from the observed data be denoted by MS_s , MS_d , MS_{sd} , and MS_w , respectively. Then we have the ANOVA shown in Table 1.2.

Equating the expressions of these means squares in terms of the variance components, we have

$$\begin{aligned} \hat{\sigma}_s^2 &= (MS_s - MS_{sd})/(rf), \quad \hat{\sigma}_d^2 = (MS_d - MS_{sd})/(rs) \\ \hat{\sigma}_{ds}^2 &= (MS_{sd} - MS_w)/r, \text{ and } \hat{\sigma}_e^2 = MS_w. \end{aligned}$$

Since σ_s^2 and σ_d^2 both equal $\frac{1}{4}\sigma_A^2$, one can combine the sums of squares due to males and females to get a pooled mean square

Table 1.2 ANOVA of North Carolina Design II

Source	df	Mean Square	E (Mean Square)	Genetic Variance Components
Males	$s - 1$	MS_s	$\sigma_e^2 + r\sigma_{sd}^2 + rf\sigma_s^2$	$\sigma_s^2 = \text{Cov}(PHS)$
Females	$f - 1$	MS_d	$\sigma_e^2 + r\sigma_{sd}^2 + rs\sigma_d^2$	$\sigma_d^2 = \text{Cov}(MHS)$
Male \times female	$(s - 1)(f - 1)$	MS_{sd}	$\sigma_e^2 + r\sigma_{sd}^2$	$\sigma_{sd}^2 = \text{Cov}(FS) - \text{Cov}(MHS)$
Error	$(r - 1)gsf$	MS_w		

$$MS_{\text{pooled}} = \frac{(s-1)MS_s + (d-1)MS_d}{s+d-2},$$

and a more precise estimate can be obtained from

$$\hat{\sigma}_s^2 = \hat{\sigma}_d^2 = \frac{s+d-2}{r(d(s-1)+s(d-1))} (MS_{\text{pooled}} - MS_{sd}).$$

Using these estimates of statistical variances, the estimates of genetic variance components and heritability of the trait arising from multiple independent genes with no epistasis are given by

$$\hat{\sigma}_A^2 = 4\hat{\sigma}_s^2 \text{ and } \hat{\sigma}_D^2 = 4\hat{\sigma}_{sd}^2.$$

1.3.3 Sets of North Carolina Design II

In this case, multiple sets, say g sets of *NCII* designs with the same number of males and females, are crossed and evaluated in a *CRD* or *RCBD* environmental design. Here the genetic sources of variation of the total variability are, as for *NCII* designs, due to males (i.e., between male *HS* family groups) on $g(s-1)$ df, females (i.e., between female *HS* family groups) on $g(f-1)$ df, and male \times female interaction on $g(s-1)(f-1)$ df—and within *FS* families on $(r-1)gsf$ df and those due to sets on $(g-1)$ df in case of *CRD*.

1.3.4 North Carolina Design III (*NCIII*)

Two inbred lines, P_1 and P_2 say, are crossed to produce an F_2 population. A number, n , of randomly selected F_2 individuals (used as males) are crossed with each of the two inbred lines as testers (taken as females). Thus, in all, there are $2n$ *FS* families, which can be evaluated using randomly selected r progenies in either replicated trials with blocks or *CRD*. The sources of variation for the ANOVA are: testers/female on 1 df, F_2 male on $(n-1)$ df, and tester \times F_2 male on $(n-1)$ df, and within *FS* families on $2n(r-1)$ df in *CRD*. Alternatively, the $2n$ *FS* families can be grouped as pairs of the resulting *FS* families when crossed with P_1 and with P_2 . Using these pairs, one can compute

mean squares due to F_2 males based on (1) their sums, and (2) their differences, and adjusting the sums of squares with the multiplier $r/2$. The mean squares based on sums will show variation in only additive effects, while those based on differences will show variation in dominance effects.

1.3.5 Line \times Tester Approach

Kempthorne (1957) proposed a line \times tester experiment, with similar mating pattern as the design II of Comstock and Robinson (1952), but with homozygous parents. This mating design provides for the estimation of additive and dominance genetic variance for both populations, that is, lines as well as testers. A number of second-order statistics, such as variance of the array means and mean variance of the arrays, similar to the statistics for diallel analysis as discussed by Hayman (1954b) presented in the following section, have been proposed for line \times tester ($L \times T$) by Patel, Christie, and Kannenberg (1984). In diallel experiments presented in sections to follow, the same lines are used as male and female parents; and a parent itself may also be included as a self-cross, but this situation does not arise in a $L \times T$ experiment. We shall discuss the above two methods in the following.

Let there be m lines and n testers and their $m \times n = v$ crosses be grown in a $RCBD$ with r replications. The response or the measurement y_{ijk} of a trait on the cross $i \times j$ grown in the k th block is modeled as:

$$y_{ijk} = \mu + l_i + t_j + (lt)_{ij} + \beta_k + e_{ijk},$$

where μ is general mean, l_i and t_j are the *gca* effects of the i th line and the j th tester, respectively, $(lt)_{ij}$ is the *sca* effect of the cross $i \times j$, β_k is the effect of the k th block (replication), and the e_{ijk} are independently and normally distributed random errors with zero means and constant variance $\sigma_e^2, i = 1, \dots, m, j = 1, \dots, n, k = 1, \dots, r$. Also, it is assumed that $l_i \sim N(0, \sigma_l^2)$, $t_j \sim N(0, \sigma_t^2)$ and $(lt)_{ij} \sim N(0, \sigma_{lt}^2)$. Further, these random variables are assumed independent of each other. The ANOVA on y_{ijk} is shown in Table 1.3.

The statistical variance components σ_l^2 , σ_t^2 , σ_{lt}^2 , and σ_e^2 can be easily estimated in terms of the mean squares MS_l , MS_t , MS_{lt} and MS_e from the ANOVA in Table 1.3. The genetic variance components are evaluated using the following expressions of the covariances between relatives:

Table 1.3 ANOVA for Line \times Tester Experiment in $RCBD$

Source	df	Mean Squares	E (Mean Squares)
Replications	$r - 1$		
Line	$m - 1$	MS_l	$\sigma_e^2 + r\sigma_{lt}^2 + mn\sigma_l^2$
Tester	$n - 1$	MS_t	$\sigma_e^2 + r\sigma_{lt}^2 + mn\sigma_t^2$
Line \times tester	$(m - 1)(n - 1)$	MS_{lt}	$\sigma_e^2 + r\sigma_{lt}^2$
Error	$(r - 1)(mn - 1)$	MS_e	

$\text{COV} (\text{HS for the line groups}) = \sigma_l^2$, the variance of *gcas* of the lines

$\text{COV} (\text{HS for the tester groups}) = \sigma_t^2$, the variance of *gcas* of the testers

$\text{COV} (\text{FS}) = \sigma_l^2 + \sigma_t^2 + \sigma_{lt}^2$, the variance of *sca*'s of the lines \times testers.

The variances σ_l^2 , σ_t^2 , and σ_{lt}^2 can be estimated by $(MS_l - MS_u)/(rn)$, $(MS_t - MS_u)/(rm)$ and $(MS_u - MS_e)/r$, respectively. A common estimate of the *gca* variance, $\hat{\sigma}_g^2$ (for the lines and the testers) is obtained by pooling the estimates of σ_l^2 and σ_t^2 as

$$\hat{\sigma}_g^2 = \frac{(m-1)MS_l + (n-1)MS_t - (m+n-2)MS_u}{r(2mn-n-m)}.$$

Further additive and dominance variances can be computed as shown for *NCII* design.

1.3.6 A Modified Line \times Tester Approach

Arunachalam (1974) presented a method of analysis of $L \times T$ experiments where the parents are also included, thus resulting in a modified line \times tester (*MLT*) design. A number of environmental designs may be chosen for growing the $m + n$ parents and the mn $L \times T$ crosses. Consider the situation where an *RCBD* has been used for accommodating the parents together as one set and crosses in another set. Here, in each block, the randomization is first done to allot a set of contiguous plots to the crosses. This is followed by allocating the remaining portion of the block to the set of parents. Further, the crosses are allotted randomly to the set of contiguous plots in the block, and the set of parents are also allocated randomly in a similar fashion. Here, the ANOVA would account for sources of variation due to parents on $m + n - 1$ df (which can be further partitioned into line parents, tester parents, and line parents versus tester parents), crosses on $mn - 1$ df, parents versus crosses on 1 df, blocks and experimental errors. The crosses sum of squares will be partitioned as per the analysis of line \times tester approach of Kempthorne (1957) as shown in Table 1.3, while the error df will now be comparatively larger $(r - 1)(m + n + mn - 1)$. This (pooled) error sum of squares can also be partitioned into errors sum of squares based on either parents information only or on crosses information only. Here also the estimation of genetic variances can proceed as shown above in Section 1.3.5.

The total number of crosses mf becomes too large even for moderately large m or f . This fact has led to the introduction of incomplete or partial diallel mating designs of type I (Hinkelmann 1966), where a sample of crosses is obtained using the incidence matrix of a balanced incomplete block (*BIB*) design in m varieties and f blocks. Here, a cross $(i \times j)$ is taken in the mating design if the treatment i is in block j of the *BIB* design.

1.4 DIALLEL CROSSES: TYPE II DESIGNS

In type II mating designs, the same set of inbred lines serve as male parents and female parents, whereas the sets of males and females are different lines in type I designs. Such crosses are called *diallel crosses*. Consider p inbred lines crossed with each other in all possible p^2 combinations. If all the crosses are sampled, the mating design is called *complete diallel crosses (CDC)*, often even after excluding the selfs. Specific situations may require designing experiments with only a subset of these crosses called *partial diallel crosses (PDC)*. Further, a complete or a partial diallel can be evaluated in a *CRD*, *RCBD*, or an *IBD*. These various cases of embedding a mating and an environmental designs for diallels will be discussed in Section 1.5.

There are primarily two approaches of analysis for type II designs.

1.4.1 Hayman Approach for Diallel Analysis

We discuss here F_1 diallel experiments, where homozygous parents are crossed to produce selfs, F_1 s and their reciprocals, and attempt to summarize the approach due to Hayman (1954a, 1954b). An analysis of variance of diallel data, called Hayman's ANOVA for diallel, is used to test additive and dominance effects from the data on the progenies of diallel crosses. A number of statistics for various components of variation are used to measure additive and dominance variation, detect nonallelic genic interaction, and to describe relative dominance of parental lines. A summary of the analysis will be illustrated through a data set on days to flowering in chickpea.

1.4.1.1 Hayman's ANOVA

Consider a set of p homozygous (inbred) lines, arranged in a $p \times p$ array with the rows representing the males and the columns representing the females, crossed in all possible combinations giving p^2 *FS* families. Let these p^2 families be grown/evaluated in r replications under *CRD* or *RCBD*. Let (i,j) denote the cross between the male in row i and the female in column j , and let η_{ij} be the mean response on a progeny of this cross and presented in a $p \times p$ diallel table, $i,j = 1, \dots, p$. Let the observed value of the response corresponding to a progeny in the cell (i,j) of the diallel table be denoted by y_{ij} . Since a diallel table can be formed from the responses of each individual complete block, as well as from the total or mean over all the blocks, y_{ij} will be used here as an estimate of η_{ij} from a given block or all the blocks. It is assumed that the trait is controlled by a set of genes at k loci. The genes at different loci are distributed independently in the parents and have additive effects for crosses and for reciprocals.

The model for Hayman's ANOVA (Hayman 1954a) is given by

$$\begin{aligned} y_{ij} &= \eta_{ij} + e_{ij} \\ \eta_{ij} &= \mu + \tau_i + \tau_j + \tau_{ij} + \rho_i - \rho_j + \rho_{ij}, \end{aligned}$$

where μ is the overall mean, τ_i is the i th parental effect measured as deviation from the overall mean, τ_{ij} is the remainder from the (i,j) th reciprocal sum ($\eta_{ij} + \eta_{ji}$), $2\rho_i$ is the difference between the effects of the i th parental line used as male parent and as female parent, and $2\rho_{ij}$ is the remainder from the reciprocal difference $\eta_{ij} - \eta_{ji}$. Following the notations of Hayman (1954a,b), the sum of squares (SS) corresponding to the parameters τ_i , τ_{ij} , ρ_i , and ρ_{ij} are denoted by a , b , c , and d , respectively. These SS are given below, where dot stands for summation over all the values from 1 to p for the omitted subscripts, and the summation is over all the values of i , or i and j , respectively:

$$\begin{aligned} a &= \sum (y_{..} + y_{..})^2 / (2p) - 2y_{..}^2 / p^2 \\ b &= \sum (y_{ij} + y_{ji})^2 / 4 - \sum (y_{..} + y_{..})^2 / (2p) + y_{..}^2 / p^2 \\ c &= \sum (y_{..} - y_{..})^2 / (2p) \\ d &= \sum (y_{ij} - y_{ji})^2 / 4 - \sum (y_{..} - y_{..})^2 / (2p) \end{aligned}$$

The a SS measures variation between the mean effects of each parental line on $n - 1$ df and is obtained by fitting τ_i s. This is used to test significance of additive genetic variation or *gca* variation. The b SS measures remaining variation in reciprocal sums, after accounting for a , on $\frac{1}{2}n(n-1)$ df and is obtained by fitting τ_{ij} and subtracting a . Thus, it can be used to test dominance or non-additive genetic variation or *sca* variation. The c SS measures variation between the mean maternal effects of each parental line on $n - 1$ df and is obtained by estimating ρ_i s. The d SS measures remaining variation in reciprocal differences after accounting for c on $\frac{1}{2}(n-1)(n-2)$ df, and thus is variation due to specific reciprocal crosses. It is obtained by fitting ρ_{ij} and subtracting c .

Since η_{ii} (or simply the η_i) measures the pure additive genetic effect, it can be further partitioned using the following representation:

$$\begin{aligned} \eta_{ij} &= \mu + \tau_i + \tau_j + \lambda + \lambda_i + \lambda_j + \lambda_{ij} + \rho_i - \rho_j + \rho_{ij} \quad (i \neq j) \\ \eta_i &= \mu + 2\tau_i - (p-1)\lambda - (p-2)\lambda_i. \end{aligned}$$

where λ is the mean dominance deviation, λ_i is the dominance deviation due to the i th parent over λ , and λ_{ij} is the remainder from the (i,j) th reciprocal sum. Their respective SS are as follows.

$$\begin{aligned} b_1 &= (y_{..} - py_{..})^2 / (p^2(p-1)) \\ b_2 &= \sum (y_{..} + y_{..} - py_i)^2 / (p(p-2)) - (2y_{..} - py_{..})^2 / (p^2(p-2)) \\ b_3 &= \sum (y_{ij} + y_{ji})^2 / 4 - \sum y_i^2 - \sum (y_{..} + y_{..} - 2y_i)^2 / (2(p-2)) \\ &\quad + (y_{..} - y_{..})^2 / ((p-1)(p-2)) \end{aligned}$$

where y_{ii} or simply y_i is the observed response in the (i,i) th cell of the diallel table. The b_1 SS on 1 df, obtained by fitting λ , measures the mean dominance deviation, that is, mean deviation of progenies from their parents. This signifies the dominance of the genes predominantly in one direction called directional dominance. The b_2 SS on $(p - 1)$ df, obtained by fitting λ_i , measures the mean dominance deviations of the parents in the crosses from the selfs, and b_3 SS on $p(p - 3)/2$ df, obtained by fitting λ_{ij} , measures the remaining dominance deviation unique to each progeny, equivalent to *sca* of Griffing (1956); see Kearsey and Pooni (1996).

Testing statistical significance of the sources contributing to the above sums of squares requires appropriate error mean squares. One approach is to generate a valid error mean square using interaction with the blocks. The above sums of squares for diallel table data can be computed for each replicate or block individually, and also for the diallel table data obtained using the sum over replicates for the corresponding cell positions. The valid error sum of squares for a given component, say a , is computed as $\text{block} \times a \text{ SS} = \text{sum over all the blocks of } a \text{ SS from block } \ell - a \text{ SS}$ computed from the diallel table of the sum over the blocks, on degrees of freedom equal to $(r - 1) \times \text{df of } a$. These error variances for individual components can be tested, and if found homogeneous, their pooled estimate can be used for testing various components a .

Example 1.1 Days to Flowering from 10×10 CDC in Chickpea. We consider a data set on the number of days to first flower from planting date from a complete diallel cross including reciprocals in $p = 10$ lines conducted using three randomized complete blocks at Tel Hadya, Aleppo, Syria during 1992/93. The data and the R-programs are given on the John Wiley website (URL: ftp://ftp.wiley.com/public/sci_tech_med/special_designs). For this data set, the Hayman ANOVA is shown in Table 1.4.

In Table 1.4, the $\text{block} \times a$ interaction is a valid error to compute the variance ratio (vr) and the p -value using an F distribution corresponding to the a SS, similarly for the other components. The a SS indicates significant variation due to additive genetic variation ($p < 0.001$). The variation in reciprocal sums

Table 1.4 Hayman ANOVA for the 10×10 Diallel Data on Days to Flowering

Item	df	SS	MS	Item	df	SS	MS	vr	p
a	9	14814.8	1646.1	$\text{Block} \times a$	18	448.2	24.9	66.11	5.99×10^{-12}
b	45	2217.0	49.3	$\text{Block} \times b$	90	1413.9	15.7	3.14	2.01×10^{-06}
b_1	1	558.9	558.9	$\text{Block} \times b_1$	2	16.9	8.5	66.11	1.48×10^{-02}
b_2	9	163.0	18.1	$\text{Block} \times b_2$	18	264.7	14.7	1.23	3.36×10^{-01}
b_3	35	1495.1	42.7	$\text{Block} \times b_3$	70	1132.3	16.2	2.64	2.77×10^{-04}
c	9	3971.5	441.3	$\text{Block} \times c$	18	434.0	24.1	18.30	2.56×10^{-07}
d	36	1172.0	32.6	$\text{Block} \times d$	72	1419.2	19.7	1.65	3.57×10^{-02}
Total	99	22175.2	224.0	$\text{Block} \times \text{total}$	198	3715.3	18.8	11.94	5.20×10^{-48}

remaining after accounting for the additive genetic variation, b , is significant and is largely due to specific combining ability. The mean deviation of the progenies from their parents, b_1 , is significant ($p = 0.0148$), and b_2 , the mean dominance deviation in the crosses from selfs, is not significant ($p = 0.336$). Finally, c indicates significant maternal differences due to the lines ($p < 0.001$).

1.4.1.2 Genetic Variance Components and $W_r - V_r$ Graph

Further to his ANOVA table, Hayman (1954b) used a “genetic algebra” on the diallel crosses data and presented a number of statistics to estimate and detect additive and dominance deviations. The expected values of various SS’s in his ANOVA table are also expressed in terms of these statistics. It is worthwhile to briefly describe the basics of the components of means and variances of a single gene effect and for several non-interacting genes in a framework of allelic frequencies using notations and the results from Hayman (1954a, 1954b) and Mather and Jinks (1982). Hayman (1954b) developed these statistics under a number of hypotheses on the genetic system of the study: *diploid segregation, absence of reciprocal differences, and independent action of non-allelic genes*. This is done under the assumptions of homozygous parents, independent distribution of genes in parents, and absence of multiple allelism.

Consider a single gene locus with alleles, A and a , and the set of inbred lines considered that show differences at this locus. Let a proportion of lines, say u , be AA and the remaining proportion $v = 1 - u$ be aa . Let the (additive) effects arising from inbred parents AA and aa be given by d and $-d$, respectively. We have the array in Table 1.5 of the F_1 progenies and their values from the cross of males and females of two inbred lines. The distribution of additive and dominance components of means arising from these inbred crosses is also given in Table 1.5 and can be used to assess the parents. The measurements are taken from the mid-parent value, that is, m of Section 1.2.1 has been set at zero, and h is called the dominance effect.

The total variance V among the progeny family means arises from the differences among maternal parents, paternal parents, and their interaction,

$$V = u^2 d^2 + 2uvh^2 + v^2 (-d)^2 - ((u-v)d + 2uvh)^2 = 2uv(d - (u-v)h)^2 + 4u^2v^2h^2.$$

Table 1.5 Distribution of Effects from a Single Gene in Inbred Crosses

	Maternal Parents		Paternal Array Means
Paternal parents	AA	aa	
	$(u,d)^a$	$(v,-d)$	
	AA	Aa	
	(u,d)	d	$ud + vh$
	aa	Aa	aa
	$(v,-d)$	h	$-d$
Maternal array means		$ud + vh$	$uh - vd$
			$(u - v)d + 2uvh$

^a (Frequency, effect).

The contribution of paternal differences can be measured using variance between the row margins given by

$$V_r = u(uv + vh)^2 + v(uh - vd)^2 - ((u - v)d + 2uvh)^2 = uv(d - (u - v)h)^2.$$

Noting the equality of variances due to paternal and maternal differences, the remaining variation $V_I = V - V_r - V_r = 4u^2v^2h^2$, therefore, measures the contribution due to interaction between paternal and maternal parents.

For the case of multiple genes that are independent in their effects (no epistasis) and uncorrelated in their distribution within population, the total heritable variation V will be the sum of the terms of above forms, one for each gene-pair. Thus, using Σ to denote the summation over the genes, we can write

$$V = \Sigma 2uv(d - (u - v)h)^2 + \Sigma 4u^2v^2h^2 = \frac{1}{2}D_R + \frac{1}{4}H_R,$$

where $D_R = \Sigma 4uv(d - (u - v)h)^2$ and $H_R = \Sigma 16u^2v^2h^2$ are the random mating forms of the additive variance and the dominance variance components D and H , respectively. When gene frequencies are equal, that is, $u = v = \frac{1}{2}$, D_R will not have dominance components, that is, it contains the effects of various genes in homozygous states only. Denoting only the additive variance by $D = \Sigma 4uvd^2$, we can rewrite $D_R = D + H_1 - H_2 - F$, where $H_1 = \Sigma 4uvh^2$ and $H_2 = \Sigma 16u^2v^2h^2$ represent dominance variations, and $F = \Sigma 8uv(u - v)dh$ gives the covariance between additive and dominance effects. Hence, D_R does not necessarily measure additive variation. The parameters, such as (1) the ratio $(H_1/D)^{1/2}$, measure degree of dominance, and (2) $H_2/(4H_1)$, the average value of uv over all loci. The component of variation due to paternal or equivalently due to maternal effects is

$$V_r = \frac{1}{4}D_R,$$

and of the interaction

$$V_I = \frac{1}{2}D_R + \frac{1}{4}H_R - 2V_r = \frac{1}{4}H_R.$$

The variation among parental lines in the above table(V_P) when summed over all the loci can be shown to equal $V_P = \Sigma ud^2 + v(-d)^2 - (ud - vd)^2 = \Sigma 4uvd^2 = D$.

The covariance W_r between parents and progeny family means, that is, a parent and its array means in Table 1.5, is

$$\begin{aligned} W_r &= \Sigma \{ud(ud + vh) + v(-d)(uh - vd) - (u - v)d((u - v)d + 2uvh)\} \\ &= \Sigma \{2uvd^2 - 2uv(u - v)dh\} = \frac{1}{2}D - \frac{1}{4}F. \end{aligned}$$

Now consider the table of progeny means having selfs in the diagonal and $p(p - 1)$ F_1 s in off-diagonal positions. Let E_P be the environmental error variance component of parental family means in diagonals, and E_F , the error variance for F_1 family means in the off-diagonals. The sufficient statistics for estimation of D , H_1 , H_2 and F are:

$$\begin{aligned} V_P &= D + E_P \\ V_{\bar{r}} &= \frac{1}{4}D + \frac{1}{4}H_1 - \frac{1}{4}H_2 - \frac{1}{4}F + (E_P + (n-1)E_F/2)/n^2 \\ V_I &= \frac{1}{4}H_2 + \frac{1}{n}\left(E_P + \frac{1}{2}(n-1)E_F\right) \\ W_{\bar{r}} &= \frac{1}{2}D - \frac{1}{4}F + \frac{1}{n}E_P. \end{aligned}$$

These statistics facilitate inferences on the distribution of alleles using $H_2/(4H_1)$, dominance using mean degree of dominance, $(H_1/D)^{1/2}$, and test for the assumptions on the genetic system. The $W_r - V_r$ graph helps in identifying the dominant/recessive parents. The above expressions for V_r and W_r are derived under the assumptions that the trait is well explained by an additive-dominance model with additive environmental effects and independence of genes in action and in distribution. Evaluating V_r and W_r for each single array as V_n and W_n , respectively, say, $i = 1, \dots, p$, we see that $W_n - V_n = (1/4)D - (1/4)H_1$, or equivalently, $W_n = (1/4)(D - H_1) + bV_n$ with $b = 1$, which is of the linear regression form: $y = a + bx$. The overall level of dominance over all loci determines the intersection of this regression line with the W_r axis. The parental lines are distributed along this regression line on the $W_r - V_r$ graph inside a limiting parabola given by $W_r^2 = V_r V_P$. The points closest to the intersection of the line and parabola determine completely dominant and recessive genes. Completely dominant genes occur near the origin with low W_r values, while recessive genes occur away from the origin with higher W_r values. The regression line of W_r on V_r will have unit slope under independent distribution of genes in parents and absence of epistasis.

A number of test statistics are used for testing the adequacy of the model. The hypotheses of genetic systems when the model is adequate lead to the constancy of $W_n - V_n$ over the arrays $i = 1, \dots, p$, and can be tested in two ways:

1. For replicated trials in blocks, $W_r - V_r$ can be computed for each block. If the constancy of $W_n - V_n$ does not hold over the arrays, the line effect is significant in the ANOVA, and the hypotheses of adequacy of the model are not tenable.
2. In unreplicated experiments, (W_r, V_r) data are used to test the departure of the slope of the regression line from unity using the t^2 statistic, where

$$t^2 = \frac{(p-2)(\text{var}(W_r) - \text{var}(V_r))^2}{4(\text{var}(V_r)\text{var}(W_r) - \text{cov}^2(V_r, W_r))},$$

follows an F distribution with 4 and $n - 2$ df.

For an adequate model, a measure of the average level of dominance is provided by the intercept of the regression line $W_n = (1/4)(D - H_1) + bV_n$, that is, when $V_r = 0$, in which case, a positive intercept ($D > H_1$) indicates partial dominance, zero intercept ($D = H_1$) indicates complete dominance, and a negative value ($D < H_1$) indicates overdominance.

The *number of genes* distributed independently that control the trait and exhibit dominance is given by h^2/H_2 under certain restrictions on equality of gene effects. Here, h is twice the difference between the mean of the parents and the mean of their p^2 progenies. As to the *dominance* and *size*, the direction of dominance is indicated by the sign of h . If the correlation between selfs data in the diagonals of the diallel table and $W_r + V_r$ computed from the pairs of values $(y_i, W_n + V_n)$, $i = 1, \dots, p$ is close to 1 (or minus one), then the recessive genes (dominant genes) are mostly positive. The ratio of total number of dominant to recessive genes in all the parents is given by:

$$\frac{(4DH_1)^{\frac{1}{2}} + F}{(4DH_1)^{\frac{1}{2}} - F}.$$

Lastly, we consider the *heritability*. If the error variance E is derived as the mean square of interaction between the family and block means, the two heritabilities in the narrow and broad senses can respectively be computed as

$$h_n^2 = \frac{\frac{1}{2}D + \frac{1}{2}H_1 - \frac{1}{2}H_2 - \frac{1}{2}F}{\frac{1}{2}D + \frac{1}{2}H_1 - \frac{1}{4}H_2 - \frac{1}{2}F + E} \quad \text{and} \quad h_b^2 = \frac{\frac{1}{2}D + \frac{1}{2}H_1 - \frac{1}{4}H_2 - \frac{1}{2}F}{\frac{1}{2}D + \frac{1}{2}H_1 - \frac{1}{4}H_2 - \frac{1}{2}F + E}.$$

Example 1.1 (Continued). The $W_r - V_r$ graph on the above data set is shown in Figure 1.1.

Regression analysis on $p = 10$ pairs of (W_r, V_r) gave a slope $b = 0.957$ with a standard error (se) of 0.0585, indicating a slope of unity supporting independent distribution of genes in parents and absence of epistasis. Thus, the model is adequate. Also, the test based on t^2 does not reject the hypothesis of adequacy of the model at 5% level of significance. The positive intercept shows partial dominance. The parents P_{10} and P_6 being near to the origin contain most dominant alleles, while parent P_2 contains the most recessive alleles for days to flowering.

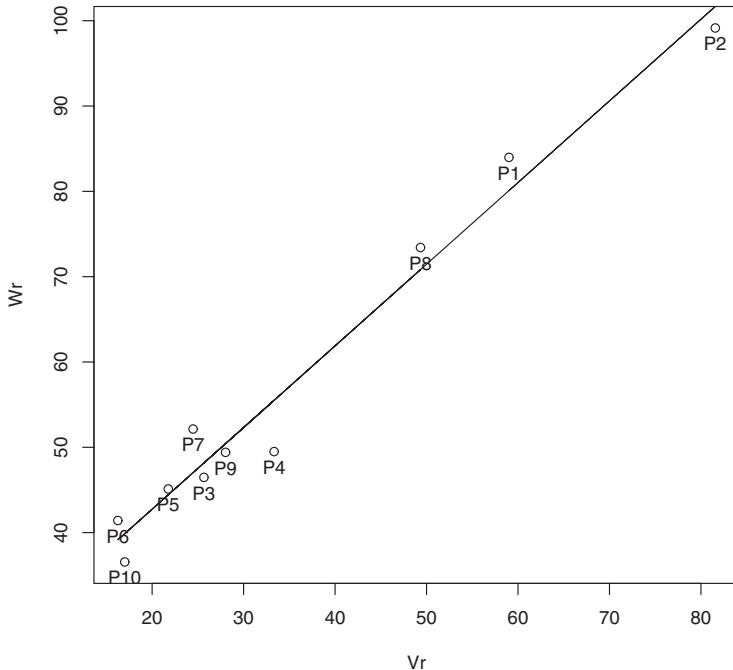


Figure 1.1 $W_r - V_r$ graph for the 10×10 diallel.

1.4.1.3 Generalization of Hayman Diallel Approach to Line \times Tester Designs

Patel, Christie, and Kannenberg (1984) generalized Hayman's approach of diallel analysis to a line \times tester situation (type I design), and derived corresponding statistics to detect and estimate additive and dominance effects. To test the departure from the additive-dominance model, they obtained expectations of the statistics based on variances for array means of line parents and tester parents, and parent-progeny means covariances under the assumptions of absence of epistasis, linkage, correlated gene distribution, multiple alleles, and maternal effects. The presence of epistasis can be detected from nonconstancy of $W_i - V_i$ over arrays $i = 1, \dots, p$ and also from the deviation of the coefficient of regression of W_i on V_i from unity.

1.4.2 Griffing's Method

Griffing (1956) considered four methods for generating the crosses among a set of p inbred lines and their analysis for comparative and exploratory experiments. Considering the environmental design as an *RCBD*, Griffing presented four mating designs, called methods, and two models by taking the genetic effects as fixed or random. The mating designs are—method 1: all possible p^2

crosses (F_1), including reciprocals and selfs; method 2: crosses and selfs; method 3: crosses and reciprocals; and method 4: crosses only. Since the crosses provide information on the parents as well, inclusion of selfs in the experiment may not provide a substantial comparative advantage over crosses only. We shall summarize here results for the two most commonly used methods, methods 3 and 4, which do not involve selfs.

Griiffing's method 3 is based on one set of F_1 s and the reciprocals, but not the parents. Let the environmental design be an $RCBD$, and y_{ij} be the mean over the r blocks for the cross between lines i and j . The model for the response is

$$y_{ij} = \mu + g_i + g_j + s_{ij} + r_{ij} + e_{ij} \quad (i, j = 1, \dots, p).$$

The parameters g_i , s_{ij} , and r_{ij} are called *gca* effect of the line i , *sca* effect of the lines i and j , and their reciprocal effect in the cross $i \times j$, respectively. Also, μ is the general mean, and e_{ij} is the mean of the r plot errors averaged over blocks.

Model I assumes all effects g_i , s_{ij} , and r_{ij} to be fixed such that $s_{ij} = s_{ji}$, $r_{ij} = -r_{ji}$, and the constraints $\sum_i g_i = 0$ and $\sum_{i \neq j} s_{ij} = 0$ (for each j). Model II assumes all these effects are random, and $g_i \sim N(0, \sigma_g^2)$, $s_{ij} \sim N(0, \sigma_s^2)$ and $r_{ij} \sim N(0, \sigma_r^2)$. Further, the mean errors are assumed to be independently distributed with $e_{ij} \sim N(0, \sigma_e^2)$. The ANOVA for method 3 under the fixed and random effects model are given in Table 1.6.

For method 3, the estimates of general mean, *gca*, and *sca*, and their estimated standard errors (*se*) under model I are as follows.

$$\begin{aligned}\hat{\mu} &= \frac{1}{p(p-1)} y_{..} \text{ with } se(\hat{\mu}) = \sqrt{\frac{1}{rp(p-1)} MS_e} \\ \hat{g}_i &= \frac{1}{2p(p-2)} [p(y_{i.} + y_{.i}) - 2y_{..}] \text{ with } se(\hat{g}_i) = \sqrt{\frac{(p-1)}{2rp(p-2)} MS_e}, \\ \hat{s}_{ij} &= \frac{1}{2} (y_{ij} + y_{ji}) - \frac{1}{2(p-2)} (y_{i.} + y_{.i} + y_{j.} + y_{.j}) + \frac{1}{(p-1)(p-2)} y_{..} \\ &\quad \text{with } se(\hat{s}_{ij}) = \sqrt{\frac{(p-3)}{2r(p-1)} MS_e} \quad (i \neq j), \\ \hat{r}_{ij} &= \frac{1}{2} (y_{ij} - y_{ji}) \text{ with } se(\hat{r}_{ij}) = \sqrt{\frac{1}{2r} MS_e} \quad (i \neq j),\end{aligned}$$

Standard errors for the comparisons:

$$\begin{aligned}se(g_i - g_{i'}) &= \sqrt{\frac{1}{r(p-2)} MS_e} \quad (i \neq i'), \\ se(\hat{s}_{ij} - \hat{s}_{ik}) &= \sqrt{\frac{(p-3)}{r(p-2)} MS_e} \quad (i \neq j, k; j \neq k),\end{aligned}$$

Table 1.6 ANOVA for Data from Method 3

Source	df	Mean Square ^a	Expected Mean Square	
			Model I	Model II
<i>gca</i>	$p - 1$	MS_g	$\sigma_e^2 + \frac{2(p-2)}{p-1} \sum g_i^2$	$\sigma_e^2 + 2\sigma_s^2 + 2(p-2)\sigma_g^2$
<i>sca</i>	$p(p-3)/2$	MS_s	$\sigma_e^2 + \frac{2}{p(p-3)/2} \sum_{i < j} s_{ij}^2$	$\sigma_e^2 + 2\sigma_s^2$
Reciprocal	$p(p-1)/2$	MS_r	$\sigma_e^2 + \frac{2}{p(p-1)/2} \sum_{i < j} r_{ij}^2$	$\sigma_e^2 + 2\sigma_r^2$
Error	$(r-1) \cdot (p^2 - p - 1)$	MS_e/r	σ_e^2	σ_e^2

^a Where:

$$MS_g = \frac{1}{p-1} \left[\frac{1}{2(p-2)} \sum_i (y_{i.} + y_{.i})^2 - \frac{2}{p(p-2)} y_{..}^2 \right],$$

$$MS_s = \frac{2}{p(p-3)} \left[\frac{1}{2} \sum_{i < j} (y_{ij} + y_{ji})^2 - \frac{1}{2(p-2)} \sum_i (y_{i.} + y_{.i})^2 + \frac{1}{(p-1)(p-2)} y_{..}^2 \right],$$

$$MS_r = \frac{1}{p(p-1)} \sum_{i < j} (y_{ij} - y_{ji})^2,$$

$$y_{i.} = \sum_{j \neq i} y_{ij}, y_{.i} = \sum_{j \neq i} y_{ji}, y_{..} = \sum_{i \neq j} y_{ij}, \text{ and}$$

MS_e is the error mean square obtained from the replicated responses on the crosses and reciprocals.

$$se(\hat{s}_{ij} - \hat{s}_{kl}) = \sqrt{\frac{(p-4)}{r(p-2)} MS_e} \quad (i \neq j, k, l; j \neq k, l; k \neq l).$$

Under model II, the variance components $\sigma_g^2 = \text{Cov}(HS)$, $\sigma_s^2 = \text{Cov}(FS) - 2\text{Cov}(HS)$ and $\sigma_r^2 = \text{Cov}(FS) - 2\text{Cov}(MHS)$ can be estimated as:

$$\hat{\sigma}_g^2 = \frac{1}{2(p-2)} (MS_g - MS_s) \text{ with } se(\hat{\sigma}_g^2) \equiv \sqrt{\frac{1}{2(p-1)(p-2)^2} MS_g^2 + \frac{1}{p(p-2)^2(p-3)} MS_s^2},$$

$$\hat{\sigma}_s^2 = \frac{1}{2} (MS_s - MS_e/r) \text{ with } se(\hat{\sigma}_s^2) \equiv \sqrt{\frac{1}{p(p-3)} MS_s^2 + \frac{1}{2r^2 v_e} MS_e^2},$$

$$\hat{\sigma}_r^2 = \frac{1}{2} (MS_r - MS_e/r) \text{ with } se(\hat{\sigma}_r^2) \equiv \sqrt{\frac{1}{p(p-1)} MS_r^2 + \frac{1}{2r^2 v_e} MS_e^2},$$

$$\hat{\sigma}_e^2 = MS_e/r \text{ with } se(\hat{\sigma}_e^2) \equiv \sqrt{\frac{2}{v_e} MS_e/r},$$

where $v_e = (r-1)(p^2 - p - 1)$.

Table 1.7 ANOVA for Data from Method 4

Source	df	Mean Square ^a	Expected Mean Square	
			Model I	Model II
gca	$p - 1$	MS_g	$\sigma_e^2 + \frac{p-2}{p-1} \sum g_i^2$	$\sigma_e^2 + \sigma_s^2 + (p-2)\sigma_g^2$
sca	$p(p-3)/2$	MS_s	$\sigma_e^2 + \frac{1}{p(p-3)/2} \sum_{i < j} s_{ij}^2$	$\sigma_e^2 + \sigma_s^2$
Error	$(r-1)(p^2-p-2)/2$	MS_e/r	σ_e^2	σ_e^2

^a Where

$$MS_g = \frac{1}{p-1} \left[\frac{1}{p-2} \sum_i y_{..}^2 - \frac{4}{p(p-2)} y_{..}^2 \right]$$

$$MS_s = \frac{2}{p(p-3)} \left[\sum_{i < j} y_{ij}^2 - \frac{1}{p-2} \sum_i y_{..}^2 + \frac{2}{(p-1)(p-2)} y_{..}^2 \right]$$

$y_{..} = \sum_{j \neq i} y_{ij}$, $y_{..} = \sum_{i < j} y_{ij}$, $y_{ij} = y_{ji}$ for $j < i$, and

MS_e is the error mean square obtained from the replicated responses on the crosses.

Among all the four methods, Griffing's method 4 requires the least number of crosses ($p(p-1)/2$) and is based on growing only F_1 's. This method is most commonly used and suitable where reciprocal effects are absent. The model for the response is

$$y_{ij} = \mu + g_i + g_j + s_{ij} + e_{ij} \quad (i < j = 1, \dots, p), \quad (1.7)$$

where the parameters are as described for method 3. The ANOVA is presented in Table 1.7.

The following expressions are given for method 4. Under model I, the estimates of general mean, gca and sca , and their estimated standard errors are:

$$\hat{\mu} = \frac{2}{p(p-1)} y_{..} \text{ with } se(\hat{\mu}) = \sqrt{\frac{2}{rp(p-1)} MS_e}$$

$$\hat{g}_i = \frac{1}{p(p-2)} [py_{..} - 2y_{..}] \text{ with } se(\hat{g}_i) = \sqrt{\frac{p-1}{rp(p-2)} MS_e}$$

$$\hat{s}_{ij} = y_{ij} - \frac{1}{p-2} (y_{..} + y_{..}) + \frac{2}{(p-1)(p-2)} y_{..} \text{ with } se(\hat{s}_{ij}) = \sqrt{\frac{p-3}{r(p-1)} MS_e}$$

Standard errors for the comparisons:

$$se(\hat{g}_i - \hat{g}_{i'}) = \sqrt{\frac{2}{r(p-2)} MS_e} \quad (i \neq i'),$$

$$se(\hat{s}_{ij} - \hat{s}_{ik}) = \sqrt{\frac{2(p-3)}{r(p-2)} MS_e} \quad (i \neq j, k; j \neq k),$$

$$se(\hat{s}_{ij} - \hat{s}_{kl}) = \sqrt{\frac{2(p-4)}{r(p-2)} MS_e} \quad (i \neq j, k, l; j \neq k, l; k \neq l).$$

Under model II, the estimates of the variance components, $\sigma_g^2 = \text{Cov}(HS)$ and, $\sigma_s^2 = \text{Cov}(FS) - 2\text{Cov}(HS)$ and their estimated standard errors are:

$$\hat{\sigma}_g^2 = \frac{1}{p-2} (MS_g - MS_s) \text{ with}$$

$$se(\hat{\sigma}_g^2) \equiv \sqrt{\frac{2}{(p-1)(p-2)^2} MS_g^2 + \frac{4}{p(p-2)^2(p-3)} MS_s^2},$$

$$\hat{\sigma}_s^2 = MS_s - MS_e/r \text{ with } se(\hat{\sigma}_s^2) \equiv \sqrt{\frac{4}{p(p-3)} MS_s^2 + \frac{2}{r^2 v_e} MS_e^2},$$

$$\hat{\sigma}_e^2 = MS_e/r \text{ with } se(\hat{\sigma}_e^2) \equiv \sqrt{\frac{2}{v_e} MS_e^2/r},$$

where $v_e = (r-1)(p^2-p-2)/2$.

In this case, the estimates of additive and dominance variance components are obtained as:

$$\hat{\sigma}_A^2 = 4\text{Cov}(HS) = 4\hat{\sigma}_g^2 \text{ and } \hat{\sigma}_D^2 = 4(\text{Cov}(FS) - 2\text{Cov}(HS)) = 4\hat{\sigma}_s^2$$

Example 1.2 Griffing Method 4. The F_1 data of 45 crosses, $p = 10$ from Example 1.1 are used to illustrate method 4. The code in R-language for carrying out this analysis is available in a supplementary file on the website. The ANOVA of Table 1.8 was carried out to get the error mean square of $MS_e/r = 3.73$, and the estimates of gca and sca are given in Table 1.9.

The estimates of statistical variance components for the analysis under model II are shown in Table 1.10, and the estimates of components of variance for additive and dominance effects are given in Table 1.11.

We note that the additive genetic effects measured by gca and the dominance effects (sca) are significant in the ANOVA table ($p < 0.001$), and their variance components σ_A^2 and σ_D^2 are positive. These results are consistent with Hayman's approach of analysis.

1.5 PARTIAL DIALLEL CROSSES: NO BLOCKING OR COMPLETE BLOCKS

In the absence of maternal effects, the set of all the $p(p-1)/2$ crosses with selfs and reciprocals excluded forms the *complete diallel cross (CDC)* (method

Table 1.8 ANOVA on Data from Method 4 and Model I

Source	df	SS	MS	vr	p-value
gca	9	2168.57	240.95	64.55	6.19×10^{-35}
sca	35	716.72	20.48	5.49	5.14×10^{-11}
Error	88	328.47	3.73		

Table 1.9 Estimates of gca and sca Effects^a from Method 4 and Model I

Lines	Lines									
	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
P_1	-0.38	9.40	-6.06	-2.95	-4.48	-2.84	-1.34	5.03	3.45	-0.21
P_2		0.34	-9.33	-3.32	-4.31	0.43	0.27	2.51	3.64	0.70
P_3			-6.42	-1.41	-0.20	3.19	4.56	4.24	-0.63	5.63
P_4				-4.76	1.60	2.86	-2.99	-4.00	8.10	2.10
P_5					-3.86	1.28	2.27	2.97	0.36	0.50
P_6						-1.90	-3.43	-0.96	0.27	-0.80
P_7							1.99	3.00	-6.07	3.70
P_8								8.75	-5.14	-7.66
P_9									9.77	-3.97
P_{10}										-3.53

^a The gca estimates are in diagonal and sca estimates are in off-diagonal positions. Their estimated standard errors are as follows:

$$se(\hat{\mu}) = 0.288$$

$$se(\hat{s}_i) = 0.648$$

$$se(\hat{s}_{ij}) = 1.704 \quad (i \neq j)$$

$$se(\hat{s}_i - \hat{g}_j) = 0.966 \quad (i \neq j)$$

$$se(\hat{s}_{ij} - \hat{s}_{ik}) = 2.556 \quad (i \neq j, k; j \neq k)$$

$$se(\hat{s}_{ij} - \hat{s}_{kl}) = 2.366 \quad (i \neq j, k, l; j \neq k, l; k \neq l).$$

Table 1.10 ANOVA for Method 4 and Model II

Source	df	Estimated Variance Component	se
gca : σ_g^2	9	27.56	41.95
sca : σ_s^2	35	16.75	4.90
Error : σ_e^2	88	3.73	0.80

Table 1.11 Genetic Variance Components for Method 4 and Model II

Genetic Component	Estimate	Estimated Standard Error
Additive variance (σ_A^2)	55.12	83.89
Dominance variance (σ_D^2)	16.75	4.90

4, Griffing 1956). With a moderately large value of p , the number of crosses $n = p(p - 1)/2$ becomes unmanageable. For such situations, a smaller sample of size n_c ($< n = p(p - 1)/2$) crosses from the *CDC*, called *partial diallel cross (PDC)*, was suggested by Gilbert (1958) as fractional diallel crosss, (see also Hinkelmann and Stern 1960, Kempthorne and Curnow 1961, and Curnow 1963). We review here various methods of construction of a *PDC*. Several methods for selecting the sample crosses are presented along with their efficiency factors by Kempthorne and Curnow (1961), Curnow (1963), Hinkelmann and Kempthorne (1963), Fyfe and Gilbert (1963), Narain, Subbarao, and Nigam (1974), Arya and Narain (1977), Narain and Arya (1981) and Arya (1983), by generally using circular samples or association schemes of partially balanced incomplete block (*PBIB*) designs. For the definition of *PBIB* design, see HK2, chapter 4. An algorithm for generating efficient *PDC* plans has been given by Singh and Hinkelmann (1988).

We present here first the method of Kempthorne and Curnow (1961) for generating the *PDCs*. Let p denote the number of distinct lines in a crossing program and s the common number of lines each line is crossed with different lines, also called the mating replication of a line. The quantities p and s are the parameters of the *PDC*. Let us denote the class of *PDC* with parameters p and s , by $PDC(p, s)$. The number of crosses in a $PDC(p, s)$ is $n = ps/2$. Some obvious restrictions on integers p and s are: $s > 2$, and both p and s both cannot be odd. Labeling randomly the p lines as $1, 2, \dots, p$, we have the sample crosses as

$$\text{line } i \times \text{lines } k+i, k+i+1, \dots, k+i-1+s, \pmod{p},$$

where $i = 1, 2, \dots, p$; s is such that $k = (p + 1 - s)/2$ is a whole number, and all numbers greater than p are reduced mod (p) to enable the line labels to be between 1 and p . Thus, $s = p - 1$ will yield the complete diallel crosses.

If these crosses are evaluated in an *RCBD* with r replications and the response from the k -th block on the cross $i \times j$ recorded, then on the mean of responses of the cross $i \times j$, one fits the model presented in Equation (1.7). The estimates of variance components can be obtained from the ANOVA given in Table 1.12.

If the lines are completely inbred, then in the absence of epistasis, $\sigma_g^2 = \text{Cov}(HS) = \frac{1}{2}\sigma_A^2$ and $\sigma_s^2 = \text{Cov}(FS) - 2\text{Cov}(HS) = \sigma_D^2$. Thus we can estimate the total genetic variance $\sigma_A^2 + \sigma_D^2$ by $2\hat{\sigma}_g^2 + \hat{\sigma}_s^2$, the average degree of

Table 1.12 ANOVA for $PDC(p, s)$

Source	df	Mean Square	Expected Mean Square
Replications	$r - 1$		
gca	$p - 1$	MS_g	$\sigma_e^2 + r\sigma_s^2 + rs(p - 2)(p - 1)^{-1}\sigma_g^2$
sca	$p(s/2 - 1)$	MS_s	$\sigma_e^2 + r\sigma_s^2$
Error	$(r - 1)(ps/2 - 1)$	MS_e	σ_e^2

dominance $\sqrt{2\sigma_D^2/\sigma_A^2}$ (when gene frequencies are 1/2) by $\sqrt{\hat{\sigma}_s^2/\hat{\sigma}_g^2}$, and the heritability in the narrow sense $h_n^2 = \sigma_A^2 / (\sigma_A^2 + \sigma_D^2 + \sigma_e^2)$ by $2\hat{\sigma}_g^2 / (2\hat{\sigma}_g^2 + \hat{\sigma}_s^2 + \hat{\sigma}_e^2)$.

With a view to estimating *gca* effects of a set of specific lines, minimizing the sum of squares over the sampled crosses, that is, $\Sigma(y_{ij} - \mu - g_i - g_j)^2$, using the constraint $\Sigma g_i = 0$ yields the equations:

$$\sum_{j=1} a_{ij} \hat{g}_j = \sum_{r=0}^{s-1} \left[y_{i,i+k+r} - \frac{2G}{ns} \right] \equiv Q_i, \quad (i = 1, 2, \dots, p),$$

where G is the grand total of cross averages. Using matrix notations, let $\mathbf{A} = (a_{ij})$ where $a_{ii} = s$ and $a_{ij} = a_{ji} = 1$ if cross $i \times j$ is included in the sample and $a_{ij} = a_{ji} = 0$ otherwise. The matrix \mathbf{A} being a circulant matrix, its inverse is also circulant and can be obtained in terms of the coefficients of \mathbf{A} and the roots of unity. Writing $\mathbf{A} = (a_{ij}) = (a_k)$ and $\mathbf{A}^{-1} = (a^{ij}) = (a^k)$, the elements of \mathbf{A}^{-1} can be expressed as:

$$a^0 = \frac{1}{p} \sum_{i=1}^p \frac{1}{\lambda_i}, \quad a^j = \frac{1}{p} \sum_{l=1}^p \frac{1}{\lambda_l} \cos \frac{j(p-l)}{p} 2\pi, \quad j = 1, 2, \dots, p-1,$$

where

$$\lambda_j = s - \frac{\sin(p-s)\frac{j\pi}{p}}{\sin\frac{j\pi}{p}}, \quad \lambda_0 = 2s, \quad j = 1, 2, \dots, p-1.$$

Thus, $\hat{\mathbf{g}} = \mathbf{A}^{-1} \mathbf{Q}$, where column vectors $\mathbf{g} = (g_1, g_2, \dots, g_p)'$ and $\mathbf{Q} = (Q_1, Q_2, \dots, Q_p)'$. In Table 1.12, the mean squares due to *gca* and *sca* are obtained using the respective sum of squares $SS_g = r \Sigma \hat{g}_i Q_i$, $SS_s = r \Sigma (y_{ij} - \bar{y} - \hat{g}_i - \hat{g}_j)^2$, where \bar{y} is the mean of cross averages. The mean square MS_e is the sampled cross \times replication or block interaction mean square.

Keeping the simplifying features of the *PDC* designs of Kempthorne and Curnow (1961) in relation to *BIB* designs, Curnow (1963) and Hinkelmann and Kempthorne (1963) extended the concept of sampling crosses using a *PBIB* design of block size 2 (for the definition of *BIB* and *PBIB* designs see HK2, Chapters 2 and 4). Using some of the two-plot block designs listed in Clatworthy (1955), Curnow (1963) evaluated the variances of *gca* effects yielding two or more distinct values. Fyfe and Gilbert (1963), using a triangular or rectangular arrangement of integers, constructed *PDCs* which were better balanced for estimation of *gcss*. Establishing a correspondence between *PDC* and *PBIB* designs with m associate classes, Hinkelmann and Kempthorne (1963) gave a general method of analysing the data if evaluated in complete blocks or no blocks. Their formal definition of the *PDC* design is as follows:

Definition 1.1. A mating system is said to be a *PDC* if it satisfies the conditions:

1. each line is crossed to s different lines; and
2. the number of times line j is crossed to line i is either zero or one.

Thus, the total number of crosses equals $ps/2$.

Now restricting the *PBIBs* with block size 2 to λ_k values of zero and one, the sampled *PDC* will be the cross between the lines within a block, where a treatment of the *PBIB* corresponds to a line of the *PDC*. The number of sampled crosses will be equal to the number of blocks in the *PBIB*, $b = tr/k = ps/2$. In order to carry out the analysis, we assume the following model for y_{ij} , the mean yield of the cross $i \times j$,

$$y_{ij} = \mu + g_i + g_j + s_{ij}, \quad (i, j = 1, 2, \dots, p; i \neq j),$$

where μ is the general mean, and g_i is the *gca* of the i th line, and s_{ij} is the residual containing specific combining ability effect and the experimental error. Further, the s_{ij} are assumed independently and identically distributed with mean zero and variance σ^2 . The above model is useful when either *sca* is absent or negligible. The normal equations for estimation of the *gca* can be written along the lines of Kempthorne and Curnow (1961) as $\mathbf{A}\hat{\mathbf{g}} = \mathbf{Q}$, where $\mathbf{A} = (a_{ij})$ and a_{ij} is 1 if cross $i \times j$ is sampled and 0 otherwise.

Writing $\mathbf{A}^{-1} = (a^{ij})$, \mathbf{A}^{-1} has at most $m + 1$ distinct elements, say a^0, a^1, \dots, a^m such that if lines i and j are k th associates, then $a^{ij} = a^k$. Further, these unknowns can be obtained explicitly from the equations:

$$\sum_{i=0}^m \sum_{j=0}^m \lambda_j p_{ij}^k a^i = \delta_{k0}, \quad (k = 0, 1, \dots, m),$$

where δ_{k0} is the Kronecker δ and the p_{ij}^k are the parameters of the second kind for the underlying *PBIB* design. The variance of difference of *gca* estimates of any two lines i and j is

$$V(\hat{g}_i - \hat{g}_j) = (a^{ii} + a^{jj} - 2a^{ij})\sigma^2 = 2(a^0 - a^{ij})\sigma^2,$$

and the average variance over all possible comparisons is

$$Av.V(\hat{g}_i - \hat{g}_j) = [(2spa^0 - 1)/(s(p-1))]\sigma^2.$$

The variance for the *CDC* will be $V(\hat{g}_i - \hat{g}_j) = [2(p-1)/(s(p-2))]\sigma_0^2$ for all i and j , where σ_0^2 is the appropriate error variance. The efficiency factor of a *PDC* relative to a *CDC* with the same numbers of lines and replicates is then given by $E = 2(p-1)^2/(p-2)(2spa^0 - 1)$, which is the ratio of average

variances of *gca* contrast under CDC to that under PDC, and $E^* = E(p - 1)/s$ represents the efficiency factor on a per cross basis. The E or E^* can be used to make comparisons between competing *PDCs*.

Hinkelmann and Kempthorne (1963) also introduced two additional classes of *PDCs*, one abbreviated as *GGD/m – PDC* based on an m associate class generalization of group divisible *PBIB* with two associate classes (Roy 1953–1954), and the other abbreviated as *EGD/(2^v – 1) – PDC* based on a $2^v - 1$ associate class *PBIB*, extension of a *PBIB* with three associate classes given by Vartak (1959) (see HK2, chapter 4).

Federer (1967) observed the *CDC* plans of Griffing (1956) as a fractional replicate of the p^2 factorial design, and used one of the interaction components of the two factors at p levels in constructing a *PDC*. Arya and Narain (1977) employed *PBIB* designs with three or four associate classes to construct *PDCs* and found that the *PDCs* using group divisible designs with three associate classes are more efficient than those based on generalized right angular designs with four associate classes.

Pederson (1980) introduced *augmented partial diallel crosses* for the situation where some of the lines (called primary lines) could be superior or well adapted, and hence of higher importance and may be represented in high proportion, while the remaining lines could be of secondary importance and may be evaluated through a partial diallel cross design. If the number of primary and secondary lines are p and q , respectively, then the design is obtained by (1) crossing each primary line to every other line resulting in $(p + q - 1)$ crosses for each primary line, and (2) using a *PDC* design, say, with $p + s$ crosses per secondary line. The number of crosses in the augmented *PDC* is $\frac{1}{2}\{p(p + q - 1) + q(p + s)\}$ when the reciprocals are excluded. He gave approximate expressions for the four variances of interline comparisons. Narain and Arya (1981) presented a circulant plan for *PDCs* for any p and s , except the trivial impossibility of both being odd. These plans are based on group divisible designs with two or three associate classes and are more efficient than those of Kempthorne and Curnow (1961), and also fill the gaps with the best designs available in the range $p = 13$ to 24.

Singh and Hinkelmann (1990) developed *PDCs* in terms of so called basic *PDCs* (*BPDC*) and derived expressions for the eigenvalues of the coefficient matrix \mathbf{A} , computed the efficiency factor and listed 112 designs in the range of $10 \leq p \leq 30$ and $4 \leq s \leq \min(p/2 \text{ or } (p - 1)/2)$. The *BPDCs* were defined as: (1) For odd p , the *BPDCs* are S_1, S_2, \dots, S_m , where $S_l = \{i \times (i + l), i = 1, 2, \dots, p, \text{mod}(p)\}, l = 1, 2, \dots, m, m = (p - 1)/2$. These *PDCs* are connected, where a design is called connected for estimation of *gca*, if every two lines can be chained through other lines having crosses in the sample (Arya 1983). For example, for $p = 8$, the set of crosses $\{1 \times 2, 2 \times 3, 3 \times 4, 4 \times 5, 5 \times 6, 6 \times 7, 7 \times 8, 1 \times 8\}$ forms a connected design, while the set $\{1 \times 5, 2 \times 6, 3 \times 7, 4 \times 8\}$ does not. (2) For even p , the $m - 1$ connected *BPDCs* are $S_1, S_2, \dots, S_{m-1}, (m = p/2)$, where $S_l = \{i \times (i + l), i = 1, 2, \dots, p, \text{mod}(p)\}, l = 1, 2, \dots, m - 1$. However, the *PDC* $S_m = \{i \times (i + m), i = 1, 2, \dots, m\}$ is disconnected.

Singh and Hinkelmann (1990) constructed a *PDC* design with s mating replications by combining $s/2$ of the *BPDCs*, and gave the following result.

Theorem 1.1. The information matrix \mathbf{A} for estimating *gca* obtained by combining the distinct *BPDCs* $S_{l_1}, S_{l_2}, \dots, S_{l_k}$ is given by

$$\mathbf{A} = \sum_{j=0}^{p-1} \lambda_j \mathbf{v}_j \mathbf{v}'_j,$$

where $\lambda_j = 2k + 2\sum_{l \in \{l_1, l_2, \dots, l_k\}} \cos(2\pi jl/p)$, $l_i = 1, 2, \dots, m$, $i = 1, 2, \dots, k$, $m = (p-1)/2$ for odd p and $(p/2-1)$ for even p , are the eigenvalues of \mathbf{A} corresponding to the eigenvector \mathbf{v}_j . The design based on $S_{l_1}, S_{l_2}, \dots, S_{l_k}$ gives a *PDC* with $s = 2k$ (even) mating replications.

When p is even, a *PDC* design can be constructed from the *BPDCs* $S_{l_1}, S_{l_2}, \dots, S_{l_{k-1}}$ and S_m ($m = p/2$). For such a design, the mating replication is $s = 2k - 1$ (odd) and the matrix \mathbf{A} will have the eigenvalues

$$\lambda_j = 2k - 1 + 2 \sum_{l \in \{l_1, l_2, \dots, l_{k-1}\}} \cos(2\pi jl/p) + \cos(2\pi jm/p), \\ l_i = 1, 2, \dots, m, i = 1, 2, \dots, k.$$

The average efficiency factor of a *PDC* compared with a *CDC* with $p(p-1)/2$ crosses is

$$E = \frac{(p-1)}{(p-2) \sum_{j=0}^{p-1} \lambda_j^{-1}} \text{ and } E^* = (p-1)E/s,$$

where E^* is the efficiency per cross basis.

Example 1.3. For $p = 10$, $s = 4$, the *BPDCs* are S_2 and S_4 , which yield the *PDC* design:

$$S_2 = \{1 \times 3, 2 \times 4, 3 \times 5, 4 \times 6, 5 \times 7, 6 \times 8, 7 \times 9, 8 \times 10, 9 \times 1, 10 \times 2\} \\ S_4 = \{1 \times 5, 2 \times 6, 3 \times 7, 4 \times 8, 5 \times 9, 6 \times 10, 7 \times 1, 8 \times 2, 9 \times 3, 10 \times 4\}.$$

In this case, $E = 0.3857$, $E^* = 0.8679$.

Example 1.4. For $p = 10$, $s = 5$, the *BPDCs* are S_1 , S_4 , and S_5 yielding the *PDC*:

$$S_1 = \{1 \times 2, 2 \times 3, 3 \times 4, 4 \times 5, 5 \times 6, 6 \times 7, 7 \times 8, 8 \times 9, 9 \times 10, 10 \times 1\} \\ S_4 = \{1 \times 5, 2 \times 6, 3 \times 7, 4 \times 8, 5 \times 9, 6 \times 10, 7 \times 1, 8 \times 2, 9 \times 3, 10 \times 4\} \\ S_5 = \{1 \times 6, 2 \times 7, 3 \times 8, 4 \times 9, 5 \times 10, 6 \times 1, 7 \times 2, 8 \times 3, 9 \times 4, 10 \times 5\},$$

with the efficiency factors $E = 0.4787$, $E^* = 0.8617$.

1.6 PARTIAL DIALLEL CROSSES IN INCOMPLETE BLOCKS

Agarwal and Das (1990) constructed a number of *PDC* in incomplete blocks using n -ary *PBIB* and *BIB* designs, and also gave an expression for estimating a missing observation. For v treatments, consider (1) a two associate *PBIB* design with parameters $v, b, r, k, \lambda_l, n_l, p_{jl}^i$, ($i, j, l = 1, 2$) such that one of the λ_l 's is zero and the other unity, and (2) a suitable *BIB* design with parameters $v' (= b), b', r', k', \lambda'$. Let the v lines of this *PDC* design correspond to the v treatments of the *PBIB* design, then all possible $k(k - 1)/2$ crosses were formed within each block of size k of the *PBIB*. Let the set of crosses from the i th block of the *PBIB* be denoted by c_i ($i = 1, 2, \dots, b$). Consider a correspondence between $v' = b$ treatments of the *BIB* design with the c_1, c_2, \dots, c_b . These c_i , when replaced by the crosses they represent, give the *PDC* in an incomplete block design.

Example 1.5. Consider $p = 9$ and the *PBIB* design with parameters: $v = 9$, $b = 9$, $r = 3$, $k = 3$, $\lambda_1 = 1$, $\lambda_2 = 0$, $n_1 = 6$, $n_2 = 2$, $p_{11}^1 = 3$, $p_{12}^1 = 2$, $p_{22}^1 = 0$, $p_{11}^2 = 6$, $p_{12}^2 = 0$, $p_{22}^2 = 1$. Table 1.13 shows the *PBIB* design and the set of crosses c_i ($i = 1, \dots, 9$) formed from each of the blocks.

Now consider the *BIB* with $v' = b = 9, b' = 12, r' = 4, k' = 3, \lambda' = 1$. Expanding the block contents of the *BIB* design in terms of c_i ($i = 1, \dots, 9$), we get the *PDC* in 12 blocks of 9 plots each. The *BIB* design and the resulting *PDC* design are shown in Table 1.14.

The design in Table 1.14 constructed by this method may be seen to require a large number of experimental plots due to the large number of replications and large block sizes, and may restrict their use in practice. Using the circulant designs for obtaining *PDCs*, Gupta, Das, and Kageyama (1995) presented a method for constructing single replicate *PDCs* and noted that the designs, being orthogonal, do not entail any loss of efficiency on the contrasts of *gcas* due to blocking.

Table 1.13 PBIB Design and the Set of Crosses from Its Blocks

Block, i	PBIB Design			Resulting Set of Crosses		
	Treatments			c_i		
1	(1	2	3)	1 × 2	1 × 3	2 × 3
2	(1	6	4)	1 × 4	1 × 6	4 × 6
3	(1	7	5)	1 × 5	1 × 7	5 × 7
4	(6	8	3)	3 × 6	3 × 8	6 × 8
5	(6	9	5)	5 × 6	5 × 9	6 × 9
6	(7	8	4)	4 × 7	4 × 8	7 × 8
7	(7	9	3)	3 × 7	3 × 9	7 × 9
8	(2	8	5)	2 × 5	2 × 8	5 × 8
9	(2	9	4)	2 × 4	2 × 9	4 × 9

Table 1.14 BIB and PDC Designs

Block	Treatments(i)	Resulting PDC Design Obtained by Replacing Treatment i of the BIB Design by the Set of Crosses c_i									
		1×2	1×3	2×3	1×4	1×6	4×6	1×5	1×7	5×7	
B_1	(1 2 3)	1×2	1×3	2×3	1×4	1×6	4×6	1×5	1×7	5×7	
B_2	(1 4 7)	1×2	1×3	2×3	3×6	3×8	6×8	3×7	3×9	7×9	
B_3	(1 5 9)	1×2	1×3	2×3	5×6	5×9	6×9	2×4	2×9	4×9	
B_4	(1 6 8)	1×2	1×3	2×3	4×7	4×8	7×8	2×5	2×8	5×8	
B_5	(2 4 9)	1×4	1×6	4×6	3×6	3×8	6×8	2×4	2×9	4×9	
B_6	(2 5 8)	1×4	1×6	4×6	5×6	5×9	6×9	2×5	2×8	5×8	
B_7	(2 6 7)	1×4	1×6	4×6	4×7	4×8	7×8	3×7	3×9	7×9	
B_8	(3 4 8)	1×5	1×7	5×7	3×6	3×8	6×8	2×5	2×8	5×8	
B_9	(3 5 8)	1×5	1×7	5×7	5×6	5×9	6×9	3×7	3×9	7×9	
B_{10}	(3 6 9)	1×5	1×7	5×7	4×7	4×8	7×8	2×4	2×9	4×9	
B_{11}	(4 5 6)	3×6	3×8	6×8	5×6	5×9	6×9	4×7	4×8	7×8	
B_{12}	(7 8 9)	3×7	3×9	7×9	2×5	2×8	5×8	2×4	2×9	4×9	

PDCs have been constructed using the analogy of incomplete block design with blocks of size 2. Although the PDCs contain fewer crosses than CDCs (Griffing method IV), it still could result in a large number of crosses to be accommodated in complete blocks. To address this aspect of the size of PDCs, Singh and Hinkelmann (1998) have presented in detail the design and analysis of partial diallel crosses in incomplete blocks. We shall include their method of construction using an imbedding of mating and environmental ($M-E$) designs and the resulting statistical analysis of data, illustrating the computation with hypothetical data.

1.6.1 Construction of Mating–Environment Designs

Using a relationship between M design and incomplete block designs (IBD) (Curnow 1963; Hinkelmann and Kempthorne 1963), we illustrate here the construction of an $M-E$ design. Consider a PBIB design with t treatments, b blocks, r replications per treatment, block size $k = 2$, n_l th associates occurring together λ_l times in the same block ($l = 1, 2, \dots, m$). Such a design is called an auxiliary design (A -design). This design results in a PDC with p lines, each line being crossed to s other lines, where

$$t = p, r = s, b = ps/2 \equiv n_c, \text{ and } s = \sum_{l=1}^m \lambda_l n_l, \text{ where } \lambda_l = 0 \text{ or } 1.$$

Let N be the treatment–block incidence matrix of the above PBIB design. The set of crosses (M design) from the above PBIB design can be described by its cross-line incidence matrix, denoted by Z with n_c rows and p columns. The elements of Z are 1 if a line appears in a particular cross, and 0 otherwise. We can easily see that

$$\mathbf{Z} = \mathbf{N}'.$$

We shall take a *PBIB* design as an *E* design and embed the above *M* design to produce the final design of a *PDC* in an incomplete block design. Let the *E* design be written in terms of its cross-block incidence matrix, ψ with n_c rows and b^* blocks, with elements equal to 1 if a particular cross occurs in a particular block, and 0 otherwise. By identifying the crosses with treatments, we can choose any incomplete block design with n_c treatments, b^* blocks, block size k^* , and replications r^* . One choice of *E* design can be the dual of the *M* design, that is, $\psi = \mathbf{N}'$. In this case, $b^* = t = p$, $r^* = 2$, $k^* = r$, and $\psi = \mathbf{Z}$.

Example 1.6. Consider the triangular association scheme with $t = 10$ treatments (see Clatworthy 1973):

$$\begin{array}{ccccc} * & 1 & 2 & 3 & 4 \\ 1 & * & 5 & 6 & 7 \\ 2 & 5 & * & 8 & 9 \\ 3 & 6 & 7 & * & 10 \\ 4 & 7 & 9 & 10 & * \end{array}$$

The association scheme is such that the treatments within the same row or same column are the first associates and others are second associates. For a block design with $\lambda_1 = 0$ and $\lambda_2 = 1$, we obtain the *PBIB* design (*A*-design) with $k = 2$, $b = 15$, $r = 3$ parameters in $t = 10$ treatments with the following incidence matrix:

$$\mathbf{N} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

The above matrix gives the following correspondence between treatments and lines, and crosses and blocks for $p = 10$, $n_c = 15$ of the *M* design:

The 15 crosses sampled in \mathbf{N} are $1 \times 8, 1 \times 9, 1 \times 10, 2 \times 6, 2 \times 7, 2 \times 10, 3 \times 5, 3 \times 7, 3 \times 9, 4 \times 5, 4 \times 6, 4 \times 8, 5 \times 10, 6 \times 9, 7 \times 8$.

Table 1.15 Incidence Matrix and Design for $p = 10$

(a)

Blocks

Cross	1	2	3	4	5	6	7	8	9	10
1 × 8	1							1		
1 × 9	1								1	
1 × 10	1									1
2 × 6		1					1			
2 × 7		1						1		
2 × 10		1								1
3 × 5			1			1				
3 × 7			1					1		
3 × 9			1						1	
4 × 5				1	1					
4 × 6					1		1			
4 × 8					1				1	
5 × 10						1				1
6 × 9							1			1
7 × 8								1	1	

(b)

Block	Crosses
1	1 × 8
2	2 × 6
3	3 × 5
4	4 × 5
5	3 × 5
6	2 × 6
7	2 × 7
8	1 × 8
9	1 × 9
10	1 × 10

Taking $\psi = \mathbf{N}'$, the 15×10 matrix gives the incidence matrix of Table 1.15a.

Rewriting Table 1.15a gives rise to the incomplete block design in Table 1.15b.

Thus, a *PBIB* design with parameters $t = p, b, r = s, k = 2, n_i (i = 1, 2, \dots, m), \lambda_i = 0, 1, \mathbf{N}$ results in an *M* design: $PDC(p, s, \mathbf{Z} = \mathbf{N}')$ and an incomplete block *E* design with parameters $n_c, b^* = p, r^* = 2, k^* = r, \psi = \mathbf{N}'$. The embedded *M-E* design therefore is a *PDC-IBD* with parameters $\mathbf{Z} = \mathbf{N}', \psi = \mathbf{N}'$. The embedded design, that is, *M-E* design, is denoted by $M-E (\mathbf{Z} = \mathbf{N}', \psi = \mathbf{N}')$ or, in short, $M-E (\mathbf{N}', \mathbf{N}')$ for the above \mathbf{N} matrix.

1.6.2 Analysis of M-E Design

1.6.2.1 Model without sca Effects

We can use the following model for the data, y_{ijl} from a *PDC-IBD*:

$$y_{ijl} = \mu + g_i + g_j + \beta_l + e_{ijl}, \quad (1.8)$$

where ij stands for the cross $i \times j$ included in the PDC, g_i is the general combining ability of line i ($i = 1, 2 \dots p$), β_l is the effect of block l ($l = 1, 2, \dots, b^* = p$), and e_{ijl} is an error term. We assume that the genetic effect of the cross $i \times j$ is expressed by the *gca* and that either the *sca* effects are negligible or are merged with error. Further, g_i and β_l are assumed fixed effects with $\Sigma g_i = 0$ and $\Sigma \beta_l = 0$. The e_{ijl} are independent and identically distributed random variables with mean 0 and variance σ_e^2 . Using a label $u = 1, 2, \dots, n$, where $n = n_c r^* = pr$, to denote the experimental units on which the offspring of a cross are grown, the above model can be written in matrix notation as

$$\mathbf{y} = \mu \mathbf{J} + \mathbf{X}_g \mathbf{g} + \mathbf{X}_\beta \boldsymbol{\beta} + \mathbf{e}, \quad (1.9)$$

where \mathbf{y} is an $n \times 1$ vector of observations; \mathbf{J} is an $n \times 1$ vector of unity elements; $\mathbf{X}_g = (x_{gui})$ is an $n \times p$ matrix with elements $x_{gui} = 1$ if one of the parents of the cross on unit u is line i , and 0 otherwise; $\mathbf{X}_\beta = (x_{\beta ui})$ is an $n \times b^*$ matrix with elements $x_{\beta ui} = 1$ if unit u is contained in block l , and 0 otherwise; $\mathbf{g} = (g_1, g_2, \dots, g_p)'$ is the vector of general combining abilities; $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_{b^*})'$ is the vector of block effects; and \mathbf{e} is the $n \times 1$ vector of errors. The least squares reduced normal equations and the information matrix \mathbf{C}_g for estimation of *gca* (\mathbf{g}) are given by

$$\mathbf{C}_g \hat{\mathbf{g}} = \mathbf{Q}_g, \quad (1.10)$$

where

$$\mathbf{C}_g = \mathbf{G} - \frac{1}{k_g} \mathbf{N}_g \mathbf{N}'_g, \quad (1.11)$$

and k_g is the number of crosses in each block, the diagonal elements of \mathbf{G} are replications of the lines and the off-diagonal elements are the replications of the crosses, and \mathbf{N}_g is the lines \times block incidence matrix for the diallel cross design, that is, it is obtained by ignoring the crosses, and thus considering $2k_g$ lines as contents of a block. Using various simplifying relationships under the *M-E* design, the information matrix \mathbf{C}_g gets simplified to

$$\mathbf{C}_g = 2\mathbf{NN}' - \frac{1}{r} \mathbf{NN}'\mathbf{NN}'.$$

We further have

$$\mathbf{Q}_g = \mathbf{X}'_g \mathbf{y} - \frac{1}{r} \mathbf{NN}'\mathbf{X}'_\beta \mathbf{y}.$$

Table 1.16 ANOVA for Model (1.9)

Source	df	SS^a	$E(MS)$
Blocks	$p - 1$	$\Sigma B_l^2/r - G^2/(pr)$	
<i>gca</i>	$p - 1$	$\hat{\mathbf{g}}' \mathbf{Q}_g$	$\sigma_e^2 + \mathbf{g}' \mathbf{C}_g \mathbf{g}/(p - 1)$
Error	$p(r - 2) + 1$	By subtraction	σ_e^2
Total	$pr - 1$	$\Sigma_{ijl} y_{ijl}^2 - G^2/(pr)$	

^a B_l , total for block l ; G , grand total.

Since the rank of \mathbf{C}_g is $\leq p - 1$, the solution to Equation (1.10) is given by

$$\hat{\mathbf{g}} = \mathbf{C}_g^- \mathbf{Q}_g,$$

where \mathbf{C}_g^- is a generalized inverse of \mathbf{C}_g . The estimator for any contrast, $\mathbf{c}'\mathbf{g}$, among the *gca* is given by

$$\mathbf{c}'\hat{\mathbf{g}} = \mathbf{c}'\mathbf{C}_g^- \mathbf{Q}_g \text{ with } \text{var}(\mathbf{c}'\hat{\mathbf{g}}) = \mathbf{c}'\mathbf{C}_g^- \mathbf{c}\sigma^2.$$

The analysis of variance for model Equation (1.9) is given in Table 1.16.

1.6.2.2 Efficiency Factors of M-E Designs

In general there may be several alternative *M-E* designs for evaluating the *gca* of p lines. A comparison of such designs can be carried out by looking at the efficiency of an *M-E* design relative to a *CDC* in an *RCBD* with the same number of replications. Singh and Hinkelmann (1995) have obtained efficiency factors of the following two classes of *M-E* designs.

1. *M-E* (N' , N') designs (see Section 1.6.1)

Denoting the average variance of the differences of the estimates of the *gca*, that is, $\hat{g}_i - \hat{g}_{i'}$, by var_C under *CDC-RCBD* and var_I under the chosen *M-E* design, the average efficiency factor of the *M-E* ($\mathbf{Z}, \boldsymbol{\psi}$) = *M-E* (N' , N') is given by

$$E_I = \frac{\text{var}_C}{\text{var}_I} = \frac{p(p-1)}{2(p-2)} \left/ \left(\sum_{i' \neq i} v_{ii'} \right) \right..$$

where $v_{ii'} = (c^{ii} + c^{i'i'} - 2c^{ii'})\sigma_e^2$, $\mathbf{C}_g^- = (c^{ii'})$. Further, let d_i ($i = 1, 2, \dots, p$) be the eigenvalues of NN' , which implies that

$$v_i = d_i \left(2 - \frac{1}{r} d_i \right),$$

($i = 1, 2, \dots, p$) are the eigenvalues of \mathbf{C}_g . Also $v_1 = 0$. This results in the average efficiency factor

$$E_I = \frac{p-1}{2(p-1) \sum_{i=1}^p \frac{1}{v_i}}.$$

Since the *CDC* and *PDC* have different number of crosses, that is, $p(p-1)/2$ versus $ps/2$, with $s = r < p - 1$, the efficiency factor on a per cross basis (Hinkelmann and Kempthorne 1963) is

$$E_I^* = \frac{p-1}{r} E_I.$$

Using the *PBIB* designs with two associate classes tabulated by Clatworthy (1973), the efficiency factors for $p \leq 27$, $s = r \leq 10$ are listed in Singh and Hinkelmann (1995).

2. Other E designs

Unlike the above case, the incidence matrices in an $M-E(\mathbf{Z}, \psi)$ design could be different. Let the E design be any general incomplete block design (e.g., *BIB* or *PBIB*) with n_c treatments, b^* blocks, r^* replications, and block size k^* . Let ψ be its $n_c \times b^*$ incidence matrix. Using the cross-line incidence matrix \mathbf{Z} , the general form of the \mathbf{C}_g matrix is

$$\mathbf{C}_g = r^* \mathbf{Z}' \mathbf{Z} - \frac{1}{k^*} \mathbf{Z}' \psi \psi' \mathbf{Z}.$$

If the E design is a *BIB* design with λ as the number of treatments common between every pair of blocks, then

$$\psi \psi' = (r^* - \lambda) \mathbf{I} + \lambda \mathbf{J} \mathbf{J}'.$$

where \mathbf{J} is the column vector of all unities of appropriate size. It can then be shown that

$$\mathbf{C}_g = r^* E_{BIB} \left(\mathbf{Z}' \mathbf{Z} - \frac{\lambda s^2}{r^* k^* E_{BIB}} \mathbf{J} \mathbf{J}' \right),$$

where $E_{BIB} = (k^* - 1)n_c/[k^*(n_c - 1)]$ is the efficiency factor of the *BIB* design. A generalized inverse for \mathbf{C}_g can be written as

$$\mathbf{C}_g^- = \frac{1}{r^* E_{BIB}} [(\mathbf{Z}' \mathbf{Z})^{-1} + \alpha \mathbf{J} \mathbf{J}'],$$

where

$$\alpha = \left(\frac{\lambda s^2}{r^* k^* E_{BIB}} \right)^2 \mathbf{J}' (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{J} - 2 \left(\frac{\lambda s^2}{r^* k^* E_{BIB}} \right).$$

The variance of the estimate of any contrast $\mathbf{c}'\hat{\mathbf{g}}$ among the *gca* values can be written as

$$\text{var}(\mathbf{c}'\hat{\mathbf{g}}) = \frac{1}{r^*E_{BIB}} \mathbf{c}'(\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{c} \sigma_e^2.$$

This results in the following expression for the efficiency factor for the *M-E* design, say E_{M-E} ,

$$E_{M-E} = E_{BIB} E_{PDC},$$

where E_{PDC} is the efficiency factor of the *PDC* as given by Kempthorne and Curnow (1961). On a per cross basis, the efficiency factor is given by

$$E_{M-E}^* = \frac{p-1}{s} E_{M-E}.$$

These have been listed for $p \leq 24, s \leq 9, k^* \leq 15$ (Singh and Hinkelmann 1995).

1.6.2.3 Model with Specific Combining Ability Effects

Taking a general block design (including variable replications and block sizes) as an *E* design for a general *PDC*, Singh and Hinkelmann (1998) used the following model to express the cross effect in terms of the *gca* and *sca* effects and carried out their estimation in two stages. The first stage considers estimation of cross (or the entry) effects by fitting the model:

$$y_{ul} = \mu + \tau_u + \beta_l + \varepsilon_{ul}, \quad (1.12)$$

where y_{ul} is the observation from the u th cross ($u = 1, 2, \dots, n^*$) in block l ($l = 1, 2, \dots, b$), μ is the general mean, τ_u is the effect of the u th cross, and β_l is the effect of the l th block, ε_{ul} are assumed independent and normally distributed with mean zero and variance σ^2 . Denoting by $\boldsymbol{\tau} = (\tau_1, \tau_2, \dots, \tau_{n^*})'$ the vector of cross effects, the least-squares estimate of $\boldsymbol{\tau}$ is obtained from the normal equations

$$E(\mathbf{Q}) = \mathbf{C}\boldsymbol{\tau}, \quad D(\mathbf{Q}) = \sigma^2 \mathbf{C}, \quad (1.13)$$

where

$$\mathbf{C} = \mathbf{r}^\delta - \psi \mathbf{k}^{-\delta} \psi', \quad \mathbf{Q} = \mathbf{T}_c - \psi \mathbf{k}^{-\delta} \mathbf{T}_b,$$

and $E(\cdot)$ and $D(\cdot)$ stand for the expectation and dispersion matrix operator of the random vector in parentheses. Further, $\mathbf{r}^\delta = \text{diag}(r_1, r_2, \dots, r_n^*)$ and $\mathbf{k}^\delta = \text{diag}(k_1, k_2, \dots, k_b)$ represent the diagonal matrices of the numbers of

replications and block sizes. The matrix $\psi = (\psi_{ul})$ is the cross versus block incidence matrix, where ψ_{ul} is the number of times the u th cross occurs in the l th block.

At this stage, the Equation (1.13) can provide the sum of squares due to crosses as $\mathbf{Q}'\mathbf{C}\mathbf{Q}$ on $\text{df} = \text{rank}(\mathbf{C})$.

At the second stage, we express the cross effects in terms of *gca* and *sca*, and to facilitate this, we define the cross and *gca* relation (*CGR*) matrix $\mathbf{Z} = (\mathbf{Z}_{ui})$, where \mathbf{Z}_{ui} is 2 if the u th cross has both parents i , 1 if the u th cross has only one parent i , and 0 otherwise. This results in

$$\boldsymbol{\tau} = \mathbf{Z}\mathbf{g} + \mathbf{s}, \quad (1.14)$$

where $\mathbf{g} = (g_1, g_2, \dots, g_p)'$, $\mathbf{s} = (s_{i_1 j_1}, s_{i_2 j_2}, \dots, s_{i_n j_n})'$ are vectors of *gca* effects (g_i) and *sca* effects (s_{ij}), respectively. Equation (1.13) can then be written as

$$E(\mathbf{Q}) = \mathbf{C} \mathbf{Z}\mathbf{g} + \mathbf{C} \mathbf{s}, D(\mathbf{Q}) = \sigma^2 \mathbf{C}. \quad (1.15)$$

Using the unified theory of least squares (Rao 1973), the estimate of \mathbf{g} is

$$\hat{\mathbf{g}} = (\mathbf{Z}'\mathbf{C}\mathbf{Z})^{-} \mathbf{Z}'\mathbf{Q},$$

with dispersion matrix $D(\hat{\mathbf{g}}) = \sigma^2 (\mathbf{Z}'\mathbf{C}\mathbf{Z})^{-}$ where a *g*-inverse $(\mathbf{Z}'\mathbf{C}\mathbf{Z})^{-}$ of $\mathbf{Z}'\mathbf{C}\mathbf{Z}$ can be computed as

$$(\mathbf{Z}'\mathbf{C}\mathbf{Z})^{-} = (\mathbf{Z}'(\mathbf{r}^\delta - \psi \mathbf{k}^{-\delta} \psi' + n^{-1} \mathbf{r} \mathbf{r}') \mathbf{Z})^{-1},$$

(for details, see Singh and Hinkelmann 1998).

The sum of squares (SS) due to *gca* is given by

$$SS(gca) = \hat{\mathbf{g}}'\mathbf{Z}'\mathbf{C}\mathbf{C}^{-}\mathbf{Q} = \mathbf{Q}'\mathbf{Z}(\mathbf{Z}'\mathbf{C}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{Q},$$

with $\text{df} = \text{rank}(\mathbf{Z}'\mathbf{C}\mathbf{Z})$. The SS due to *sca* equals

$$SS(sca) = SS \text{ due to cross} - SS \text{ due to gca} = \mathbf{Q}'(\mathbf{C}^{-} - \mathbf{Z}(\mathbf{Z}'\mathbf{C}\mathbf{Z})^{-}\mathbf{Z}')\mathbf{Q},$$

with $\text{df} = \text{rank}(\mathbf{C}) - \text{rank}(\mathbf{Z}'\mathbf{C}\mathbf{Z})$. The ANOVA is given in Table 1.17.

Table 1.17 Analysis of Variance for a General M-E Design

Source of Variation	df	Sum of Squares
Blocks	$b - 1$	$b\mathbf{T}_b'\mathbf{k}^{-\delta}\mathbf{T}_b - T_0^2/n$
Crosses (adjusted for blocks)	$\text{rank}(\mathbf{C})$	$\mathbf{Q}'\mathbf{C}\mathbf{Q}$
<i>gca</i>	$\text{rank}(\mathbf{Z}'\mathbf{C}\mathbf{Z})$	$\mathbf{Q}'\mathbf{Z}(\mathbf{Z}'\mathbf{C}\mathbf{Z})^{-}\mathbf{Z}'\mathbf{Q}$
<i>sca</i>	$\text{rank}(\mathbf{C}) - \text{rank}(\mathbf{Z}'\mathbf{C}\mathbf{Z})$	$\mathbf{Q}'(\mathbf{C}^{-} - \mathbf{Z}(\mathbf{Z}'\mathbf{C}\mathbf{Z})^{-}\mathbf{Z}')\mathbf{Q}$
Residual	by subtraction	by subtraction
Total	$n - 1$	$\mathbf{y}'\mathbf{y} - T_0^2/n$

where T_0 = grand total of all the n observations

Table 1.18 ANOVA for PDC in Incomplete Blocks

Source	df	SS	MS	Vr	Prob.
Blocks (unadjusted)	2	45.9			
Cross (adjusted)	42	7079.5	168.6	17.1	1.43×10^{-24}
<i>gca</i>	9	5006.8	556.3	56.5	1.47×10^{-29}
<i>sca</i>	33	2072.7	62.8	6.4	2.09×10^{-11}
Error	74	729.2	9.9		
Total	118	7854.6			

1.6.3 An Example of PDC in Incomplete Blocks

Example 1.7. A subset of data was arbitrarily created from the data in Example 1.2 by dropping the crosses 1×10 , 2×9 , and 4×5 from all the three replications, 1×8 from replications 1 and 2, 9×10 from replication 1, and the crosses 6×9 , 6×10 , 7×8 , and 7×9 from replication 3. The three incomplete blocks 1, 2, and 3 are of sizes 40, 41, and 38, respectively. Table 1.18 gives the ANOVA and Table 1.19 the variance-covariance matrix of the *gca* estimates.

The estimated *gca* values and their standard errors are given in Table 1.20.

1.6.4 Other M-E Designs

Ghosh and Divecha (1997) constructed *PDCs* in incomplete blocks using a *PBIB* design, where one of the λ parameters is equal to zero, and by taking all possible distinct pairs of treatments in each block to form crosses, a scheme similar to Agarwal and Das (1990). Such designs may, however, result in a large number of crosses, for example, use of a group divisible design with parameters $v = 12$, $b = 9$, $r = 3$, $k = 4$, $\lambda_1 = 0$, $\lambda_2 = 1$ requires 54 crosses out of 66 crosses of the *CDC* design. Further, the designs are not efficient as we discuss next. Let \mathbf{N} be the incidence matrix of the original block design, then the line \times block incidence matrix of the diallel cross design will be $(k - 1)\mathbf{N}$, that is, $\mathbf{N}_d = (k - 1)\mathbf{N}$. The information matrix of the original block design is given by

$$\mathbf{C} = \mathbf{R} - \frac{1}{k} \mathbf{NN}',$$

where \mathbf{R} is the diagonal matrix of replication numbers of the original block design and

$$\mathbf{NN}' = \begin{bmatrix} r_1 & \lambda_{12} & \cdots & \lambda_{1v} \\ \lambda_{21} & r_2 & \cdots & \lambda_{2v} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{v1} & \lambda_{v2} & \cdots & r_v \end{bmatrix}.$$

The information matrix Equation (1.11) of the resulting diallel cross block design is then

Table 1.19 Variance–Covariance Matrix for *gca* Estimates

Line	1	2	3	4	5	6	7	8	9	10
1	0.4636	-0.0385	-0.0294	-0.0368	-0.0339	-0.0338	0.0110	-0.0485	0.0074	
2	-0.0385	0.4442	-0.0290	-0.0348	-0.0289	-0.0288	-0.0349	0.0336	-0.0384	
3	-0.0294	-0.0290	0.4062	-0.0279	-0.0247	-0.0254	-0.0279	-0.0359	-0.0113	
4	-0.0368	-0.0348	-0.0279	0.4459	0.0553	-0.0314	-0.0311	-0.0334	-0.0452	-0.0410
5	-0.0368	-0.0348	-0.0279	0.0353	0.4459	-0.0314	-0.0311	-0.0334	-0.0452	-0.0410
6	-0.0339	-0.0289	-0.0247	-0.0314	-0.0314	0.4220	-0.0267	-0.0314	-0.0140	-0.0129
7	-0.0338	-0.0288	-0.0254	-0.0311	-0.0311	-0.0267	0.4224	-0.0097	-0.0150	-0.0363
8	0.0110	-0.0349	-0.0279	-0.0334	-0.0334	-0.0314	-0.0097	0.4412	-0.0440	-0.0361
9	-0.0485	0.0336	-0.0359	-0.0452	-0.0452	-0.0140	-0.0150	-0.0440	0.5062	-0.0233
10	0.0074	-0.0384	-0.0113	-0.0410	-0.0410	-0.0129	-0.0363	-0.0361	-0.0233	0.4784

Table 1.20 The gca Estimates and Standard Errors

Parents	Estimate	Estimated Standard Error
1	-0.43	0.681
2	0.33	0.666
3	-6.11	0.637
4	-4.40	0.668
5	-3.50	0.668
6	-1.50	0.650
7	2.50	0.650
8	8.56	0.664
9	9.56	0.711
10	-3.68	0.692

$$\mathbf{C}_d = \mathbf{G}_d - \frac{1}{k_d} \mathbf{N}_d \mathbf{N}'_d,$$

where $k_d = k(k - 1)/2$. The diagonal elements of \mathbf{G}_d are the replication numbers of the lines and the off diagonal elements are the replication numbers of the crosses, that is,

$$\begin{aligned} \mathbf{G}_d &= \begin{bmatrix} (k-1)r_1 & \lambda_{12} & \dots & \lambda_{1v} \\ \lambda_{21} & (k-1)r_2 & \dots & \lambda_{2v} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{v1} & \lambda_{v2} & \dots & (k-1)r_v \end{bmatrix}. \\ &= \mathbf{N}\mathbf{N}' + (k-2)\mathbf{R}. \end{aligned}$$

Hence, $\mathbf{C}_d = (k-2)\mathbf{C}$. Thus the resulting diallel cross block design may seem to retain the desirable statistical properties of the original block design. However, such diallel cross designs are not efficient as the lines appear 0 or $k - 1$ times in each block, while for optimality (see Section 1.7), the frequencies should be $[2k/p]$ or $[2k/p] + 1$, where $[z]$ denotes the largest integer not exceeding z .

Choi and Gupta (2000) gave a method of constructing variance balanced *CDC* designs using a Greco-Latin square. Their method can be explained as follows. Consider the Greco-Latin square for, say, $p = 5$, where the Latin letters and the Greek letters are both replaced by the symbols 1, 2, ..., p , and the \times symbol is inserted between the two symbols:

$$\begin{array}{ccccc} 1\times 1 & 2\times 2 & 3\times 3 & 4\times 4 & 5\times 5 \\ 2\times 1 & 3\times 4 & 4\times 5 & 5\times 1 & 1\times 2 \\ 3\times 1 & 4\times 1 & 5\times 2 & 1\times 3 & 2\times 4. \\ 4\times 1 & 5\times 3 & 1\times 4 & 2\times 5 & 3\times 1 \\ 5\times 1 & 1\times 5 & 2\times 1 & 3\times 2 & 4\times 3 \end{array}$$

In particular, for p a prime or a prime power, there exists a variance balanced *CDC* design in p blocks, each of size $p - 1$.

Choi, Gupta, and Son (2002), using the pattern of the \mathbf{C} -matrix for estimation of the *gca*, structured partial balancedness in the *PDCs* and defined a partially balanced *PDC* block design. They proved that the existence of a resolvable *PBIB* design implies the existence of a partially balanced *PDC* block design. Their method of construction is explained with the help of the following example.

Example 1.8. Consider the following replication sets of a resolvable group divisible *PBIB* design with parameters $v = 6, b = 12, r = 4, k = 2, \lambda_1 = 0, \lambda_2 = 1$:

Replication set 1: (1, 3), (2, 5), (4, 6) Replication set 2: (1, 4), (2, 6), (3, 5)

Replication set 3: (1, 5), (2, 4), (3, 6) Replication set 4: (1, 6), (2, 3), (4, 5).

The partially balanced *PDC* design in four blocks is then obtained by forming crosses among the two parents within each block and taking the four replication sets as four blocks:

Block 1: $1 \times 3, 2 \times 5, 4 \times 6$ Block 2: $1 \times 4, 2 \times 6, 3 \times 5$

Block 3: $1 \times 5, 2 \times 4, 3 \times 6$ Block 4: $1 \times 6, 2 \times 3, 4 \times 5$.

Choi, Gupta, and Son (2002) constructed *PDC* block designs using two-associate class group divisible and triangular designs, and listed them for $p \leq 24$ along with the two distinct efficiencies of estimation of *gca* comparisons.

1.7 OPTIMALITY

Various optimality criteria and their relationships in the context of experimental designs have been discussed by Kiefer (1958, 1959, 1975) (see also HK2, chapter 1). Optimality of *M-E* designs suitable for estimation of *gca* effects has been discussed by Singh and Hinkelmann (1998). Suppose that an *M-E* design d is connected for *gca* effects, and let \mathbf{L} denote the $p \times (p - 1)$ matrix of coefficients of $p - 1$ orthonormal contrasts of *gca*. Then the variance-covariance matrix of the estimate $\mathbf{L}'\hat{\mathbf{g}}$ is given by

$$\sigma^2 \mathbf{V}_d = \sigma^2 (\mathbf{L}' \mathbf{C}_{dg} \mathbf{L})^{-1},$$

where σ^2 is the error variance in the model effects, and \mathbf{C}_{dg} is the coefficient matrix of the normal equations for estimating *gca* effects using d ; see Equation (1.11). In order to obtain optimal designs, there are three commonly used criteria that require minimization of (1) $\det(\mathbf{V}_d)$ for *D*-optimality, (2) $\text{trace}(\mathbf{V}_d)$ for *A*-optimality, and maximum eigenvalue of \mathbf{V}_d for *E*-optimality. Let $D(p, b, k)$ (or simply D) denote the class of all connected block designs having p lines and b blocks of k crosses each. Kiefer (1975) proved the following result on

universal optimality and hence avoided the search for optimal designs by computing V_D for each design in its class.

Theorem 1.2. A design $d^* \in D$ is universally optimal in D if it satisfies the following:

1. $\mathbf{C}_{d_g^*}$ is completely symmetric; and
2. $\text{trace}(\mathbf{C}_{d_g^*}) = \max_{d \in D} \text{trace}(\mathbf{C}_{dg}).$

The following general result is due to Singh and Hinkelmann (1998).

Theorem 1.3. Consider a class of $M-E$ designs where the mating design M requires that every line is crossed with an equal number of different lines, and the environment design E is equi-replicate, equiblock sized and binary. In this class of designs, a *CDC* embedded in a *BIB* design is universally optimal.

The \mathbf{C}_{dg} of a design of the above theorem is of the form $a\mathbf{I} + b\mathbf{J}$ and hence is universally optimal from Theorem 1.2.

1.7.1 Optimal *CDC* Designs for Estimation of *gca*

We review here the methods of constructing universally optimal block designs for *CDC* (i.e., type IV mating designs of Griffing 1956), where *sca* effects are negligible or merged with random error. Henceforth, by optimal design we mean universally optimal diallel cross block design. As before, the number of lines and the number of crosses will be denoted by p and n_c respectively.

A novel method of constructing diallel cross designs is due to Gupta and Kageyama (1994), and most of the subsequent developments in optimal design are essentially based on their method. They relaxed the previous attention on crosses as treatments in an incomplete block E design to lines as treatments. The lines within each block are then paired to form crosses of an M design. The pairing of the lines within the blocks of E will result in a *CDC* if an irreducible *BIB* design with block size 2 is nested within E in the sense of Preece (1967). Such E designs are called nested balanced incomplete block (*NBIB*) designs with subblock size 2. Gupta and Kageyama (1994) proved that a *NBIB* design with subblock size 2 yields an optimal *CDC* block design. For clarity, we define *NBIB* designs.

Definition 1.2. An arrangement of p treatments each replicated r_1 times in two systems of blocks is said to be an *NBIB* design with parameters $(v, b_1, b_2 = mb_1, r_1, k_1 = mk_2, k_2, \lambda_1, \lambda_2)$ if it satisfies the following conditions:

1. The second system of b_2 blocks is nested within the first system of b_1 blocks such that each block of the first system (block) contains exactly m blocks of the second system (subblocks).

2. The first system (ignoring the second) forms a *BIB* design with b_1 blocks each of k_1 units with λ_1 concurrences.
3. The second system (ignoring the first) forms a *BIB* design in b_2 blocks each of k_2 units with λ_2 concurrences.

The parameters of a *NBIB* design satisfy the following relationships

$$pr_1 = b_1 k_1 = mb_1 k_2 = b_2 k_2, (p-1)\lambda_1 = (k_1 - 1)r_1, (p-1)\lambda_2 = (k_2 - 1)r_1.$$

We now give an example to illustrate the method of Gupta and Kageyama (1994).

Example 1.9. Let the *E* design be a *BIB* with $p = b_1 = 5$, $r_1 = k_1 = 4$, $\lambda_1 = 3$ having blocks $\{1, 2, 3, 4\}$, $\{1, 2, 3, 5\}$, $\{1, 2, 4, 5\}$, $\{1, 3, 4, 5\}$, $\{2, 3, 4, 5\}$. These blocks are further divided into subblocks, with $k_2 = 2$ and $\lambda_2 = 1$ yielding the *NBIB* design $\{(1, 2), (3, 4)\}$, $\{(1, 3), (2, 5)\}$, $\{(1, 5), (2, 4)\}$, $\{(1, 4), (3, 5)\}$, $\{(2, 3), (4, 5)\}$. Then, for instance, the first block $\{(1, 2), (3, 4)\}$ contains two crosses 1×2 and 3×4 , which, in fact, constitute two blocks of the nested design. Thus, this *NBIB* design yields an optimal *CDC* design with $p = b = 5$, $k = 2$, and each cross replicated only once.

In general, a *NBIB* design d with parameters $v = p$, b_1 , b_2 , r_1 , k_1 , $k_2 = 2$, λ_1 , λ_2 yields an optimal design over $D = D(p, b, k)$, the class of all connected *CDC* designs in p lines, b blocks containing k crosses each, with $n_c = p(p-1)/2$ crosses replicated $r = b_2/n_c$ times. The designs are also variance balanced for estimation of *gca* effects. In a variance balanced design, all the elementary contrasts of treatment effects are estimated with the same variance. As Example 1.9 shows, $\lambda_2 = 1$, or equivalently, $b_1 k_1 = p(p-1)$, yields an optimal *CDC*, with each cross replicated only once.

Gupta and Kageyama (1994) obtained two series of optimal designs presented below.

Series 1.1. $D(p = 2t, b = 2t - 1, k = t)$, $t \geq 2$, obtained by developing the following initial block modulo $(2t - 1) \pmod{(2t - 1)}$:

$$\{1 \times p, 2 \times (p-1), \dots, t \times (p-t+1)\},$$

where the parental line p is treated as invariant.

The crosses of Series 1.1 are orthogonally blocked in the sense of Gupta, Das, and Kageyama (1995). This will be further discussed in the context of *PDC*'s considered in the next section. The following example illustrates Series 1.1.

Example 1.10. The initial block of Series 1.1 for $t = 3$, $p = 6$ is given by $\{1 \times 6, 2 \times 5, 3 \times 4\}$. Developing this initial block mod (5) and keeping the symbol

6 invariant yields the optimal design $D(p = 6, b = 5, k = 3)$ with block contents:

$$(1 \times 6, 2 \times 5, 3 \times 4) \quad (2 \times 6, 3 \times 1, 4 \times 5) \quad (3 \times 6, 4 \times 2, 5 \times 1) \\ (4 \times 6, 5 \times 3, 1 \times 2) \quad (5 \times 6, 1 \times 4, 2 \times 3).$$

Series 1.2. $D(p = b = 2t+1, k = t)$ obtained by developing the initial block $\{1 \times (2t), 2 \times (2t-1), \dots, t \times (t+1)\}, \text{mod } (2t+1)$.

The optimal design given earlier in Example 1.9 belongs to Series 1.2.

Dey and Midha (1996) proved that a two-associate class triangular *PBIB* design (see HK2, chapter 4) with parameters $v = p(p-1)/2, b, r, k, \lambda_1, \lambda_2$ and treatments labeled by (i, j) provides a variance balanced *CDC* design when treatment (i, j) is replaced by the cross $i \times j, i < j; i, j = 1, 2, \dots, p$. The information matrix (Eq. 1.11) of this design can be shown to be

$$\mathbf{C}_d = \theta(\mathbf{I}_p - p^{-1}\mathbf{J}_p),$$

where $\theta = pk^{-1}\{r(k-1) - (p-2)\lambda_1\}$. Therefore, any elementary comparison among *gca* effects is estimated with a variance $2\sigma^2/\theta$.

Example 1.11. For $p = 5$, consider the triangular *PBIB* design having parameters $v = 10, b = 15, r = 3, k = 2, \lambda_1 = 0, \lambda_2 = 1$ with block contents $(1, 8) (1, 9) (1, 10) (2, 6) (2, 7) (2, 10) (3, 5) (3, 7) (3, 9) (4, 5) (4, 6) (4, 8) (5, 10) (6, 9)$, and $(7, 8)$. The association scheme is:

Row/Column	1	2	3	4	5
1	*	1	2	3	4
2	1	*	5	6	7
3	2	5	*	8	9
4	3	6	8	*	10
5	4	7	9	10	*

The labeling of treatments by row and column identification is shown in the following table:

Treatment	Label	Treatment	Label
1	(1,2)	6	(2,4)
2	(1,3)	7	(2,5)
3	(1,4)	8	(3,4)
4	(1,5)	9	(3,5)
5	(2,3)	10	(4,5)

Now, replacing the treatment label (i, j) by $i \times j$ yields a variance balanced *CDC* design with parameters $p = 5, b = 15, k = 2$. The resulting block contents are:

$$(1 \times 2, 3 \times 4) \quad (1 \times 2, 3 \times 5) \quad (1 \times 2, 4 \times 5) \quad (1 \times 3, 2 \times 4) \quad (1 \times 3, 2 \times 5) \\ (1 \times 3, 4 \times 5) \quad (1 \times 4, 2 \times 3) \quad (1 \times 4, 2 \times 5) \quad (1 \times 4, 3 \times 5) \quad (1 \times 5, 2 \times 3) \\ (1 \times 5, 2 \times 4) \quad (1 \times 5, 3 \times 4) \quad (2 \times 3, 4 \times 5) \quad (2 \times 4, 3 \times 5) \quad (2 \times 5, 3 \times 4).$$

This design also satisfies the condition of Result 1.1 presented below, and hence is optimal as well.

Further variance balanced *CDC* designs can be obtained using triangular designs listed in Clatworthy (1973). The next two results on optimality are due to Dey and Midha (1996) and Das, Dey, and Dean (1998), respectively. They also listed 21 optimal designs obtained using triangular designs given in Clatworthy (1973). In all the cases, the basic idea is to develop the block design from a set of initial blocks (see HK2, chapter 3).

Result 1.1. A *CDC* design obtained using a triangular design with $\lambda_1 = 0$ is optimal.

Result 1.2. A *CDC* design derived from a triangular design with parameters $v = p(p - 1)/2, b, r, k, \lambda_1, \lambda_2$ is optimal if

$$p(p-1)(p-2)1 = bx\{4k - p(x+1)\},$$

where $x = [2k/p]$, and $[z]$ denotes the largest integer not exceeding z .

The next two series of optimal designs are due to Das, Dey, and Dean (1998).

Series 1.3. For $p = 2t + 1, t \geq 1$, a prime or a prime power, an optimal design $D(p = 2t + 1, b = t(2t + 1), k = 2)$, is obtained by developing the following t initial blocks mod $(2t + 1)$, where x is a primitive element of $\text{GF}(2t + 1)$:

$$\{(0, x^{i-1}), (x^i, x^{i+1})\}; \quad i = 1, 2, \dots, t.$$

Example 1.12. The three initial blocks of Series 1.3 for $t = 3, p = 7$ are given by $\{(0 \times 1), (3 \times 2)\}; \{(0 \times 3), (2 \times 6)\}; \{(0 \times 2), (6 \times 4)\}$. Developing these initial blocks, mod (7) yields the optimal design $D(p = 7, b = 21, k = 2)$ with block contents:

$$(0 \times 1, 2 \times 3) \quad (0 \times 3, 2 \times 6) \quad (0 \times 2, 4 \times 6) \quad (1 \times 2, 3 \times 4) \quad (1 \times 4, 0 \times 3) \quad (1 \times 3, 0 \times 5) \\ (2 \times 3, 4 \times 5) \quad (2 \times 5, 1 \times 4) \quad (2 \times 4, 1 \times 6) \quad (3 \times 4, 5 \times 6) \quad (3 \times 6, 2 \times 5) \quad (3 \times 5, 0 \times 2) \\ (4 \times 5, 0 \times 6) \quad (0 \times 4, 3 \times 6) \quad (4 \times 6, 1 \times 3) \quad (5 \times 6, 0 \times 1) \quad (1 \times 5, 0 \times 4) \quad (0 \times 5, 2 \times 4) \\ (0 \times 6, 1 \times 2) \quad (2 \times 6, 1 \times 5) \quad (1 \times 6, 3 \times 5).$$

Series 1.4. For $12t + 7, t \geq 0$, a prime or a prime power, an optimal design $D(p = 12t + 8, b = (3t + 2)(12t + 7), k = 2)$, is obtained by developing the fol-

lowing $3t + 2$ initial blocks mod $(12t + 7)$, where x is a primitive element of $GF(12t + 7)$ and ∞ is an invariant symbol:

$$\{(1 \times \infty), (x^{3t+2} \times x^{6t+3})\}, \{(x^i \times x^{i+3t+1}), (x^{i+3t+2} \times x^{i+6t+3})\}, i = 1, 2, \dots, 3t + 1.$$

For $t = 0, 1, 2, 3$, optimal designs $D(8, 14, 2)$, $D(20, 95, 2)$, $D(32, 248, 2)$, and $D(44, 473, 2)$, respectively, can be constructed using Series 1.4. Das, Dey, and Dean (1998) presented two more series of optimal designs that were later generalized by Parsad, Gupta, and Srivastava (1999) to the following series.

Series 1.5. For $p = 2ut + 1$, a prime or prime power, $u \geq 2$, $t \geq 1$, an optimal design, $D(p = 2ut + 1, b = t(2ut + 1), k = 2)$ is obtained by developing the following t initial blocks mod (p) , where x is a primitive element of $GF(p)$:

$$\{(x^i \times x^{i+ut}), (x^{i+t} \times x^{i+(u+1)t}), \dots, (x^{i+(u-1)t} \times x^{i+(2u-1)t})\}, i = 0, 1, \dots, t - 1.$$

Taking $u = 1$ above leads to Series 1.2, while optimal designs for $u = 2, 3$ give the other two series of Das, Dey, and Dean (1998).

Example 1.13. The two initial blocks for $u = 2$, $t = 2$, $p = 9$ given by Das, Dey, and Dean (1998) are $\{(1 \times 2), (2x + 1 \times x + 2)\}$, $\{x \times 2x\}$, $\{2x + 2 \times x + 1\}$, where x is a primitive element of $GF(3^2)$, $0, 1, 2, x, x + 1, x + 2, 2x, 2x + 1, 2x + 2$ are the elements of $GF(3^2)$, and $(2x + 1 \times x + 2)$ denotes a cross between the lines coded as $2x + 1$ and $x + 2$, etc. The full optimal design $D(p = 9, b = 18, k = 2)$ obtained by adding successively the nonzero elements of $GF(3^2)$, mod (p) , to the contents of the initial blocks is given below, where the lines have been relabeled 1 through 9 using the correspondence $0 \rightarrow 1, 1 \rightarrow 2, 2 \rightarrow 3, x \rightarrow 4, x + 1 \rightarrow 5, x + 2 \rightarrow 6, 2x \rightarrow 7, 2x + 1 \rightarrow 8, 2x + 2 \rightarrow 9$:

$$\begin{array}{ccccc} (2 \times 3, 6 \times 8) & (1 \times 3, 4 \times 9) & (1 \times 2, 5 \times 7) & (5 \times 6, 2 \times 9) & (4 \times 6, 3 \times 7) \\ (4 \times 5, 1 \times 8) & (8 \times 9, 3 \times 5) & (7 \times 9, 1 \times 6) & (7 \times 8, 2 \times 4) & (4 \times 7, 5 \times 9) \\ (5 \times 8, 6 \times 7) & (6 \times 9, 4 \times 8) & (1 \times 7, 3 \times 8) & (2 \times 8, 1 \times 9) & (3 \times 9, 2 \times 7) \\ (1 \times 4, 2 \times 6) & (2 \times 5, 3 \times 4) & (3 \times 6, 1 \times 5) & & \end{array}$$

Das, Dey, and Dean (1998) also considered a generalization of *NBIB* designs to nested balanced block (*NBB*) designs in the same manner as *BIB* designs have been generalized to balanced block designs. They presented the following series based on *NBB* designs.

Series 1.6. $D(p = 2t + 1, b = t, k = 2t + 1)$, $t \geq 1$, obtained by developing the following initial block mod $(2t + 1)$:

$$\{(j \times 2t + 1 - j), (1 + j \times 1 - j), (2 + j \times 2 - j), \dots, (2t + j \times 2t - j)\}, j = 1, 2, \dots, t$$

The information matrix C_d (see Eq. 1.11) of a design of the above series can be shown to be

$$\mathbf{C}_d = (2t-1)(\mathbf{I}_p - p^{-1}\mathbf{J}_p).$$

The design is thus variance balanced for estimation of *gca* effects. Further, since each line appears in a block two times, which is constant for all the blocks, the design satisfies the orthogonality property of Gupta, Das, and Kageyama (1995) and is thus optimal.

We now present the following result due to Parsad, Gupta, and Srivastava (1999).

Result 1.3. Suppose there exist a *BIB* design with parameters $p, b_0, r_0, k_0, \lambda_0$, and an *NBIB* design with parameters $k_0, b_1, b_2, k_1, k_2 = 2, r_1, \lambda_1, \lambda_2$. Then, writing each of the block contents of the *BIB* design as an *NBIB* design, we get an *NBIB* design with parameters $p, b_0b_1, b_0b_2, k_1, k_2 = 2, r_0r_1, \lambda_0\lambda_1, \lambda_0\lambda_2$, and hence an optimal *CDC* design $D(p, b = b_0b_1, k = k_1/2)$.

The next two series of designs are based on the above result.

Series 1.7. The existence of a *BIB* design $(p, b_0, r_0, k_0 = 2t, \lambda_0)$ and a *NBIB* design $(2t, 2t-1, t(2t-1), r_1 = 2t-1, k_1 = 2t, k_2 = 2, \lambda_1 = 2t-1, \lambda_2 = 1)$ implies the existence of an optimal *CDC* design $D(p, b = b_0(2t-1), k = t)$.

Example 1.14. Consider a *BIB* design $p = 16, b = 20, r = 5, k = 4, \lambda = 1$ and an *NBIB* design with parameter, $4, b_1 = 3, b_2 = 6, k_1 = 4, k_2 = 2, r = 3, \lambda_1 = 3, \lambda_2 = 1$. Then Series 1.7 yields an optimal design $D(16, 60, 2)$.

Series 1.8. The existence of a *BIB* design $(p, b_0, r_0, k_0 = 2t+1, \lambda_0)$ and an *NBIB* design $(2t+1, 2t+1, t(2t+1), r_1 = 2t, k_1 = 2t, k_2 = 2, \lambda_1 = 2t+1, \lambda_2 = 1)$ implies the existence of an optimal *CDC* design $D(p, b = b_0(2t-1), k = t)$.

Example 1.15. Consider a *BIB* design $p = b = 6, r = k = 5, \lambda = 4$ and a corresponding *NBIB* design with parameters $2t+1 = 5, b_1 = 5, b_2 = 10, k_1 = 4, k_2 = 2, r = 4, \lambda_1 = 3, \lambda_2 = 1$. Then Series 1.8 yields an optimal design $D(6, 30, 2)$.

The following series of designs is due to Choi and Gupta (2000).

Series 1.9. $D(p, b = p-1, k = p)$. The i th block of the design is obtained by developing the cross $1 \times i + 1, \text{mod } (p)$, $i = 1, 2, \dots, p-1$.

Example 1.16. For $p = 6$, the above series yields the following design, $D(p = 6, b = 5, k = 6)$, where rows give the block contents, and for $i > j$ cross, $i \times j$ is replaced by $j \times i$:

$$\begin{aligned} &\{1 \times 2, 2 \times 3, 3 \times 4, 4 \times 5, 5 \times 6, 6 \times 1\} \\ &\{1 \times 3, 2 \times 4, 3 \times 5, 4 \times 6, 5 \times 1, 6 \times 2\} \\ &\{1 \times 4, 2 \times 5, 3 \times 6, 4 \times 1, 5 \times 2, 6 \times 3\} \\ &\{1 \times 5, 2 \times 6, 3 \times 1, 4 \times 2, 5 \times 3, 6 \times 4\} \\ &\{1 \times 6, 2 \times 1, 3 \times 2, 4 \times 3, 5 \times 4, 6 \times 5\}. \end{aligned}$$

As noted by Parsad, Gupta, and Gupta (2005), when p is odd, the above series can be improved by taking $i = 1, 2, \dots, (p-1)/2$, resulting in an optimal design $D(p, b = (p-1)/2, k = p)$. Each cross is replicated only once in this design. Choi and Gupta (2000) also studied constructions of optimal designs using *BIB* designs with nested rows and columns (Singh and Dey 1979), and gave several series of optimal designs. Shaffer and Srivastav (2009) gave some further designs based on this method.

Srivastav and Shankar (2007) gave two methods to construct optimal *CDC* designs (i.e., in $n_c = p(p-1)/2$ crosses of the p lines) as *NBIB* designs with subblock size 2 using (1) binary pairwise balanced designs, and (2) strongly equineighbored designs. In a binary pairwise balanced design, every (unordered) pair of treatments occurs in a constant number of blocks. For strongly equineighboured design, see Martin and Eccleston (1991) and Street (1992). We describe the method of construction using the following example based on a binary pairwise balanced design.

Example 1.17. For $p = 7$, a binary pairwise balanced design is shown in Table 1.21. Consider the following trivial *NBIB* design with $p = 4$: $\{(1, 2), (3, 4)\}, \{(1, 3), (2, 4)\}, \{(1, 4), (2, 3)\}$, where there are three complete blocks, and the nested design is with block size 2. The method involves constructing this *NBIB* design for $p = 4$ using the contents of, say, the first block of the pairwise balanced design in Table 1.21. This is done for each of the blocks of the pairwise balanced design. The resulting *CDC* block design is also shown in Table 1.21.

Sharma and Fanta (2010) obtained optimal *CDCs* using two associate class *PBIB* designs such that none of the λ s is zero. These designs were in addition to those of Dey and Midha (1996). Their method requires first constructing an auxiliary design by selecting a *PBIB* design from the tables of Clatworthy (1973), in $v = p$ (treatments or) lines and a two associate class *PBIB* design with parameters $v = b, r = k, \lambda_1, \lambda_2, n_1, n_2, p_{jk}^i (i, j, k = 1, 2)$ such that none of the λ s is equal to zero, and any pair of treatments does not occur more than once in any column of the design, where blocks are the rows and block contents are

Table 1.21 Binary Pairwise Balanced Design and *CDC* Design

Binary Pairwise Balanced Design		<i>CDC</i> Block Design		
Block	Treatments	Block 1	Block 2	Block 3
1	(1, 2, 3, 6)	(1 × 2, 3 × 6)	(1 × 3, 2 × 6)	(1 × 6, 2 × 3)
2	(1, 2, 5, 7)	(1 × 2, 5 × 7)	(1 × 5, 2 × 7)	(1 × 7, 2 × 5)
3	(1, 3, 4, 5)	(1 × 3, 4 × 5)	(1 × 4, 3 × 5)	(1 × 5, 3 × 4)
4	(1, 4, 6, 7)	(1 × 4, 6 × 7)	(1 × 6, 4 × 7)	(1 × 7, 4 × 6)
5	(2, 3, 4, 7)	(2 × 3, 4 × 7)	(2 × 4, 3 × 7)	(2 × 7, 3 × 4)
6	(2, 4, 5, 6)	(2 × 4, 5 × 6)	(2 × 5, 4 × 6)	(2 × 6, 4 × 5)
7	(3, 5, 6, 7)	(3 × 5, 6 × 7)	(3 × 6, 5 × 7)	(3 × 7, 5 × 6)

arranged in k columns. Then take all possible distinct pairs of treatments in each block, starting from the first treatment of the block. This will result into $bk(k - 1)/2$ pairs (crosses), which may be larger than the number of crosses in a *CDC*. The cross from the i th pair formed in a block is allocated to block i of the diallel cross block design, $i = 1, 2, \dots, k(k - 1)/2$. This results in $k(k - 1)/2$ blocks of v crosses each. For $\lambda_1 > \lambda_2$, there will be $n_1(\lambda_1 - \lambda_2)/2$ repeated blocks. The requisite *CDC* block design is obtained by retaining only the distinct blocks. While for $\lambda_2 > \lambda_1$, there will be $n_2(\lambda_2 - \lambda_1)/2$ repeated crosses in some blocks instead of repeated blocks themselves. In this case, the design is obtained by retaining only the distinct crosses in each block. The next two examples illustrate the two situations.

Example 1.18. Consider the design C12 of Clatworthy (1973) with parameters: $v = 5, b = 5, r = 3, k = 3, \lambda_1 = 1, \lambda_2 = 2, n_1 = 2, n_2 = 2$. Table 1.22 shows the *PBIB* design and the derived blocks of crosses.

Since $\lambda_2 = 2$, the crosses formed from the second associate treatments appear in both of the two blocks, *block 2* and *block 3*. Thus, by removing one of these two blocks, *block 3* say, we get the equiblock sized *CDC* in two blocks *block 1* and *block 2*.

Example 1.19. Consider the design R42 of Clatworthy (1973) with parameters: $v = 6, b = 6, r = 3, k = 3, \lambda_1 = 2, \lambda_2 = 1, n_1 = 3, n_2 = 2$. Table 1.23 shows the *PBIB* design and the derived blocks of crosses.

Table 1.22 PBIB Design and the CDC

PBIB Design		Resulting Blocks of Crosses		
Block	Treatments	Block 1	Block 2	Block 3
1	(1, 2, 4)	1 × 2	1 × 4	2 × 4
2	(2, 3, 5)	2 × 3	2 × 5	3 × 5
3	(3, 4, 1)	3 × 4	1 × 3	1 × 4
4	(4, 5, 2)	4 × 5	2 × 4	2 × 5
5	(5, 1, 3)	1 × 5	3 × 5	1 × 3

Table 1.23 PBIB Design and the CDC

PBIB Design		Resulting Blocks of Crosses		
Block	Treatments	Block 1	Block 2	Block 3
1	(1, 2, 4)	1 × 2	1 × 4	2 × 4
2	(2, 3, 5)	2 × 3	2 × 5	3 × 5
3	(3, 4, 6)	3 × 4	3 × 6	4 × 6
4	(4, 5, 1)	4 × 5	1 × 4	1 × 5
5	(5, 6, 2)	5 × 6	2 × 5	2 × 6
6	(6, 1, 3)	1 × 6	3 × 6	1 × 3

Since $\lambda_1 = 2$, the crosses formed from the first associate treatments appear twice in *block* 2. Thus, by retaining only the distinct crosses in this block, we get the following *CDC* in blocks of sizes 2 and 3:

$$\text{Block 1: } \{1 \times 2, 2 \times 3, 3 \times 4, 4 \times 5, 5 \times 6, 1 \times 6\}$$

$$\text{Block 2: } \{1 \times 4, 2 \times 5, 3 \times 6\}$$

$$\text{Block 3: } \{2 \times 4, 3 \times 5, 4 \times 6, 1 \times 5, 2 \times 6, 1 \times 3\}.$$

1.7.2 Optimal Partial Diallel Crosses

For a *PDC* in a *CRD*, the observations are represented with *gca* terms in the model as

$$\mathbf{Y} = \mu \mathbf{J}_n + \Delta'_1 \mathbf{g} + \mathbf{e},$$

where \mathbf{Y} is the $n \times 1$ vector of observed responses, μ is the general mean, \mathbf{g} is the vector of *gca* effects, Δ'_1 is a $n \times p$ matrix such that its (h, t) th element is 1 if the h th observation pertains to the t th line and 0 otherwise. Let $\mathbf{D}_0(p, n)$ denote the class of all *PDC* in a *CRD* with p lines and n crosses. For a design $d_0 \in \mathbf{D}_0(p, n)$, the information matrix of the normal equations for estimating linear functions of *gca* effects under the above model is

$$\mathbf{C}_0 = \mathbf{G}_0 - \frac{1}{n} \mathbf{s}_0 \mathbf{s}'_0, \quad (1.16)$$

where $\mathbf{s}'_0 = (s_{01}, s_{02}, \dots, s_{0p})$, s_{0i} is number of times the i th line appears in d_0 , $\mathbf{G}_0 = (g_{0ij})$ is a symmetric matrix, g_{0ij} is the number of replications of the cross $i \times j$ in d_0 , and $g_{0ii} = s_i$, ($i < j = 1, 2, \dots, p$). Mukerjee (1997) gave the following result for obtaining *E-optimal PDC* in a *CRD*.

Result 1.4. Let $p = mn$, $m \geq 2$ $n \geq 3$, be partitioned into m mutually exclusive and exhaustive classes each of size n . For the i th class, take all possible $\binom{n}{2}$ pairs. Repeat this process for all the classes $t = 1, 2, \dots, m$. In this way, we get $mn(n-1)/2$ pairs. Then the *PDC* obtained by considering these pairs as crosses is *E-optimal* over $D_0(p, n)$. Further, for $n = 3$, the *PDC* is *D-optimal* as well.

Example 1.20. For $p = 12$, $m = 4$, $n = 3$, the arrangement:

$m \downarrow n \rightarrow$	1	2	3
1	1	5	9
2	2	6	10
3	3	7	11
4	4	8	12

yields the *PDC* with crosses 1×5 , 1×9 , 5×9 , 2×6 , 2×10 , 6×10 , 3×7 , 3×11 , 7×11 , 4×8 , 4×12 , and 8×12 , which is *E*- and *D*-optimal in $D_0(12, 12)$.

Mukerjee (1997) also showed that for $m \geq 2$, $n \geq 4$, $p = mn \leq 30$, the *E*-optimal *PDC*'s have high values of *A*- and *D*-efficiencies.

We shall now consider optimal and efficient blocking of *PDCs*. For this, we consider again the model (Eq. 1.8) and the class of diallel cross block designs $D(p, b, k)$ with $n = bk$. Ignoring the blocks, the set of $n = bk$ crosses involved in $d \in D(p, b, k)$ form a diallel cross completely randomized design $d_0 \in D_0(p, bk)$. Thus, to every block design $d \in D(p, b, k)$, there corresponds a *CRD* $d_0 \in D_0(p, bk)$. Following Gupta, Das, and Kageyama (1995), define d_0 to be an orthogonal block design if the i th line appears in every block of the design s_{0i}/b times $i = 1, \dots, p$, that is,

$$\mathbf{N}_d = \frac{1}{b} \mathbf{s}_0 \mathbf{J}'_b,$$

where \mathbf{Nd} is the lines versus blocks incidence matrix of design $d \in D(p, b, k)$, and s_{0i} and \mathbf{s}_0 are as defined before. Using $\mathbf{N}_d \mathbf{J}_b = \mathbf{s}_0$, the information matrix of d can be written as

$$\mathbf{C}_d = \mathbf{C}_0 - \frac{1}{k} \mathbf{N}_d \left[\mathbf{I}_b - \frac{1}{b} \mathbf{J}_b \mathbf{J}'_b \right] \mathbf{N}'_d. \quad (1.17)$$

Thus, $\mathbf{C}_d \leq \mathbf{C}_0$, where for a pair of non-negative definite matrices \mathbf{A} and \mathbf{B} , $\mathbf{A} \leq \mathbf{B}$, implies that $\mathbf{B} - \mathbf{A}$ is non-negative definite. The equality is achieved when $\mathbf{N}_d [\mathbf{I}_b - \frac{1}{b} \mathbf{J}_b \mathbf{J}'_b] \mathbf{N}'_d = \mathbf{O}$, that is, if and only if $\mathbf{N}_d = \frac{1}{b} \mathbf{s}_0 \mathbf{J}'_b$, the condition for an orthogonal block design.

Now consider a nonincreasing optimality criterion ϕ , that is, $\phi(\mathbf{B}) \leq \phi(\mathbf{A})$ whenever $\mathbf{B} - \mathbf{A}$ is non-negative definite. Then $\phi(\mathbf{C}_{d^*}) = \phi(\mathbf{C}_{d_0^*}) \leq \phi(\mathbf{C}_d)$ for any design $d \in D(p, b, k)$ and corresponding $d_0 \in D_0(p, n = bk)$. Thus we have the following theorem.

Theorem 1.4. If a *CRD* $d_0^* \in D_0(p, n = bk)$ is ϕ -optimal and corresponds to an orthogonal block design $d^* \in D(p, b, k)$, then d^* is also ϕ -optimal.

In other words, if a *PDC* in a *CRD* is optimal on the basis of any of the usual optimality criteria, then, the corresponding orthogonally blocked *PDC* is also optimal. Eccleston and Hedayat (1974) introduced the *MS*-optimality criterion that generally provides designs with high values of *A*-efficiency. However, *MS*-optimal designs may not be *A*-, *D*- or *E*-optimal. Das, Dean, and Gupta (1998) gave the following characterization of *MS*-optimal *PDC* in a *CRD*.

Result 1.5. A *PDC* $d_0^* \in D_0(p, n)$ with $\mathbf{s}_0 = s \mathbf{J}_p$ and $s = 2n/p$, an integer, is *MS*-optimal if $g_{d_0^*ij}$, the number of times cross $i \times j$ appears in d_0^* , satisfies

$|g_{g_{0ij}^*} - s/p - 1| < 1$ for $i \neq j = 1, \dots, p$, that is, $g_{d_{0ij}^*} = \lambda$ or $\lambda + 1$, where $\lambda = [s/(p-1)]$ and $[z]$ is the integer portion of z .

Using Result 1.5, the *PDC* designs, in which every line appears $s = 2n/p$ times and in which each cross appears either $\lambda = s/(p-1)$ or $\lambda + 1$ times, are *MS-optimal*. A method of constructing *MS-optimal* designs can be described as follows:

1. p even: For any integer $t \geq 2$ and $p = 2t$ lines, construct the following set of $p-1$ blocks $B_j, j = 1, 2, \dots, p-1$, each of size $k = p/2$:

$$B_j : \{(j \times 2t - 3 + j), (1 + j \times 2t - 4 + j), \dots, (t - 2 + j \times t - 1 + j), (j - 1 \times \infty)\}.$$

In each block, the symbols are reduced $\text{mod } (p-1)$, and ∞ is an invariant symbol. An *MS-optimal* design $d^* \in D(p, b, k)$, with $p = 2t, t \geq 2, k = p/2, b < p-1$, is obtained by selecting any b distinct blocks. Such a design has $n = pb/2$ with $s = b$.

2. p odd: For any integer $t \geq 1$ and $p = 2t+1$ lines, construct the following $t = (p-1)/2$ blocks $B_j, j = 1, 2, \dots, t$, each of size $k = p$:

$B_j : \{(j \times 2t + 1 - j), (1 + j \times 1 - j), (2 + j \times 2 - j), \dots, (2t - 1 + j \times 2t - 1 - j), (2t + j \times 2t - j)\}$. In each block, the symbols are reduced $\text{mod } p = (2t+1)$. An *MS-optimal* design $d^* \in D(p, b, k)$ with $p = k = 2t+1, b < t, t \geq 1$ is obtained by selecting any b of the t blocks. Such a design has $n = ps/2$ with $s = 2b$.

Example 1.21. For $p = 8$ lines, $t = 4$, a design with $b = 4$ blocks of $k = 4$ crosses can be obtained by selecting any four blocks, say the blocks $j = 1, 2, 3, 5$. Leaving the symbol ∞ fixed and reducing all other symbols $\text{mod } (p-1) = \text{mod } (7)$, we get the design:

$$\text{Block 1: } (1 \times 6, 2 \times 5, 3 \times 4, 0 \times \infty)$$

$$\text{Block 2: } (2 \times 0, 3 \times 6, 4 \times 5, 1 \times \infty)$$

$$\text{Block 3: } (3 \times 1, 4 \times 0, 5 \times 6, 2 \times \infty)$$

$$\text{Block 4: } (5 \times 3, 6 \times 2, 0 \times 1, 4 \times \infty).$$

This design is *MS-optimal* in $D(8, 4, 4)$. An *MS-optimal* design in $D(8, 11, 4)$ with $b = 11$ blocks of size $k = 4$ can be obtained by appending the full set of $p-1 = 7$ blocks to the above design.

For tables of such *MS-optimal* designs, see Das, Dean, and Gupta (1998) and Parsad, Gupta, and Gupta (2005). The latter authors also showed that singular or semi-regular group divisible designs provide disconnected *PDCs*. Das, Dean, and Gupta (1998) also considered orthogonal blocking of *MS-optimal PDCs* in *CRD* based on Theorem 1.4.

Mathur and Narain (1976) obtained A -efficient PDC in a CRD through a computer search, which in view of Result 1.5 are also MS -optimal. They generated 31 A -efficient plans for $5 \leq p \leq 18$, tabulated in Narain (1990), for which lower bounds to A - and D -efficiencies can be found in Parsad, Gupta, and Gupta (2005). Some of these designs are better with respect to A -efficiency than those listed by Das, Dean, and Gupta (1998). Singh and Hinkelmann (1995) provided designs with high A -efficiency that are also MS -optimal.

Chai and Mukerjee (1999) extended the concept of optimality of a PDC for estimating $gcas$ to include estimation of sca s as well using $PBIB$ designs with triangular association schemes. Das and Dey (2005) obtained existence conditions for optimal designs for estimation of gca and sca .

Choi et al. (2002) considered orthogonally blocked $CDCs$ for estimating gca using a model that includes sca , and proved that these designs remain optimal even in the presence of sca . They reported three new series of orthogonally blocked optimal designs based on the following theorem.

Theorem 1.5. A binary diallel cross design with parameters $v = p(p - 1)/2$, r_c , b , k such that it is orthogonally blocked with respect to the lines, such that each line occurs in every block $r = b$ times where $r = r_c(p - 1)$, and r_c is the number of times each cross is repeated in the design, is optimal for the estimation of gca comparisons even in the presence of sca effects. Here, a *binary* design means that a cross appears, at most, once in a block.

Das (2003) extended the concept of designs for control versus test treatment (see HK2, chapter 6) in the diallel cross settings and obtained A - and MV -optimal designs for comparing the gca of a standard line (could be a well or widely adapted line) with the $gcas$ of test lines and tabulated 322 designs in the range of $3 \leq p \leq 30$ along with their efficiency factors. Here, an A -optimal (MV -optimal) design minimizes the sum (maximum) of variances of estimates of the difference between gca 's of a test line and the control line. An A - and MV -optimal design for $p = 5$ test lines and 0 being the control line is given below, where each row represents a block:

$$\begin{aligned} & \{1 \times 4, 2 \times 5, 3 \times 1, 4 \times 2, 5 \times 0, 3 \times 0\} \\ & \{2 \times 3, 3 \times 4, 4 \times 5, 5 \times 1, 1 \times 0, 2 \times 0\} \\ & \{2 \times 5, 3 \times 1, 4 \times 2, 5 \times 3, 1 \times 0, 4 \times 0\} \\ & \{3 \times 4, 4 \times 5, 5 \times 1, 1 \times 2, 2 \times 0, 3 \times 0\} \\ & \{5 \times 3, 1 \times 2, 1 \times 4, 2 \times 3, 4 \times 0, 5 \times 0\}. \end{aligned}$$

Choi, Gupta, and Kageyama (2004) applied the concept of supplemented balance (Pearce 1960) to diallel experiments with two different precisions for comparing gca : one for comparing a noncontrol line with a control line, and the other for comparing two noncontrol lines. Expressions for the variances

for these two types of comparisons of *gcas* have been obtained for unblocked and blocked situations and tabulated for various designs for lines $p \leq 25$. They also constructed such designs and called them *S*-type designs. The following is an example for $p = 6$ lines in 10 blocks of size 2, where line 6 is the control:

$$(1 \times 6, 2 \times 4), (1 \times 6, 3 \times 4), (1 \times 3, 2 \times 6), (2 \times 6, 4 \times 5), (1 \times 2, 4 \times 6), \\ (1 \times 5, 4 \times 6), (2 \times 4, 3 \times 6), (1 \times 5, 3 \times 6), (1 \times 3, 5 \times 6), (2 \times 3, 5 \times 6).$$

Das, Gupta, and Kageyama (2006) gave a sufficient condition for *A*-optimality of *control* line versus *test* line comparisons. They also characterized a class of *S*-type designs, provided a lower bound to *A*-efficiency, and showed that these *S*-type designs are highly efficient for *test* versus *control* comparisons.

Ghosh and Das (2005) presented the problem of predicting the yielding capacity of a cross arising from randomly chosen inbred lines (Curnow 1963) using the best linear unbiased predictor, and characterized *A*-optimal *CDC* designs and some efficient *PDC* designs.

1.7.3 Estimation of Heritability

In some cases, parental lines are randomly selected from a population of lines resulting in a random effects model. A diallel cross from a random sample of inbred lines provide information on heritability, a key determinant for the response to selection. Ghosh and Das (2004) and Ghosh, Das, and Midha (2005) considered the case where specific combining ability effects are negligible. With an aim to obtain optimal designs for the estimation of heritability, they obtained an estimator of the ratio of the variance components using the optimality criterion, which minimizes the sum of the variances of the estimators of the variance components, or the variance of the ratio of the estimators of the variance components. The analysis is carried out under the random effects model

$$Y_{ijl} = \mu + g_i + g_j + e_{ijl}, \quad i < j,$$

where Y_{ijl} is the observation on the l -th replication of the cross (i, j) , g_i is the i -th line effect with $E(g_i) = 0$, $\text{Var}(g_i) = \sigma_g^2 \geq 0$, $\text{Cov}(g_i, g_j) = 0$, and e_{ijl} is the random error, uncorrelated with g_i , with expectation zero and variance $\sigma_e^2 > 0$, $1 \leq i < j \leq p$. Here μ , σ_e^2 , and σ_g^2 are unknown parameters.

The *heritability* in the narrow sense, h_n^2 , is defined as $h_n^2 = 4\sigma_g^2 / (2\sigma_g^2 + \sigma_e^2)$. An estimator for h_n^2 is $4T/2T + 1$, where T is an unbiased estimator of σ_g^2/σ_e^2 . Ghosh, Das, and Midha (2005) provide an unbiased estimator T , which leads to an asymptotically unbiased estimator for h_n^2 , and give an expression for the variance of T . The following theorem by Ghosh and Das (2005) forms the basis of identifying optimal designs for the estimation of heritability.

Theorem 1.6. For a design d with p lines, b blocks each of size k , an unbiased estimator for σ_g^2/σ_e^2 is $T = [(n - b - p - 1)(SSL/SSE) - p + 1]/\text{tr}(C_d)$, with variance

$$V(T; d, \sigma_g^2, \sigma_e^2) = \alpha \left\{ (n - b - p - 1)\sigma_g^4 \frac{\text{tr}(C_d^2)}{\text{tr}^2(C_d)} + 2t\sigma_e^2\sigma_g^2 \frac{1}{\text{tr}(C_d)} \right. \\ \left. + t(p - 1)\sigma_e^4 \frac{1}{\text{tr}^2(C_d)} + \sigma_g^4 \right\},$$

where SSL = sum of squares due to lines, SSE = the sum of squares due to error, C_d = coefficient matrix C_g of the normal equations for *gca* effects for design d (see Eq. 1.11), $\alpha = 2/(n - b - p - 3)\sigma_e^4$, and $t = n - b - 2$.

The results for estimation of σ_g^2/σ_e^2 and the corresponding variance expression under unblocked diallel cross experiments can be obtained as a special case of the above results by taking the number of blocks as one. For example, the unbiased estimator of σ_g^2/σ_e^2 under an unblocked model, using a design d_0 with p lines and n crosses, reduces to $(n - p - 2)(SSL/SSE) - p + 1/\text{tr}(C_{0d_0})$, where

$$C_{0d_0} = \mathbf{G} - \frac{1}{n} \mathbf{s}\mathbf{s}',$$

is the C_g -matrix of d_0 , $\mathbf{G} = (G_{ij})$ = number of replications s_i of line i when $i = j$, and the number of times the cross $i \times j$ appears in the design when $i \neq j$ (see Section 1.7.2).

Example 1.24. The following design with blocks as rows is optimal in the class of designs with ($p = 8, b = 4, k = 4$):

$$(1 \times 6, 2 \times 5, 3 \times 4, 0 \times 7) \\ (2 \times 0, 3 \times 6, 4 \times 5, 1 \times 7) \\ (3 \times 1, 4 \times 0, 5 \times 6, 2 \times 7) \\ (5 \times 3, 6 \times 2, 0 \times 1, 4 \times 7).$$

Ghosh and Das (2004) constructed optimal designs for estimating confidence interval for heritability, by introducing an *L*-optimality criterion, which minimizes the maximum expected normalized length of a set of confidence intervals. They found that the existence of a nested *BIB* block design with parameters $v = p, b_1 = b, b_2 = bk, k_1 = 2k, k_2 = 2$ implies the existence of an *L*-optimal design.

Example 1.25. The following diallel design, with rows as blocks, is *L*-optimal in the class of designs with $p = 12, b = 3, k = 6$:

$$(1 \times 2, 3 \times 4, 5 \times 6, 7 \times 8, 9 \times 10, 11 \times 12)$$

$$(1 \times 3, 2 \times 4, 5 \times 7, 6 \times 8, 9 \times 11, 10 \times 12)$$

$$(1 \times 4, 2 \times 3, 5 \times 8, 6 \times 7, 9 \times 12, 10 \times 11).$$

1.8 ROBUSTNESS

Even in a well-planned experiment, other stages of conduct of the experiment may unavoidably result in missing observations. The remaining observations may not necessarily provide the statistical properties of optimality and estimation of parameters as planned. Wherever possible, it is thus desirable to choose experimental designs so that the properties of the estimates may be least affected due to missing observations, meaning that the design chosen shows a certain degree of *robustness*. The issue of robustness of experimental design for missing observations has been dealt with in Baksalary and Tabis (1987), Dey and Dhall (1988), Das and Kageyama (1992), Ghosh (1982), Ghosh, Rao, and Singh (1983), Hedayat and John (1974), Mukerjee and Kageyama (1990), Srivastava, Gupta, and Dey (1990), and Mansson and Prescott (2002). Often-used criteria for evaluating robustness of an experimental design include connectedness, variance balancedness, and *A*-efficiency of the residual design. For example, Mukerjee and Kageyama (1990) and Srivastava, Gupta, and Dey (1990) have assessed robustness of *BIB* designs, group divisible designs, and Youden designs in terms of *A*-efficiency when one block is missing (see also HK2, section 2.8).

The missing observation issue is of crucial importance in diallel experiments, as a missing cross reduces information on *gcas* of the two lines involved in the cross. From the diallel crosses perspective, the effect of a missing observation, for example, a cross or even a whole block, has been addressed by several authors. Ghosh and Desai (1998) evaluated robustness of a *CDC* plan constructed from an *BIB* design for the loss of a block. Let the parameters of the *BIB* design be $v = p, b, r, k (>2)$, and λ , where $2 \leq k \leq 15$ and $r \leq 15$. If one block of such a design is lost, then the *A*-efficiency, which is defined as the ratio of the average variance of pairwise contrasts of the *gca* effects under the *CDC* design (i.e., no loss of observations) to that under the residual design, is given by

$$E = 1 - \frac{k(k-1)}{\lambda v(v-1) - k(v-k)}.$$

Thus, the reduction in the efficiency due to the loss of a block is

$$1 - E = \frac{k(k-1)}{\lambda v(v-1) - k(v-k)}.$$

For the *CDC* derived from a *BIB* design with $p = v = 13, b = 13, r = 4, k = 3, \lambda = 1$ having an efficiency of 1.0 (fully efficient) for the estimation of *gca*, the E is 0.9 for the residual design. Thus, losing 1 of the 13 blocks results in reduction of efficiency from 1.0 to 0.9, and the residual design can, therefore, be defended as being robust in terms of efficiency. Further, for the 143 fully efficient designs listed for $7 \leq p \leq 211$, where $2 \leq k \leq 15$ and $r \leq 15$, E exceeded 0.9 in 138 out of 143 designs, and there were five cases where E was in the range 0.8–0.9, thus the *CDC* plans are fairly robust to the unavailability of any block.

Ghosh and Desai (1999) introduced a *CDC* with unequal replications. Consider the situation of a *CDC* with $p(p - 1)/2$ crosses, of which $p(n - 1)/2$ crosses (integer $n < p$) are repeated λ_1 times, and the remaining $p(p - n)/2$ crosses are repeated λ_2 times. Let such an M design be evaluated in an incomplete block design yielding two distinct variances for comparing *gca* effects. This is a case of two different numbers of replications. They have listed 124 designs with efficiency factors of the residual designs. It is worth noting that in cases where the lost block represents less than 10% of the total blocks, the reduction in efficiency is also less than 10% in the majority of cases.

Ghosh and Biswas (2000) showed that the optimal designs developed by Dey and Midha (1996) using triangular *PBIB* designs with two associate classes are fairly robust against the loss of one block. Of the 12 designs listed for p in the range 5–10, 10 designs have efficiency exceeding 0.95, and two have 0.91 and 0.92.

Binary balanced block designs have been shown to be robust under the following situations: one missing observations or one missing block (Dey, Srivastava, and Parsad 2001); exchange of a cross (Panda, Parsad, and Sharma 2003); interchange of a pair of crosses (Panda, Parsad, and Sharma 2004); and presence of an outlier (Sarkar, Parsad, and Gupta 2005).

Prescott and Manson (2004) have examined the efficiency of the resulting design when one or more observations scattered throughout the layout, or one or two crosses are lost from a *CDC* or a *PDC* design. For the selected *CDC* and *PDC* class of designs, they have given expression for the variances of difference between the *gca* effects of: (1) the two lines in the missing cross, or two lines in the affected block but not part of the missing cross, or two lines not in the affected block (denoted by V_1); (2) one of the lines in the missing cross and a line in the affected block (V_2); (3) a line in the affected block (but not in the missing cross) and a line not in the block (V_3); and (4) a line in the missing cross and a line not in the affected block (V_4). Conclusions have been drawn on the basis of several sets with diverse design parameters, and they noted that for small designs or for large number of lines, the efficiency for individual line comparisons can reduce substantially, stressing that greater care should be taken to avoid the possible loss of observations.

Example 1.26. Consider the following design for 13 lines generated from a *BIB* with 13 treatments in 13 blocks of six crosses per block (Ghosh and Desai 1998).

Block	Crosses					
1	1×2	1×4	1×10	2×4	2×10	4×10
2	2×3	2×5	2×11	3×5	3×11	5×11
3	3×4	3×6	3×12	4×6	4×12	6×12
4	4×5	4×7	4×13	5×7	5×13	7×13
5	5×6	5×8	1×5	6×8	1×6	1×8
6	6×7	6×9	2×6	7×9	2×7	2×9
7	7×8	7×10	3×7	8×10	3×8	3×10
8	8×9	8×11	4×8	9×11	4×9	4×11
9	9×10	9×12	5×9	10×12	5×10	5×12
10	10×11	10×13	6×10	11×13	6×11	6×13
11	11×12	1×11	7×11	1×12	7×12	1×7
12	12×13	2×12	8×12	2×13	8×13	2×8
13	1×13	3×13	9×13	1×3	1×9	3×9

For this design, if any of the crosses is lost from one of the blocks, Prescott and Manson (2004) computed the four variances for the differences between pairs of line effects to be $V_1 = 0.30772$, $V_2 = 0.34252$, $V_3 = 0.31642$ and $V_4 = 0.31642$. The loss of efficiency is highest, at just over 10%, for the comparison of one of the lines in the missing cross and another line from the affected block. For the designs listed by Ghosh and Desai (1998), these four types of variances and their efficiencies are presented. The highest loss of efficiency was found when an observation was lost from relatively small designs, and the efficiencies were in excess of 90% for most cases.

1.9 THREE- OR HIGHER-WAY CROSSES

1.9.1 Triallel or Three-Way Crosses

Triallel mating designs are useful in estimating the additive and dominance genetic variances, but under the absence of epistatic effects. If epistatic effects are present, the variance components due to interaction between additive and additive (σ_{AA}^2), additive and dominance (σ_{AD}^2), and dominance and dominance (σ_{DD}^2) effects can be estimated by raising triallel crosses or three-way and double crosses or four-way crosses introduced by Rawlings and Cockerham (1954a, 1954b). To understand three-way crosses, let there be two distinct lines A and B , and let AB be their F_1 hybrid, and C be an unrelated line. Then a cross between AB and C is a three-way hybrid or three-way cross, denoted by $(AB)C$. Here, lines A and B are grandparents or half-parents, and C is the parent of the three-way cross. If there are p lines in the experiment, then in the absence of reciprocal effects and maternal-paternal interactions (i.e., $(AB)C$ and $(BA)C$ are the same three-way cross), there will be $p(p - 1)(p - 2)/2$ three-way crosses for a complete triallel crosses. For complete triallel

crosses evaluated in complete blocks, Rawlings and Cockerham (1962a) presented a linear model leading to an orthogonal analysis of variance (see also the double hybrid crosses in Section 1.9.2), and expectation of various mean squares in terms of statistical variance components, relationships between of genetic and statistical variance components, and presented tests for a number of genetic hypotheses.

However, the number of all possible three-way crosses would be unmanageable even for a moderate number of lines p . For $p = 10$ it will be required to develop 360 triallel crosses. To handle this situation, Hinkelmann (1965) proposed partial triallel crosses (*PTC*) and constructed a number of *PTCs* using their relationship with generalized partially balanced incomplete block design (Shah 1959), and method of their analysis.

Example 1.27. An example of *PTC* with $p = 10$, with 1/12th of the possible crosses, is:

(8, 9)1	(5, 8)4	(3, 8)7
(8, 10)1	(6, 8)4	(1, 4)8
(9, 10)1	(3, 4)5	(1, 7)8
(6, 7)2	(3, 10)5	(4, 7)8
(6, 10)2	(4, 10)5	(1, 3)9
(7, 10)2	(2, 4)6	(1, 6)9
(5, 7)3	(2, 9)6	(3, 6)9
(5, 9)3	(4, 9)6	(1, 2)10
(7, 9)3	(2, 3)7	(1, 5)10
(5, 6)4	(2, 8)7	(2, 5)10

For the analysis, a model with the following parametrization, which is different from that of Rawlings and Cockerham (1962a), is

$$y_{(ij)k,l} = \mu + h_i + h_j + g_k + d_{ij} + s_{(i)k} + s_{(j)k} + t_{(ij)k} + r_l + e_{(ij)k,l},$$

where $y_{(ij)k,l}$ is response from triple cross $(ij)k$ from the block l , μ is the general mean, g_i general effect of line i as full parent (or general effect of first kind), h_i are general effect of line i as half parent or general effect of second kind, ds and ss are two-line specific effects, and t 's are three-line specific effects. In case of *PTC*, the above model is further simplified by retaining only single line effects as follows for an *RCBD* situation,

$$y_{(ij)k,l} = \mu + h_i + h_j + g_k + r_l + e_{(ij)k,l},$$

where r_l is the effect of l -th block, and $e_{(ij)k,l}$ s are assumed independently normally distributed with mean zero and variance σ^2 . The above parametrization does not permit an orthogonal partitioning of the between crosses sum of

squares. In case of a random effects model, Hinkelmann (1965) has shown that a number of covariances between various order line effects do exist, and covariances among relatives for three-way crosses in terms of statistical and genetic variances and covariances are available in Hinkelmann (1975), which can be used for the estimation of various components of variances of genetic effects and their interactions. Ponnuswamy, Das, and Handoo (1974) provided the analysis for $p = 5$ maize lines data where detailed expressions of variance component estimates and their standard errors are available. Srinivasan and Ponnuswamy (1993) presented a systematic mathematical approach for estimation of variance components using an alternative model.

1.9.2 Double- or Four-Way Crosses

A double cross is a cross between two unrelated F_1 hybrids, say denoted by $(ij)(kl)$, where i, j, k , and l are denoting the grandparents and no two of them are the same. Ignoring reciprocal crosses, with p grandparents, there will be $3\binom{p}{4}$ double or four-way crosses. With double hybrid crosses, one can obtain information on epistatic interactions that are either assumed absent or ignored in single cross hybrids analysis. Rawlings and Cockerham (1962b) have introduced the double crosses and presented an orthogonal ANOVA from a complete set of doubled hybrid crosses evaluated in an *RCBD*, expressing the contributions of two-line, three-line, and four-line interaction for lines appearing together irrespective of arrangements and interactions due to specific arrangements. Relationships between the statistical variances of the terms in the linear model and genetic variances arising due to covariances between the hybrid relatives are given. Estimates of statistical variances are presented in terms of the ANOVA mean squares. Here, the model used is described, and tests for selected genetic variance components are mentioned. With p grandparent lines, the following model is used for analysis:

$$y_{(ij)(kl)m} = \mu + \rho_m + \gamma_{(ij)(kl)} + \varepsilon_{(ij)(kl)m},$$

where $y_{(ij)(kl)m}$ is the observation on double cross $(ij)(kl)$ from replication m , $m = 1, 2, \dots, r$; $i, j, k, l = 1, 2, \dots, p$; i, j, k , and l are all distinct, μ is the general mean, ρ_m is the effect of replication m , $\gamma_{(ij)(kl)}$ is genotypic effect of the double cross hybrid $(ij)(jk)$, and $\varepsilon_{(ij)(kl)m}$ s are random errors assumed independent and normally distributed with mean zero and variance σ_e^2 . The genotypic effect $\gamma_{(ij)(kl)}$ can be expressed as a linear function of random and uncorrelated effects and interactions due to the lines as follows:

$$\begin{aligned} \gamma_{(ij)(kl)} &= (g_i + g_j + g_k + g_l) \\ &+ (s_{2ij} + s_{2ik} + s_{2il} + s_{2jk} + s_{2jl} + s_{2kl}) + (s_{3ijk} + s_{3ijl} + s_{3ikl} + s_{3jkl}) + (s_{4ijkl}) \\ &+ (t_{2ij} + t_{2kl} + t_{2i,k} + t_{2i,l} + t_{2j,k} + t_{2j,l}) + (t_{3ij,k} + t_{3ij,l} + t_{3kl,i} + t_{3kl,j}) + (t_{4ijkl}). \end{aligned} \quad (1.18)$$

where g_i is the average effect of line i , s_{2ij} is the two-line interaction effect of lines i and j appearing together irrespective of arrangement, s_{3ijk} is the three-line interaction effect of lines i, j , and k appearing together irrespective of the arrangement, s_{4ijkl} is the four-line interaction effect of lines i, j, k , and l appearing together irrespective of the arrangement, t_{2ij} is the two-line interaction effect of lines i and j due to the particular arrangement $(ij)(--)$ where a dash indicates the lines not common with the members of each pair in the double hybrid cross, $t_{2i,j}$ is the two-line interaction effect of lines i and j due to the particular arrangement $(i-)(j-)$, $t_{3ij,k}$ is the three-line interaction effect of lines i, j , and k due to the particular arrangement $(ij)(k-)$, and $t_{4ij,kl}$ is the four-line interaction effect of lines i, j, k , and l due to the particular arrangement $(ij)(kl)$. The effects g_i , called general effects, are random variables assumed independent and follow $N(0, \sigma_g^2)$, and effects s_{2ij}, s_{3ijk} , and s_{4ijkl} , called specific effects, are random variables assumed independent and independent of g_i 's and follow $N(0, \sigma_{s_2}^2), N(0, \sigma_{s_3}^2)$, and $N(0, \sigma_{s_4}^2)$, respectively. The effects $t_{2ij}, t_{2i,j}, t_{3ij,k}$, and $t_{4ij,kl}$, called arrangement effects, are such that their sums are zeros across all arrangements for each combination of lines, for example, $t_{2ij} + 2t_{2i,j} = 0$, $t_{3ij,k} + t_{3ik,j} + t_{3jk,i} = 0$, and $t_{4ij,kl} + t_{4ik,jl} + t_{4il,jk} = 0$. The expected values of the squares and products of the arrangement effects are expressed as

$$\begin{aligned} E[t_{2ij}^2] &= 4E[t_{2i,j}^2] = -2E[t_{2ij}t_{2i,j}] = (4/9)\sigma_{t_2}^2 \\ E[t_{3ij,k}^2] &= -2E[t_{3ij,k}t_{3ik,j}] = (2/3)\sigma_{t_3}^2 \\ E[t_{4ij,kl}^2] &= -2E[t_{4ij,kl}t_{4ik,jl}] = (2/3)\sigma_{t_4}^2. \end{aligned}$$

Rawlings and Cockerham (1962b) estimated the statistical variance components $\sigma_g^2, \sigma_{s_2}^2, \sigma_{s_3}^2, \sigma_{s_4}^2, \sigma_{t_2}^2, \sigma_{t_3}^2$, and $\sigma_{t_4}^2$ in terms of the mean squares from the ANOVA on the observations. The covariances between relatives can be expressed in terms of genetic variance components required for the interpretation of the nature of gene action, and under the necessary conditions are given by

$$\Sigma_{a,d} \alpha^a \delta^d \sigma_{ad}^2 = \alpha \sigma_A^2 + \delta \sigma_D^2 + \alpha^2 \sigma_{AA}^2 + \alpha \delta \sigma_{AD}^2 + \delta^2 \sigma_{DD}^2 + \dots$$

where σ_{ad}^2 is the genetic variance component for additive effects entering a times and dominance effects entering d times. Thus, $\sigma_A^2, \sigma_D^2, \sigma_{AA}^2, \sigma_{AD}^2$, and σ_{DD}^2 are components of genetic variance due to additive effects, dominance effects, additive \times additive interaction, additive \times dominance interaction, and dominance \times dominance interaction, respectively. The coefficients α and δ are correlation between additive deviations and correlation between dominance deviations of the two relatives. Assuming that all the lines have the same coefficient of inbreeding, Rawlings and Cockerham (1962b) have tabulated the relationships between genetic and statistical variance components, and

derived F distribution based tests for the seven composite hypotheses on genetic variances. The test for a composite hypothesis of genetic variances was obtained as a simple hypothesis on a statistical variance component. For example, the null hypotheses $\sigma_g^2 = 0$ tests the composite null hypothesis that $\sigma_A^2 = \sigma_{AA}^2 = \sigma_{AAA}^2 = \dots = 0$, that is, the additive genetic variance and epistatic genetic variances of all-additive types are zero, and thus reflecting the absence of additive effects and any interactions of all-additive types. A test of null hypothesis $\sigma_{s_2}^2 = 0$ is a test of a strictly additive model, that is, except σ_A^2 , all the other components of genetic variance are zero. Presence of additive \times dominance epistatic variance or any of the higher-factor epistatic variance components except all-dominance types can be tested by the test for $\sigma_{s_3}^2 = 0$. Another form of the model (Eq. 1.18), as well as a reduced model, are discussed in Hinkelmann (1975).

1.10 COMPUTATION

Analysis of data from diallel experiments can be carried out using software which have tools for fitting linear models and variance components. The list of statistical software is limited for analysis of diallel experiments. There are specifically written programs in statistical software environments, such as *Genstat*[®], *SAS*[®] and *R*-language. *Genstat* has a library of procedures for carrying out statistical analysis of data from Griffing's and Hayman's approaches (Payne 2009). Singh and Chaudhary (1979) have presented formulae and clear step-by-step computations of the analysis of the most commonly used genetic experiments. These include formulae for covariances among relatives, diallel analyses—Hayman's approach and Griffing's approach, partial diallel, three-way cross analysis, analysis of double cross hybrids, line \times tester analysis, North Carolina designs analyses. Several *SAS* codes for the analyses dealing with a number of diallel experiments conducted in multienvironments can be found in the articles published in Kang (2003). The computations for the statistical analyses shown in Sections 1.4 and 1.6 were carried out using the *R*-codes, which are available on the John Wiley website (URL: ftp://ftp.wiley.com/public/sci_tech_med/special_designs).

ACKNOWLEDGMENTS

The authors are thankful to Ms Suhaila Arslan, Assistant Manager, International Nursery, Biodiversity and Integrated Germplasm Management Program (BIGM), ICARDA, Syria for providing the data from her diallel experiment, Dr. R.S. Malhotra, Senior Chickpea Breeder and Dr S. Udupa, BIGM, ICARDA for reviewing an earlier version of the manuscript, and Dr Ashish Das, Indian Institute of Technology, Mumbai, India, for help with Section 1.7.3.

REFERENCES

- Agarwal, S.C. and M.N. Das (1990). Use of n-ary block designs in diallel crosses evaluation. *J. Appl. Stat.*, **17**, 125–131.
- Arunachalam, V. (1974). The fallacy behind the use of a modified line x tester design. *Indian J. Genet. Plant Breed.*, **34**, 280–287.
- Arya, A.S. (1983). Circulant plant for partial diallel crosses. *Biometrics*, **39**, 43–52.
- Arya, A.S. and P. Narain (1977). Practical diallel crosses based on some association schemes with three or four associate classes. *Sankhya Ser. B*, **39**, 394–399.
- Baksalary, J.K. and Z. Tabis (1987). Conditions for the robustness of block designs against the unavailability of data. *J. Stat. Plan. Inference*, **16**, 49–54.
- Brownman, K.W. (2005). The genomes of recombinant inbred lines. *Genetics*, **169**, 1133–1146.
- Chai, F.S. and R. Mukerjee (1999). Optimal designs for diallel crosses With specific combining abilities. *Biometrika*, **86**, 253–458.
- Choi, K.C. and S. Gupta (2000). On constructions of optimal complete diallel crosses. *Utilitas Math.*, **58**, 153–160.
- Choi, K.C., K. Chatterjee, A. Das, and S. Gupta (2002). Optimality of orthogonally blocked diallels with specific combining abilities. *Stat. Prob. Lett.*, **57**, 145–150.
- Choi, K.C., S. Gupta, and Y.N. Son (2002). Partial diallel cross design. *Ars Combinatoria*, **64**, 51–64.
- Choi, K.C., S. Gupta, and S. Kageyama (2004). Designs for diallel crosses for test versus control Comparisons. *Utilitas Math.*, **65**, 167–180.
- Clatworthy, W.H. (1955). Partially balanced incomplete block designs with two associate classes and two treatments per block. *J. Res. Nat. Bur. Stand.*, **54**, 177–190.
- Clatworthy, W.H. (1973). *Tables of Two-Associate-Class Partially Balanced Designs*. Washington, DC: U.S. Department of Commerce.
- Cockerham, C.C. (1963). Estimation of genetic variances. *Statistical Genetics and Plant Breeding. Natl. Acad. Sci. Natl. Res. Council Publ.*, **982**, 53–94.
- Comstock, R.E. and H.F. Robinson (1952). Estimation of average dominance of genes. In: *Heterosis*, J.W. Gowen. Ames: Iowa State College Press, pp. 494–516.
- Curnow, R.N. (1963). Sampling the diallel cross. *Biometrics*, **19**, 287–306.
- Das, A. (2003). Efficient control-test designs for diallel cross experiments. *Sankhya*, **65**, 678–688.
- Das, A. and A. Dey (2005). Designs for diallel cross experiments with specific combining abilities. *J. Indian Soc. Agric. Stat.*, **57**, 247–256.
- Das, A. and S. Kageyama (1992). Robustness of BIB and extended BIB designs against the unavailability of any number of observations in a block. *Comput. Stat. Data Anal.*, **14**, 343–358.
- Das, A., A.M. Dean, and S. Gupta (1998). On optimality of some partial diallel cross designs. *Sankhya Ser. B*, **60**, 511–524.
- Das, A., A. Dey, and A.M. Dean (1998). Optimal block designs for diallel cross experiments. *Stat. Prob. Lett.*, **36**, 427–436.
- Das, A., S. Gupta, and S. Kageyama (2006). A-optimal diallel crosses for test versus control comparisons. *J. Appl. Stat.*, **33**, 601–608.

- Dey, A. and S.P. Dhall (1988). Robustness of augmented BIB designs. *Sankhya Ser. B*, **50**, 376–381.
- Dey, A. and C.K. Midha (1996). Optimal block designs for diallel crosses. *Biometrika*, **83**, 484–489.
- Dey, A., R. Srivastava, and R. Parsad (2001). Robustness of block designs for diallel crossing plans against missing observations. *J. Indian Soc. Agric. Stat.*, **54**, 376–384.
- Doerge, R.W., B.S. Wier, and Z.-B. Zeng (1997). Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Stat. Sci.*, **12**, 195–219.
- Dunwell, J.M. (2010). Haploids in flowering plants: origins and exploitation. *Plant Biotech. J.*, **8**, 377–424.
- Eccleston, J.A. and A. Hedayat (1974). On the theory of connected designs: characterization and optimality. *Ann. Stat.*, **2**, 1238–1255.
- Falconer, D.S. and T.F.C. MacKay (1996). *Quantitative Genetics*. Harlow: Addison Wesley.
- Federer, W.T. (1967). Diallel cross designs and their relation to fractional replication. *Appl. Genet.*, **37**, 174–178.
- Forster, B.P. and W.T.B. Thomas (2010). Doubled haploids in genetics and plant breeding. In: *Plant Breeding Reviews*, vol. 25. J. Janick (ed.). Oxford: John Wiley.
- Fyfe, J.L. and N. Gilbert (1963). Partial diallel crosses. *Biometrics*, **19**, 278–286.
- Ghosh, D.K. and P.C. Biswas (2000). Robust designs for diallel crosses against the missing of one block. *J. Appl. Stat.*, **27**, 715–723.
- Ghosh, D.K. and N.R. Desai (1998). Robustness of complete diallel crosses plans to the unavailability of one block. *J. Appl. Stat.*, **25**, 827–837.
- Ghosh, D.K. and N.R. Desai (1999). Robustness of a complete diallel crosses plan with an unequal number of crosses to the unavailability of one block. *J. Appl. Stat.*, **26**, 563–577.
- Ghosh, D.K. and J. Divecha (1997). Two associate class partially balanced incomplete block designs and partial diallel crosses. *Biometrika*, **84**, 245–248.
- Ghosh, H. and A. Das (2004). Optimal diallel cross designs for the interval estimation of heredity. *Stat. Prob. Lett.*, **67**, 47–55.
- Ghosh, H. and A. Das (2005). Optimal designs for best linear unbiased prediction in diallel crosses. *Comm. Stat. Theory Methods*, **34**, 1579–1586.
- Ghosh, H., A. Das, and C.K. Midha (2005). Optimal designs for estimation of ratio of variance components in diallel crosses. *Sankhya*, **67**, 785–794.
- Ghosh, S. (1982). Robustness of designs against the unavailability of data. *Sankhya Ser. B*, **44**, 50–62.
- Ghosh, S., S.B. Rao, and N.M. Singhi (1983). On robustness property of PBIBD. *J. Stat. Plan. Inference*, **8**, 355–363.
- Gilbert, N. (1958). Diallel cross in plant breeding. *Heredity*, **12**, 477–492.
- Griffing, B. (1956). Concepts of general and specific combining ability in relation to diallel crossing systems. *Aust. J. Biol. Sci.*, **9**, 463–493.
- Gupta, S. and S. Kageyama (1994). Optimal complete diallel crosses. *Biometrika*, **81**, 420–424.

- Gupta, S., A. Das, and S. Kageyama (1995). Single replicate orthogonal block designs for circulant partial diallel crosses. *Comm. Stat. Theory Methods*, **24**, 2601–2607.
- Hayman, B.I. (1954a). The analysis of variance of diallel tables. *Biometrics*, **10**, 235–244.
- Hayman, B.I. (1954b). The theory and analysis of diallel crosses. *Genetics*, **39**, 789–809.
- Hedayat, A. and P.W.M. John (1974). Resistant and susceptible BIB designs. *Ann. Stat.*, **2**, 148–158.
- Hinkelmann, K. (1965). Partial triallel crosses. *Sankhya Ser. A*, **27**, 173–196.
- Hinkelmann, K. (1966). Unvollständige diallel Kreuzungspläne. *Biom. Z.*, **8**, 242–265.
- Hinkelmann, K. (1975). Design of genetical experiments. In: *A Survey of Statistical Design and Linear Models*, J.N. Srivastava (ed.). Amsterdam: North Holland, pp. 243–269.
- Hinkelmann, K. and O. Kempthorne (1963). Two classes of group divisible partial diallel crosses. *Biometrika*, **50**, 281–291.
- Hinkelmann, K. and K. Stern (1960). Kreuzungspläne zur Selektionszüchtung bei Waldbäumen. *Silvae Genet.*, **9**, 121–133.
- Jensen, J. (1989). Estimation of recombination parameters between a quantitative trait locus (QTL) and two marker gene loci. *Theor. Appl. Genet.*, **78**, 613–618.
- Kang, M.S. (ed.). (2003). *Handbook of Formulas and Software for Plant Geneticists and Breeders*. New York: Food Products Press, The Haworth Reference Press.
- Kearsey, M.J. and H.S. Pooni (1996). *The Genetical Analysis of Quantitative Traits*. London: Chapman and Hall.
- Kempthorne, O. (1956). The theory of the diallel crosses. *Genetics*, **41**, 451–459.
- Kempthorne, O. (1957). *An Introduction to Genetic Statistics*. New York: John Wiley and Sons.
- Kempthorne, O. and R.N. Curnow (1961). The partial diallel cross. *Biometrics*, **17**, 229–250. [See also correction, *Biometrics*, 18 (1962), 128].
- Kiefer, J. (1958). On the nonrandomized optimality and randomized nonoptimality of symmetric designs. *Ann. Math. Stat.*, **29**, 675–699.
- Kiefer, J. (1959). Optimal experimental designs. *J. R. Stat. Soc. B*, **21**, 272–319.
- Kiefer, J. (1975). Construction and optimality of generalized Youden designs. In: *A Survey of Statistical Designs and Linear Models*, J.N. Srivastava (ed.). Amsterdam: North-Holland, pp. 333–353.
- Knapp, S.J. (1991). Using molecular markers to map multiple quantitative trait loci: Models for backcross, recombinant inbreed and doubled haploid progeny. *Theor. Appl. Genet.*, **81**, 333–338.
- Lander, E.S. and D. Botstein (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185–199.
- Lynch, M. and B. Walsh (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland, USA: Sinauer Associates, Inc.
- Mansson, R.A. and P. Prescott (2002). Missing observations in Youden square designs. *Comput. Stat. Data Anal.*, **40**, 329–338.
- Maria-Joao, P., M. Boer, X. Huang, M. Koornneef, and F. Eeuwijk (2008). A mixed model QTL analysis for a complex cross population consisting of a half diallel of

- two-way hybrids in *Arabidopsis thaliana*: Analysis of simulated data. *Euphytica*, **161**, 107–114.
- Martin, R.J. and J.A. Eccleston (1991). Optimal incomplete block designs for general dependence structures. *J. Stat. Plan. Inference*, **28**, 6781.
- Mather, K. and J.L. Jinks (1982). *Biometrical Genetics* (2nd ed.). London: Chapman and Hall Ltd.
- Mathur, S.N. and P. Narain (1976). Some optimal plans for partial diallel crosses. *Indian J. Genet.*, **36**, 301–308.
- Matzinger, D.F. and O. Kempthorne (1956). The modified diallel table with partial inbreeding and interactions with environment. *Genetics*, **41**, 822–833.
- Monaghan, F. and A. Corcos (1986). Tschermark: A non-discover of Mendelism. I. An historical note. *J. Hered.*, **77**, 468–469.
- Monaghan, F. and A. Corcos (1987). Tschermark: A non-discover of Mendelism. II. A critique. *J. Hered.*, **78**, 208–210.
- Mukerjee, R. (1997). Optimal partial diallel crosses. *Biometrika*, **84**, 939–948.
- Mukerjee, R. and S. Kageyama (1990). Robustness of group divisible designs. *Comm. Stat. Theory Methods*, **19**, 189–3203.
- Narain, P. (1990). *Statistical Genetics*. New Delhi: Wiley Eastern.
- Narain, P. and A.S. Arya (1981). Truncated triangular association scheme and related partial diallel crosses. *Sankhya Ser. B*, **43**, 93–103.
- Narain, P., C. Subbarao, and A.K. Nigam (1974). Partial diallel crosses based on extended triangular association scheme. *Ind. J. Genet.*, **34**, 309–317.
- Panda, D.K., R. Parsad, and V.K. Sharma (2003). Robustness of complete diallel crossing plans against exchange of one cross. *J. Appl. Stat.*, **30**, 21–35.
- Panda, D.K., R. Parsad, and V.K. Sharma (2004). Robustness of block designs for complete diallel crosses against interchange of a pair of crosses. In: *Recent Advances in Mating Designs*, L.S. Kaushik and R.C. Hasija (eds.). New Delhi: Dhanpat Rai and Company, pp. 1.11–1.28.
- Parsad, R., V.K. Gupta, and R. Srivastava (1999). Optimal designs for diallel crosses. *J. Soc. Stat. Comput. Appl.*, **1**, 35–52.
- Parsad, R., V.K. Gupta, and S. Gupta (2005). Optimal designs for experiments on two-line and four-lines crosses. *Utilitas Math.*, **68**, 11–32.
- Patel, J.D., B.R. Christie, and L.W. Kannenberg (1984). Line \times tester Crosses: A new approach of analysis. *Can. J. Genet. Cytol.*, **26**, 523–527.
- Payne, R.W. (2009). GenStat. *Wiley Interdiscip. Rev. Comput. Stat.*, **1**, 255–258.
- Pearce, S.C. (1960). Supplemented balance. *Biometrika*, **47**, 263–271.
- Pederson, D.G. (1980). The augmented partial diallel cross. *Heredity*, **44**, 327–331.
- Ponnuswamy, K.N., M.N. Das, and M.I. Handoo (1974). Combining ability type of analysis for triallel crosses in maize. *Theor. Appl. Genet.*, **45**, 170–175.
- Preece, D.A. (1967). Nested balanced incomplete block designs. *Biometrika*, **54**, 479–486.
- Prescott, P. and R. Manson (2004). Robustness of diallel cross designs to the loss of one or more observations. *Comput. Stat. Data Anal.*, **47**, 91–109.

- Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. New York: John Wiley.
- Rawlings, J.O. and C.C. Cockerham (1962a). Triallel analysis. *Crop Sci.*, **2**, 228–231.
- Rawlings, J.O. and C.C. Cockerham (1962b). Analysis of double hybrid populations. *Biometrics*, **18**, 229–244.
- Roy, R.M. (1953–1954). Hierarchical group divisible incomplete block designs with m associate classes. *Sci. Cult.*, **19**, 210–211.
- Sarkar, S., R. Parsad, and V.K. Gupta (2005). Outliers in block designs for diallel crosses. *Metron. Int. J. Stat.*, **63**, 177–191.
- Schmidt, J. (1919). La valeur de l'individu à titre de générateur apprécié suivant méthode du croisement diallel. *Compt. Rend. Lab. Carlsberg*, **14**(6), 1–33.
- Sen, S. and G.A. Churchill (2001). A statistical framework for quantitative trait mapping. *Genetics*, **159**, 371–387.
- Shaffer, J.G. and S.K. Srivastav (2009). A simple technique for constructing optimal complete diallel cross designs. *J. Stat. Prob. Lett.*, **79**, 1181–1185.
- Shah, B.V. (1959). A generalization of partially balanced incomplete block designs. *Ann. Math. Stat.*, **30**, 1041–1050.
- Sharma, M.K. and S. Fanta (2010). Optimal block designs for diallel crosses. *Metrika*, **71**, 361–372.
- Singh, M. and A. Dey (1979). Block designs with nested rows and columns. *Biometrika*, **66**, 321–326.
- Singh, M. and K. Hinkelmann (1988). On generation of efficient partial diallel crosses plans. Technical report No. 88-24. Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg.
- Singh, M. and K. Hinkelmann (1990). On generation of efficient partial diallel crosses plans. *Biometrical J.*, **32**, 177–187.
- Singh, M. and K. Hinkelmann (1995). Partial Diallel Crosses in Incomplete Blocks. *Biometrics*, **51**, 1302–1314.
- Singh, M. and K. Hinkelmann (1998). Analysis of partial diallel crosses in incomplete blocks. *Biometrical J.*, **40**, 165–181.
- Singh, R.K. and B.D. Chaudhary (1979). *Biometrical Methods in Quantitative Genetic Analysis*. New Delhi: Kalyani Publishers.
- Sprague, G.F. and L.A. Tatum (1942). General vs. specific combining ability in single crosses of corn. *Am. Soc. Agron.*, **34**, 923–932.
- Srinivasan, M.R. and K.N. Ponnuswamy (1993). Estimation of variance components based on triallel mating design. *Theor. Appl. Genet.*, **85**, 593–597.
- Srivastav, S. and A. Shankar (2007). On the construction and existence of a certain class of diallel cross designs. *Statist. Prob. Lett.*, **77**, 111–115.
- Srivastava, R., V.K. Gupta, and A. Dey (1990). Robustness of some designs against missing observations. *Comm. Stat. Theory Methods*, **19**, 121–126.
- Stam, P. (1993). Construction of integrated genetic linkage map by means of a computer package: Joinmap. *Plant J.*, **5**, 739–744.
- Street, D.J. (1992). A note on strongly equineighboured designs. *J. Stat. Plan. Inference*, **30**, 99–105.

- Vartak, M.N. (1959). The non-existence of certain PBIB designs. *Ann. Math. Stat.*, **30**, 1051–1062.
- Verhoeven, K.J., J. Jannink, and L. McIntyre (2006). Using mating designs to uncover QTL and the genetic architecture of complex traits. *Heredity*, **96**, 139–149.
- Willham, R.L. (1963). The covariance between relatives for characters composed of components contributed by related individuals. *Biometrics*, **19**, 18–27.
- Zeng, Z.-B. (1994). Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457–1468.

C H A P T E R 2

Design of Gene Expression Microarray Experiments

Dan Nettleton

2.1 INTRODUCTION

The quantity of messenger ribonucleic acid (mRNA) produced by a gene is referred to as a gene's *expression level*. Microarray technology allows researchers to simultaneously measure the expression levels of thousands of genes in multiple biological samples. By understanding how expression levels change across experimental conditions, researchers gain clues about gene function and learn how genes work together to carry out biological processes.

Since Schena et al. (1995) first described the use of microarray technology to measure gene expression, many thousands of microarray experiments have been conducted. The Gene Expression Omnibus (GEO) database (Edgar, Domrachev, and Lash 2002) contains publicly available data from well over 10,000 microarray experiments involving more than 500 different types of organisms (Barrett et al. 2009). The aim of many of these experiments is to identify genes whose expression level distributions change in response to treatment. Such genes are called *differentially expressed*. This chapter focuses on the design of microarray experiments aimed at identifying differentially expressed genes.

In Section 2.2, we introduce gene expression microarray technology and explain how microarrays are used to obtain quantitative measures of gene expression levels. Section 2.3 introduces methods for preprocessing the raw expression measures to obtain data suitable for statistical analysis. Section 2.4 provides an introduction to the principal topic of this chapter, design of gene expression microarray experiments. Sections 2.5 through 2.8 cover

Design and Analysis of Experiments: Special Designs and Applications, First Edition. Edited by Klaus Hinkelmann.

© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

specific design topics, including two-treatment experiments, experiments involving more than two treatments, multifactor experiments, and more complex experiments that involve multiple blocking factors, multiple observations per experimental unit, and/or multiple sizes of experimental unit.

2.2 GENE EXPRESSION MICROARRAY TECHNOLOGY

2.2.1 Introduction

Although many different types of microarrays have been used to identify differentially expressed genes over the past 15 years, the various technologies share many common characteristics that will be described in this section. For more detailed information about the biology of gene expression and various aspects of gene expression microarray technology, statistically trained readers are encouraged to see the excellent article by Nguyen et al. (2002).

2.2.2 Definition of a Microarray

A *microarray* is a glass microscope slide, nylon membrane, silicone chip, or other small solid support onto which DNA sequences from thousands of genes are arranged in an array with a fixed number of rows (r) and a fixed number of columns (c). Each array element contains many copies of a particular DNA sequence corresponding to a gene. Following the nomenclature of Phimister (1999), we will refer to an array element as a *probe*. In the simplest case, there is one unique probe for each gene represented on the microarray. The probe sequences and their locations are known by the user.

In the usual case, each microarray is a disposable measuring device that can only be used once to obtain gene expression measurements. Thus, multiple copies of a microarray are needed to measure experimental units in a microarray experiment. Due to manufacturing constraints, the array position of each probe is kept constant across copies of a microarray. Hence, for any $i = 1, \dots, r$ and $j = 1, \dots, c$, the probe in the i th row and j th column is the same across all copies of a microarray. This means that array position is completely confounded with gene. While not ideal, this confounding is not as problematic as it might seem given that within-gene (rather than between-gene) comparisons are of primary interest.

2.2.3 Using Microarrays to Measure Gene Expression

Microarray technology relies on complementary base pairing to obtain gene expression measurements. Recall that each strand of DNA is a sequence of nucleotides. Each nucleotide has one of four bases: adenine (A), thymine (T), guanine (G), and cytosine (C). The bases A and T are complementary, as are G and C. This means, for example, that the DNA strand GATTACA will bind

with the complementary DNA strand CTAATGT (G will bind with C, A will bind with T, T will bind with A, etc.) to form a double-stranded molecule. This binding process is known as *hybridization*.

The single-stranded DNA probes fixed to a microarray slide are free to hybridize with nucleic acid sequences derived from an mRNA sample of interest. By quantifying the amount of hybridization to each probe, gene-specific measures of the sample's mRNA content can be obtained. The exact protocols used at this step vary across microarray technologies, but the basic idea is as follows.

Messenger RNA molecules are extracted from a biological sample of interest. A fluorescent label, also known as a dye, is attached to nucleic acid molecules derived from the extracted mRNA molecules. This fluorescently labeled *target* sample consists of unknown sequences whose identities will be discovered by hybridization to a microarray slide. If a particular gene (say gene *g*) produced mRNA sequences present in the biological sample, then the target sample will contain labeled nucleic acid molecules that are complementary to the microarray probe for gene *g*. To determine which labeled nucleic acid sequences in the target sample are complementary to which probes, labeled nucleic acid molecules are placed on the microarray slide and allowed to hybridize to complementary probes. Labeled sequences that are unable to find a complementary probe on the microarray slide are washed from the microarray slide. (Such sequences might correspond to genes that were not represented by probes on the microarray slide.) A laser is then used to excite the fluorescent dye attached to the remaining nucleic acid molecules that were able to hybridize with DNA probes. A scanner is used to capture an image, and the level of fluorescence intensity emitted by labeled nucleic acid molecules bound to each microarray probe is quantified. The higher the measurement of fluorescence intensity, the greater the number of dye molecules at a particular probe. Thus, a high fluorescence intensity measurement at the probe for gene *g* suggests that a large number of nucleic acid molecules were able to bind to the DNA probe for gene *g*. This indicates that gene *g* produced many messenger RNA sequences in the biological sample from which the target sample was derived. Analogously, a low fluorescence intensity measurement at the probe for gene *g* suggests a low level of expression for gene *g* in the biological sample of interest.

2.2.4 Types of Gene Expression in Microarrays

Gene expression microarrays can be broadly classified as either single-channel or dual-channel. The original microarray described by Schena et al. (1995) is an example of a dual-channel microarray. Dual-channel microarrays are also known as *two-color microarrays* because they use two different dyes—usually Cyanine 3 (green) and Cyanine 5 (red)—to measure expression. The use of two dyes allows two different nucleic acid samples, dyed with different dyes, to be hybridized together to the same microarray slide. Because the two dyes

fluoresce at different wavelengths, it is possible to obtain two separate measurements of fluorescence intensity for each probe. Two-color microarray images are often displayed with spots, corresponding to probes, colored according to the intensity of the green and red signals. As we shall discuss subsequently, the ability to pair samples on a single array makes experimental design for two-color microarray experiments especially interesting from a statistical perspective.

Single-channel microarrays use only one dye. Thus, only one nucleic acid sample can be measured on each microarray. The most common single-channel microarrays are manufactured by Affymetrix™ and are known as Affymetrix™ GeneChips®. Each gene on Affymetrix GeneChip is represented by multiple probes. Each probe consists of millions of copies of a short probe-specific DNA sequence (usually 25 bases) known as an oligonucleotide. The collection of probes corresponding to a single gene is known as a *probe set*. As discussed in the next section, the fluorescence intensity signals from the probes in a probe set can be combined together in many ways to obtain a gene-specific measure of expression.

2.3 PREPROCESSING OF MICROARRAY FLUORESCENCE INTENSITIES

2.3.1 Introduction

The raw fluorescence intensities obtained from gene expression microarrays are usually processed prior to use. As discussed by Wu and Irizarry (2007), this preprocessing can be divided into three stages: background correction, normalization, and summarization. In this section, we briefly describe the idea behind these preprocessing steps and provide references to relevant literature. A full treatment of these topics is beyond the scope of this chapter and is somewhat tangential to our main focus on microarray experimental design. Thus, the material in this section can be skipped by readers primarily interested in experimental design.

2.3.2 Background Correction

Background correction involves an attempt to remove any portion of a raw fluorescence intensity measurement that is not attributable to fluorescence from target nucleic acid molecules hybridized to their complementary probe. Example sources of fluorescence other than hybridized target nucleic acid molecules include fluorescence in the microarray slide itself, fluorescence from neighboring probe spots, or fluorescence from unbound labeled nucleic acid sequences or other stray particles not washed from the slide. Due to these and related factors, some regions of a microarray slide may tend to be brighter than others. Thus, adjustments for variation in brightness across a microarray slide might lead to better measures of gene expression.

For two-color microarrays, image processing software provides a measure of the fluorescence intensity around each probe spot that is often used as a measure of local background fluorescence. Subtracting the measure of local background fluorescence from the raw fluorescence intensity measurement of a microarray probe spot is one of the simplest forms of background correction. One drawback of this approach is that a background fluorescence measurement is sometimes greater than the corresponding fluorescence measurement of the probe spot. Thus, background subtraction results in a negative fluorescence value that is not sensible as a measure of gene expression.

Silver, Ritchie, and Smyth (2009) described a more sophisticated approach for obtaining background corrected measures of expression. Following Irizarry et al. (2003), they modeled the raw fluorescence intensity minus background at a given spot as a convolution of a normally distributed error and an independent exponentially distributed signal due to fluorescence from bound complementary target nucleic acid molecules. Using data from all probes, they described how to obtain maximum likelihood estimates of model parameters and probe-specific estimates of the conditional expected value of the exponential signal, given the observed difference between the raw fluorescence intensity and the local background. The estimated conditional expectation (which is guaranteed to be nonnegative) serves as the background corrected measure of expression for each probe. Ritchie et al. (2007) found advantages for this approach over seven other background correction methods for two-color microarrays.

Background correction is somewhat different for Affymetrix GeneChip data because no measures of local background fluorescence are available due to the compact arrangement of Affymetrix probes. Affymetrix preprocessing algorithms use the following background correction method. First, each GeneChip is divided into 16 rectangular zones. Next, the lowest 2% of intensities in each zone are averaged to form zone-specific background values. The background for a given probe is then calculated as the weighted average of the zone-specific background values, with weights that are inversely proportional to a constant plus the squared distances between the probe's location on the GeneChip and the zone centers. This approach can help adjust for spatial trends in fluorescence across the surface of a GeneChip.

As an alternative, Irizarry et al. (2003) proposed the normal-exponential convolution model that was later improved and extended to two-color arrays by Ritchie et al. (2007) and Silver, Ritchie, and Smyth (2009) as discussed above. Other approaches for background correction of Affymetrix GeneChip data are discussed by Irizarry, Wu, and Jaffee (2006), who showed that the background correction step can have a big impact on the performance of an Affymetrix preprocessing algorithm.

2.3.3 Normalization

The main aim of normalization is to remove technical artifacts from the fluorescence intensities that are unrelated to biological variation in mRNA levels.

For example, if the distribution of log-scale fluorescence intensities across all genes for one experimental unit (e.g., animal, plant, bacterial colony, etc.) is shifted by several units relative to another, it often makes sense to align these distributions at a common centerpoint before analyzing data for individual genes. Researchers attempt to measure the same total amount of labeled target for each experimental unit. Thus, shifts in the entire distribution of expression measurements from experimental unit to experimental unit are not biologically meaningful and usually are attributed to technical variation.

Sources of technical variation include variation across replicate microarray slides resulting from the manufacturing process, variation in the preparation of target samples, dye variation (especially for two-color arrays), and variation across various steps in the measurement process, such as hybridization, washing, and microarray image acquisition. Of course, if appropriate experimental designs are used, these technical nuisance factors will not be confounded with factors of interest. However, variation across these factors contributes to measurement error, and thus, adjustments that mitigate these factors can increase power for identifying genes whose expression level distributions change across treatments.

If only a single gene were measured for each experimental unit, it would be impossible to adjust for many technical nuisance factors without additional information. However, because thousands of genes are measured simultaneously for each experimental unit, it is possible to make reasonable adjustments for most technical nuisance factors whenever their effects are similar across all genes or across all genes with similar expression levels. The details of such adjustments are covered in a large literature on normalization methods for microarray data that we will not attempt to review here. Papers by Yang et al. (2002) and Smyth and Speed (2003) are good starting points for readers who wish to learn more about normalization methods for two-color microarray data sets.

Normalization for Affymetrix GeneChip data and related methods for expression measure construction have also received substantial attention in the literature. Although the robust multiarray average (RMA) method proposed by Irizarry et al. (2003) is widely used, there are many other options. Irizarry, Wu, and Jaffee (2006) described a Web tool that can be used to compare the performance of Affymetrix GeneChip preprocessing methods using benchmark data sets. More than 31 different methods were compared by the authors, and many additional methods have since been evaluated using the Web tool.

2.3.4 Summarization

When a gene is represented by more than one probe on a microarray slide, the probe-specific measures of expression are typically combined together to produce one gene-specific measure of expression for each target sample. For the sake of illustration, consider a hypothetical single-channel microarray

where each of thousands of genes is represented by p probes. Let y_1, \dots, y_p denote expression measures (perhaps background corrected and normalized) for a given gene on a given microarray slide. Because the length of any probe DNA sequence is far shorter than the DNA sequence of the gene that it measures, it is typical for the p probes for any one gene to be designed to match different portions of the gene sequence. As a result, the p probes differ in sequence and provide measures of gene expression of variable quality.

One obvious summary of the p measures of expression is the average $\bar{y} = \sum_{j=1}^p y_j$. However, heterogeneity in probe quality often makes the simple average far from ideal. Thus, many more complex strategies have been proposed for summarizing data from multiple probes. Much of the work in this area has been carried out for Affymetrix GeneChips, where multiple different probes per gene is the norm.

The summarization algorithm recommended by Affymetrix is based on a one-step version of Tukey's bi-weight estimator (see chapter 10 of Mosteller and Tukey 1977). The details are as follows. Let m equal the median of y_1, \dots, y_p , and let M denote the median of $|y_1 - m|, \dots, |y_p - m|$. For $j = 1, \dots, p$, let

$$t_j = (y_j - m) / (5M + 0.0001).$$

Define

$$B(t) = \begin{cases} (1-t^2)^2 & \text{for } |t| < 1, \\ 0 & \text{for } |t| \geq 1. \end{cases}$$

Then the summarized measure of gene expression is given by

$$\frac{\sum_{j=1}^p B(t_j) y_j}{\sum_{j=1}^p B(t_j)}.$$

This summarization technique heavily weights probes whose expression measures are close to the median across the p probes and downweights probes with expression measures far from the median. This approach is quite useful for automatic outlier removal if, for example, a manufacturing or processing error causes a probe to produce an errant measurement.

The popular RMA procedure also uses a summarization strategy that attempts to reduce the impact of outlying observations through the use of Tukey's median polish procedure (see chapter 9 of Mosteller and Tukey 1977). See Irizarry et al. (2003) for a detailed description of the RMA procedure. For a comparison of the Affymetrix and RMA strategies with other summarization strategies, see Cope et al. (2004), Choe et al. (2005), and Irizarry, Wu, and Jaffee (2006).

2.4 INTRODUCTION TO GENE EXPRESSION MICROARRAY EXPERIMENTAL DESIGN

In the remaining sections of this chapter, we consider the design of microarray experiments aimed at identifying differentially expressed genes. Although the response in a microarray experiment is a high-dimensional vector obtained via the procedures described in Sections 2.2 and 2.3, it is useful to focus on the data for a single gene when studying the design of microarray experiments. Fortunately, most experimental design questions can be investigated as if only a single response variable is to be measured, because the same principles that lead to a good experimental design for a single response also apply to experiments with thousands of response variables. (The topic of power and sample size calculation is one exception. See e.g., Gadbury et al. 2004; Ruppert, Nettleton, and Hwang 2007.)

Because we will be concerned with the analysis of only a single gene in this chapter, we will suppress gene subscripts for all variables and model parameters. However, it should be understood that all model parameters are expected to vary from gene to gene, and it is important to note that the very large and growing body of literature on microarray data analysis focuses primarily on the simultaneous analysis of thousands of genes. Strategies for multiple testing and for combining information across thousands of genes are two very important and interesting analysis topics that space considerations do not allow us to address in this design-focused chapter.

As noted by Kerr (2003) and Jarrett and Ruggiero (2008), microarray experiments are *two-phase experiments* in the sense of McIntyre (1955). In the first phase, an experiment is conducted using the fundamental principles of experimental design (replication, randomization, and blocking) to assign treatments to experimental units and to obtain tissue samples for measurement with microarrays. The design of the first-phase experiment could be any one of many possible designs, including a completely randomized design, a randomized complete block design, an incomplete block design, a Latin square design, a split-plot design, and so on depending on the questions of interest and available resources. For example, in an experiment to compare the effects of two treatments on gene expression in plants, a randomized complete block design with 10 blocks might be a natural choice if 10 pots, each with two plants, comprise the available experimental material.

The second phase of a microarray experiment involves measuring the mRNA content of the tissue samples using microarrays. It is the design issues that arise in the second phase that make experimental design for microarray experiments an interesting topic from a statistical perspective. Two-color microarray experiments in particular offer some unique challenges that will be explored in the remaining sections of this chapter. The opportunity to hybridize two target samples to the same microarray slide enables a comparison of two treatments on a single microarray slide and under identical washing and hybridization conditions. This local control or blocking can eliminate the impact of slide-to-slide variation and variation in washing and hybridization

conditions on estimates of treatment effects. However, as discussed in Section 2.2.4, target samples must be dyed with two different dyes in order to be measured together on a single microarray slide. Although normalization methods attempt to remove global differences due to the dyes (see e.g., Dudoit et al. 2002; Yang et al. 2002), gene-specific dye effects cannot be entirely removed by normalization (see e.g., Yang and Speed 2002, or the data presented by Landgrebe, Bretz, and Brunner 2006). Thus, the elimination of some nuisance factors (microarray slide, washing conditions, hybridization conditions) introduces a new nuisance factor (dye). As illustrated in the following sections, the second-phase design—that is, determining how best to pair experimental units on microarray slides and how to assign dyes to each member of each pair—can be quite challenging, especially if the first phase of the experiment involves a complex experimental design.

2.5 TWO-TREATMENT EXPERIMENTS USING TWO-COLOR MICROARRAYS

The simplest two-color microarray experiments involve a comparison of two treatments using either a balanced, that is, equirePLICATE, and completely randomized design (CRD) or a randomized complete block design (RCBD) for the first phase of the experiment. First, suppose the number of experimental units per treatment r matches the number of available two-color microarray slides. If a CRD is used in the first phase of the experiment, then experimental units from treatment group 1 should be randomly paired with experimental units in treatment group 2 and assigned to microarray slides with one pair per slide. Suppose the two possible dye assignments (green for treatment 1 and red for treatment 2 and the reverse) are randomly assigned to the r slides with r_1 slides for dye assignment 1 and r_2 slides for dye assignment 2 ($r_1 + r_2 = r$).

An example of this design with $r_1 = r_2 = 4$ is depicted in Figure 2.1 using a two-color microarray experimental design symbolism motivated by Kerr and

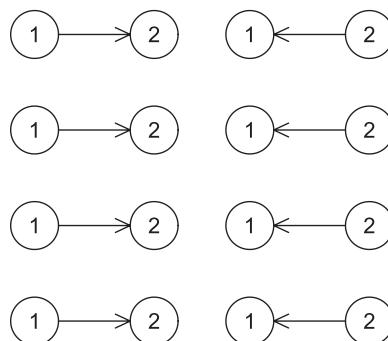


Figure 2.1 Recommended design for comparing two treatments with 8 slides.

Churchill (2001). Each experimental unit is represented by a circle. The labels on each circle indicate the treatment assigned to the experimental unit. An arrow connecting two experimental units signifies co-hybridization of target samples from the experimental units on a single microarray slide. The direction of the arrow indicates the dye assignment, with the convention that the arrow emanates from the green-dyed sample and points to the red-dyed sample. Variations of this symbolism are common throughout the microarray literature. Unfortunately, some of the past work on microarray experimental design has not been careful to distinguish repeated measurement of a single experimental unit (known as technical replication in the microarray literature) from the measurement of multiple independent experimental units treated alike (known as biological replication in the microarray literature). By insisting on a one-to-one correspondence between circles and experimental units in our version of the symbolism, we hope to make this distinction clear.

If an RCBD is used in the first phase of the experiment, the same strategy depicted in Figure 2.1 should be used except that random pairing of the treatment 1 and treatment 2 experimental units is neither necessary nor desirable. The blocks from the first phase of the experiment are used to define the pairs for the second phase of the experiment. This completely confounds blocks from the first phase of the experiment with microarray slides in the second phase, but this confounding is typically desirable because block and slide are each nuisance factors. It is important to control for variation across the levels of these nuisance factors, but in the usual case, it is not necessary to obtain separate estimates of variation across the levels of these factors. By intentionally confounding them, it is possible to control for variation across levels of both factors simultaneously.

A model for the log-scale expression data for a single gene obtained from the design in Figure 2.1 is

$$y_{ij(k)} = \mu + s_i + \delta_j + \tau_k + e_{ij(k)}, \quad (2.1)$$

where μ is a parameter common to all observations; s_1, \dots, s_r are slide effects; δ_1 and δ_2 are dye effects; τ_1 and τ_2 are treatment effects; and the $e_{ij(k)}$ terms are independent random errors with mean zero and variance σ^2 . If an RCBD is used in the first phase of the experiment, each s_i effect is the sum of a block effect from phase 1 and slide effect from phase 2. Regardless of whether the first phase of the experiment is conducted as a CRD or an RCBD, the effects s_1, \dots, s_r may be treated as either fixed or random independent effects with mean zero and variance σ_s^2 . If the effects s_1, \dots, s_r are treated as random, they are assumed to be independent of the random errors.

With slide and dye as blocking factors, the design in Figure 2.1 is a Latin rectangle design, and model (2.1) is equivalent to the model for a Latin rectangle design presented in HK1, section 10.4. The design and modeling also match a two-period crossover (or change-over) design assuming no carryover effects (see HK1, section 10.7.1 and, more generally, HK2, chapter 19). Each

treatment occurs once on each slide and $r_1 = r_2 = 4$ times with each dye. We will use the terminology *dye-balanced* to describe designs for which the number of experimental units per dye is the same within each treatment.

Recall that the goal of the experiment is to determine whether each gene is differentially expressed. Assuming that model (2.1) holds for a given gene, the null hypothesis of no differential expression is

$$H_0 : \tau_1 = \tau_2.$$

Assuming that the random errors in model (2.1) are normally distributed, it is straightforward to test this null hypothesis using standard linear model analysis methods. However, it is common practice in the analysis of two-color microarray data to use the log of the red-to-green expression ratios from each slide as the response. This is equivalent to analyzing the differences

$$d_i \equiv y_{i2(k_i)} - y_{i1(k_i^*)},$$

for $i = 1, \dots, r$, where (k_i, k_i^*) is either $(1, 2)$ or $(2, 1)$ depending on the dye assignment for slide i .

For $i = 1, \dots, r$, let $a_i \in \{1, 2\}$ denote the dye assignment number for the i th slide, where $a_i = 1$ denotes green dye for treatment 1 and red for treatment 2, and $a_i = 2$ denotes the reverse. Let

$$\mu_1 = \delta_2 - \delta_1 - (\tau_1 - \tau_2), \mu_2 = \delta_2 - \delta_1 + (\tau_1 - \tau_2), \text{ and } \eta^2 = 2\sigma^2.$$

Then, model (2.1) implies that the differences d_1, \dots, d_r are independent and that d_i has mean μ_{a_i} and variance η^2 for $i = 1, \dots, r$. If we assume that the $e_{ij(k)}$ terms in model (2.1) are normally distributed, it follows that

$$d_i \sim N(\mu_{a_i}, \eta^2),$$

and that a standard two-sample t -test can be used to test the null hypothesis

$$H_0 : \mu_1 = \mu_2.$$

In this case, the differences from slides with dye assignment 1 form one sample while the differences from slides with dye assignment 2 form the other. Note that

$$H_0 : \mu_1 = \mu_2 \text{ is equivalent to } H_0 : \tau_1 = \tau_2.$$

Thus, the two-sample t -test provides a test for differential expression.

The test of $H_0 : \tau_1 = \tau_2$ based on direct linear model analysis of the data in Equation (2.1) is identical to the two-sample t -test of $H_0 : \tau_1 = \tau_2$ based on the differences, provided that the slide effects are modeled as fixed. If the slide

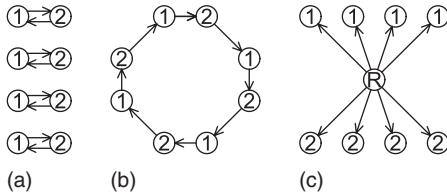


Figure 2.2 Alternative designs for comparing two treatments.

effects are modeled as random, then the results of the two analyses will be identical as long as the REML estimate (see HK2, section 1.11.2) of σ_s^2 is strictly positive. In either case, the best linear unbiased estimator of $\tau_1 - \tau_2$ is

$$\widehat{\tau_1 - \tau_2} \equiv (\bar{d}_2 - \bar{d}_1)/2, \text{ where } \bar{d}_\ell \equiv \sum_{\{i: a_i = \ell\}} d_i / r_\ell \text{ for } \ell = 1, 2.$$

Note that

$$\text{var}(\widehat{\tau_1 - \tau_2}) = \eta^2 (1/r_1 + 1/r_2)/4.$$

Thus, using a dye-balanced design ($r_1 = r_2$) will minimize $\text{var}(\widehat{\tau_1 - \tau_2})$ whenever r is even. If r is odd, choosing $r_1 = \lfloor r/2 \rfloor$ and $r_2 = \lfloor r/2 \rfloor + 1$ (or vice versa) will be most efficient.

Three alternative designs that have been proposed for measuring experimental units in two-treatment two-color microarray experiments are depicted in Figure 2.2. The design Figure 2.2a is an example of what is often referred to as a *dye swap design*. Each pair of experimental units is measured on two slides, once with each of the two dye assignments. This design is a good choice if the number of available experimental units matches the number of available two-color microarray slides. However, it is often the case that the cost of a microarray slide is substantially more than the cost of an experimental unit. This makes the number of microarray slides the limiting factor when determining the size of the experiment. Dobbin, Shih, and Simon (2003) proved that measuring each pair of experimental units on only one slide minimizes the variance of the best linear unbiased estimator of $\tau_1 - \tau_2$ for any fixed number of slides. Thus, for a given number of microarray slides, it is better to maximize the number of experimental units by measuring each pair of experimental units on only one slide as in Figure 2.1.

Kerr (2003) described the design in Figure 2.2b as an *alternating loop design* and showed that the variance of the best linear unbiased estimator of $\tau_1 - \tau_2$ is the same for this design and the dye swap design in Figure 2.2a. Jarrett and Ruggiero (2008) argued that although the variances of the best linear unbiased estimators are the same, the variance can typically be better estimated when using the dye swap design in Figure 2.2a than when using the alternating loop design in Figure 2.2b. Furthermore, the dye swap design in Figure 2.2a is clearly

preferred over the alternating loop design in Figure 2.2b if the first phase of the experiment is an RCB. In this case, two dye-swapped slides should be used to measure the experimental units within each phase 1 block. For this design, it can be shown that the natural linear model-based test for differential expression is equivalent to a paired-data *t*-test, using the average of the log-scale expression measurements for each experimental unit as individual data points and the phase 1 blocks to define pairs.

The design in Figure 2.2c is known as a *reference design*. The single circle labeled with R in the center of the Figure 2.2c represents a reference pool of nucleic acid to which each experimental unit is compared on a single slide. The reference design has been popular among practitioners due to its simplicity. By subtracting the log-scale expression measurement of the reference pool from the log-scale expression measurement for each experimental unit, a set of data can be obtained for which slide effects have been removed and other nuisance effects (dye and the reference pool expression level) are constant across observations. This allows for a simple analysis strategy that depends on only the design of the first phase of the microarray experiment. However, many statistical researchers have pointed out the inefficiency of the reference design for identifying differentially expressed genes. Common sense suggests that it is unwise to take 50% of all measurements on a reference pool that is not of direct interest. For more discussion of the reference design and variations and formal comparisons with more efficient alternatives, see, for example, Kerr and Churchill (2001), Churchill (2002), Yang and Speed (2002), Dobbin, Shih, and Simon (2003), Kerr (2003), Tempelman (2005), and Altman and Hua (2006).

The designs in Figure 2.2 require one microarray slide for each experimental unit. As noted above, a microarray slide is often more expensive than an experimental unit. Thus, it is common to have more experimental units than microarray slides. The design in Figure 2.1 is appropriate when there are twice as many experimental units as microarray slides. What if more experimental units are available? Kendziorski et al. (2003) recommended forming multiple independent pools for each treatment, where each pool is obtained by combining samples from experimental units treated alike. They provided formulas that show efficiency advantages of this approach over measuring individual experimental units when the number of microarray slides is fixed but the number of experimental units per treatment is not. However, the derivations of Kendziorski et al. (2003) treated pooling as mathematically similar to averaging log-scale expression values. Zhang et al. (2007) argued that pooling is more similar to randomly weighted averaging of original-scale expression levels. Their calculations indicate that pooling can be advantageous but not as advantageous as the formulas of Kendziorski et al. (2003) suggest.

In summary, a two-treatment two-color microarray experiment is best conducted using a dye-balanced design in which independent pairs of differently treated experimental units are hybridized together to microarray slides as depicted in Figure 2.1. If the number of slides is fixed but more experimental

units can be obtained, then each individual experimental unit in Figure 2.1 can be replaced by a pool of experimental units treated alike. However, the number of experimental units in each pool should be the same for all pools so that homogeneous variance models can be used for analysis. Regardless of whether individual experimental units or pools are used, a standard two-sample t -test with slide-specific log-scale expression differences (red–green), as data can be used to test for differential expression. If the number of experimental units is fixed and one microarray slide is available for each experimental unit, then the dye swap design in Figure 2.2a is a good option, particularly when the first phase of the experiment is conducted as an RCBD. In this case, it is straightforward to test for differential expression using a paired-data t -test.

2.6 TWO-COLOR MICROARRAY EXPERIMENTS INVOLVING MORE THAN TWO TREATMENTS

In this section, we will focus on the design of two-color microarray experiments involving $t > 2$ treatments. We will assume that the first phase of the experiment is conducted using a CRD or an incomplete block design with blocks of size 2. In the latter case, all the designs we consider will confound incomplete blocks from the first phase of the experiment with microarray slides from the second phase. The advantage of this strategy was discussed for the case of two treatments in Section 2.5. Phase 2 design for more complex phase 1 designs is discussed in Section 2.8. Furthermore, we will assume that the total number of experimental units is equal to twice the number of two-color microarray slides so that each experimental unit can be measured exactly once. Under these conditions, the task at hand is to determine the best strategy for pairing experimental units on slides and determining dye assignments for each slide.

As in the two-treatment case, we will consider Equation (2.1), except that in the t -treatment case, k ranges from 1 to t . A gene is considered differentially expressed if

$$H_0 : \tau_1 = \dots = \tau_t,$$

is false, where τ_1, \dots, τ_t represent the effects associated with the t treatments. We will assume that pairwise comparison of treatments is of interest and that all possible comparisons are of equal interest unless otherwise noted.

In the two-treatment case, the recommended design was the Latin rectangle design depicted in Figure 2.1. When $t > 2$, a Latin rectangle design cannot be used because the number of dyes is fixed at two. (Although using more than two dyes is conceptually possible, technical challenges have prevented common use of more than two dyes.) However, it is still possible to control for two sources of heterogeneity (microarray slides and dyes) by using an extended incomplete Latin square design (see HK1, section 10.5). The extended incom-

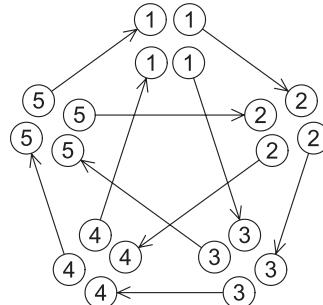


Figure 2.3 An interwoven loop design for all pairwise comparisons among five treatments.

plete Latin square designs useful for two-color microarray experiments have $r = \alpha t$ rows (α integer) corresponding to microarray slides and two columns corresponding to the two dyes. Each treatment appears α times in each column, which guarantees that the design will be dye-balanced. The rows of the design form a BIBD, which ensures that each pair of treatments can be compared with equal precision.

An example of such a design is depicted in Figure 2.3 for the case of $t = 5$, $\alpha = 2$, and $r = 10$. Each of the $\binom{5}{2} = 10$ pairwise comparisons of treatments occurs on a single slide. This is an example of an *interwoven loop design* proposed by Kerr and Churchill (2001), and found to be at least near optimal in a wide variety of situations by Wit, Nobile, and Khanin (2005). The name “interwoven loop” comes from the two loops visible in the design: an outer loop that circles through the treatments in the order 1,2,3,4,5, and back to 1, and an inner loop that moves through the treatments in the order 1,3,5,2,4, and back to 1. Kerr and Churchill (2001) recommended the design in Figure 2.3 as the design that minimizes the average of the variances of all $\binom{t}{2}$ estimates of treatment differences for the case of $t = 5$ treatments and $r = 10$ slides. Kerr and Churchill (2001) referred to this as an *A*-optimal design; however, as pointed out by Wit, Nobile, and Khanin (2005), it would be more accurate to describe this as a design that is *L*-optimal for all pairwise treatment comparisons. (See also HK2, section 1.13 and Pukelsheim 1993 for definitions and discussion of various design optimality criteria.)

Kerr and Churchill (2001) also presented designs *L*-optimal for all pairwise comparisons of treatments for $t = 6, \dots, 10$ and $r = 2t$, as well as designs *L*-optimal for all pairwise comparisons of treatments among designs with an even number of experimental units per treatment for $t = 6, \dots, 13$ and $r = t + 2$. Using Euler’s result that a connected graph has a circuit that traverses every edge exactly once if and only if the degree of every node in the graph is even, Kerr and Churchill (2001) pointed out that having an even number of experimental units per treatment is a necessary and sufficient condition for dye balance to be possible.

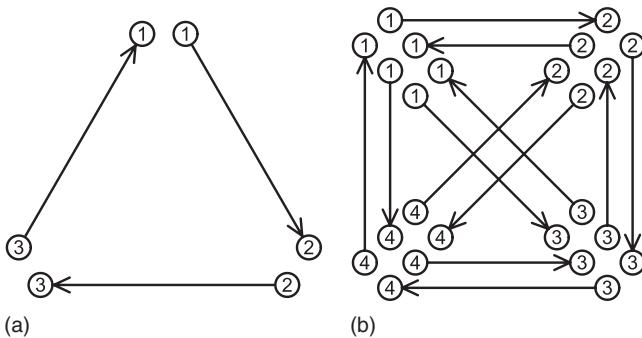


Figure 2.4 Extended incomplete Latin square designs for $t = 3$ (a) and $t = 4$ (b) treatments.

Aside from the design depicted in Figure 2.3, none of the designs proposed by Kerr and Churchill (2001) are extended incomplete Latin square designs. Figure 2.4 shows extended incomplete Latin square designs for the cases of $t = 3$ and $t = 4$. The design for $t = 3$ is an example of a *loop design* proposed by Kerr and Churchill (2001) for the case where the number of treatments t equals the number of slides r . For $t = r > 3$, loop designs clearly do not have a BIBD property because only treatments with adjacent numeric labels (along with treatments 1 and t) are compared on a slide. However, as pointed out by Kerr and Churchill (2001), loop designs are an attractive choice when the levels of treatment correspond to time points. In such cases, it is natural to focus the power of the design on identifying changes between adjacent time points. The extended incomplete Latin square design for the case of $t = 4$ includes two loops through the treatments 1 through 4 (one clockwise and one counterclockwise), along with comparisons between treatments 2 and 4 and between treatments 1 and 3 to satisfy the BIBD property. Alternatively, this design can be viewed as a series of four loops, each involving three treatments: 1,2,3; 2,1,4; 3,4,1; 4,3,2.

The designs depicted in Figure 2.4 illustrate a basic strategy for deriving dye-balanced extended incomplete Latin square designs for $t > 4$. If t is odd as in Figure 2.4a, one slide can be used for each of the $\binom{t}{2}$ treatment comparisons. This will result in an even number $(t - 1)$ of experimental units for each treatment. By Euler's result, dye balance for the entire design can be achieved. If t is even as in Figure 2.4b, two slides can be used for each of the $\binom{t}{2}$ treatment comparisons. This will result in an even number $(2t - 2)$ of experimental units for each treatment so that dye balance can be guaranteed. Whether t is even or odd, any of the dye-balanced extended incomplete Latin square designs can be repeated as many times as resources permit to obtain larger dye-balanced extended incomplete Latin square designs. For example, the entire design depicted in Figure 2.4a could be repeated by looping in the

opposite direction to obtain a dye-balanced extended incomplete Latin square design with four, rather than two, replications per treatment. However, note that even the smallest dye-balanced extended incomplete Latin square designs may be quite expensive for large t , especially when t is even.

2.7 MULTIFACTOR TWO-COLOR MICROARRAY EXPERIMENTS

2.7.1 Introduction

Section 2.6 considered multiple-treatment designs in which all pairwise comparisons among the levels of the treatment factor were of equal interest. In this section, we consider multiple-treatment designs where treatments are defined by combinations of levels from two or more factors. In such multifactor experiments, it is often the case that only a subset of all possible treatment comparisons is of interest. Furthermore, not all the treatment comparisons of interest may be equally important to the investigators. In the following subsections, we review three different strategies for evaluating candidate two-color microarray designs for such experiments.

2.7.2 Admissible Designs

Glonk and Solomon (2004) proposed the consideration of *admissible designs* for multifactor two-color microarray experiments. We introduce their main ideas using the following hypothetical example.

Suppose researchers are interested in studying the gene expression response of maize plants to a virus infection under drought conditions. A balanced, completely randomized design is used to assign 12 plants to the four combinations of infection type (mock infection versus infection with a virus) and watering protocol (simulated drought conditions versus normal). Table 2.1 provides notation for the log-scale expression means corresponding to each of the four treatments in this completely randomized, two-factor experiment. For simplicity, dye effects have been excluded from the means in Table 2.1, but these will be considered subsequently.

Table 2.1 Log-Scale Expression Means for the Four Treatments

Treatment	Infection Type	Watering Protocol	Mean
1	Mock	Drought	$\mu + \tau_1$
2	Virus	Drought	$\mu + \tau_2$
3	Mock	Normal	$\mu + \tau_3$
4	Virus	Normal	$\mu + \tau_4$

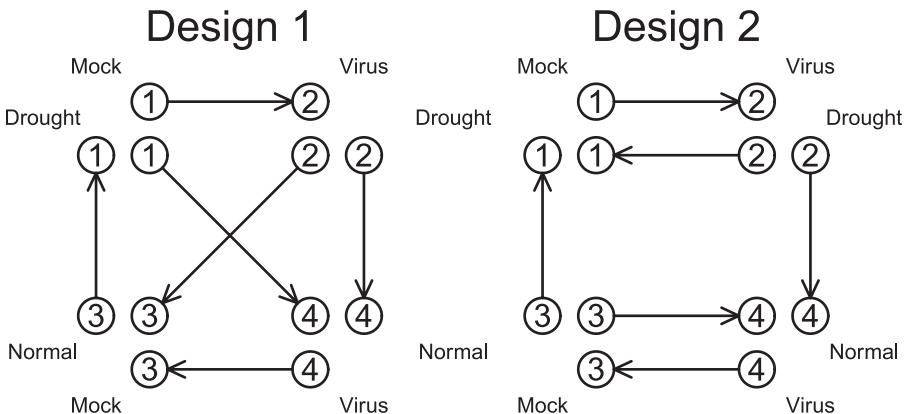


Figure 2.5 Candidate designs for an example two-factor experiment.

Suppose past research indicates that maize plants show no adverse effects of virus infection under normal conditions but show clear signs of viral damage under drought conditions. Based on this observation, the researchers are primarily interested in understanding gene expression differences between mock-infected and virus-infected plants under drought conditions, and in identifying genes whose change in expression in response to virus infection differs between normal and drought conditions. In other words, the researchers wish to test for a simple effect of the factor infection type under drought conditions, and for interaction between the factors watering protocol and infection type. The null hypotheses in terms of the mean parameters in Table 2.1 corresponding to these tests of interest are

$$H_{01} : \tau_1 - \tau_2 = 0 \text{ and } H_{02} : \tau_1 - \tau_2 - \tau_3 + \tau_4 = 0.$$

Figure 2.5 shows two possible designs for measuring expression in the 12 experimental units using six two-color microarray slides. Given the researchers' objectives, which of these designs is most appropriate? To address this question, we will consider an analysis using the red-green log-scale expression difference from each slide as the response as in Section 2.5. For each design $i = 1, 2$, let \mathbf{d}_i denote the response vector of differences, \mathbf{X}_i denote the design matrix, and \mathbf{e}_i denote the vector of independent, mean zero, constant variance ($\eta^2 > 0$) errors. We can write a linear model for the data from design i as

$$\mathbf{d}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{e}_i, \quad (2.2)$$

where $\boldsymbol{\beta} = (\delta_2 - \delta_1, \tau_1, \tau_2, \tau_3, \tau_4)'$ is a vector of unknown parameters whose entries are the difference between dye effects ($\delta_2 - \delta_1$) along with parameters defined in Table 2.1. Note that because all the observations in the response vector involve differences between two treatments, the mean of the response vector

does not involve the parameter μ in Table 2.1, and this parameter is not included in the vector β . The design matrices corresponding to designs 1 and 2 in Figure 2.5 can be written as

$$\mathbf{X}_1 = \begin{bmatrix} 1 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & 1 & -1 \\ 1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 1 \\ 1 & 0 & -1 & 1 & 0 \end{bmatrix} \text{ and } \mathbf{X}_2 = \begin{bmatrix} 1 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & 1 & -1 \\ 1 & 1 & 0 & -1 & 0 \\ 1 & 1 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 1 \end{bmatrix}.$$

For design $i = 1, 2$, the variance of the best linear unbiased estimator of any estimable linear combination $\mathbf{c}'\beta$ is $\eta^2 \mathbf{c}'(\mathbf{X}'_i \mathbf{X}_i)^{-1} \mathbf{c}$ (see HK1, section 4.16.2). Because our interest centers on the estimable linear combinations $(0, 1, -1, 0, 0)'\beta = \tau_1 - \tau_2$ and $(0, 1, -1, -1, 1)'\beta = \tau_1 - \tau_2 - \tau_3 + \tau_4$, we should examine

$$v_{ij} = \mathbf{c}'_j (\mathbf{X}'_i \mathbf{X}_i)^{-1} \mathbf{c}_j$$

for $i = 1, 2$ and $j = 1, 2$, where $c_1 = (0, 1, -1, 0, 0)'$ and $c_2 = (0, 1, -1, -1, 1)'$. Straightforward calculations show that

$$v_{11} = 0.5 > 0.4375 = v_{21} \text{ and } v_{12} = 1.0 > 0.75 = v_{22}.$$

Thus, design 2 produces a lower variance estimator than design 1 for each of the estimable linear combinations of interest. Because design 2 uses the same number of slides as design 1 to achieve superior estimates of the parameters of interest, design 1 can be excluded from further consideration as a potential microarray design for the second phase of the experiment. In the terminology of Glonek and Solomon (2004), design 1 is said to be *inadmissible*.

In general, a design using r slides and design matrix \mathbf{X} is said to be *admissible* if there exists no other design with r slides and design matrix \mathbf{X}_* such that $v_j \geq v_j^*$ for all $j = 1, \dots, J$ with strict inequality for at least one j , where

$$v_j = \mathbf{c}'_j (\mathbf{X}' \mathbf{X})^{-1} \mathbf{c}_j \text{ and } v_j^* = \mathbf{c}'_j (\mathbf{X}'_* \mathbf{X}_*)^{-1} \mathbf{c}_j,$$

and $\mathbf{c}'_1 \beta, \dots, \mathbf{c}'_J \beta$ denote all estimable linear combinations of interest. A design that is not admissible is said to be *inadmissible*. Glonek and Solomon (2004) recommended choosing a design that is admissible among the class of designs that permit estimation of all linear combinations of interest and meet restrictions on the number of experimental units and the number of slides set by the experimenter.

In our hypothetical example, we assume that the number of experimental units per treatment is fixed at 3 and the number of microarray slides is fixed

at $r = 6$. Neither design 1 nor design 2 is admissible among the class of designs that use three experimental units for each of the four treatment and six slides. An admissible design can be obtained by replacing the two vertical arrows in design 2 with an additional horizontal arrow from treatment 1 to treatment 2 and an additional horizontal arrow from treatment 3 to treatment 4. This design sacrifices the ability to estimate the simple and main effects of watering protocol to obtain greater precision for estimating infection type effects and the interaction between watering protocol and infection type.

Glonek and Solomon (2004) provided tables of admissible designs for various scenarios, but placed special emphasis on two-factor experiments in which each factor has two levels, as in the example presented here. In contrast to our presentation, Glonek and Solomon (2004) did not consider the possibility of gene-specific dye effects. This simplifies their presentation to some extent and makes it easier to generate an exhaustive list of admissible designs for a given scenario.

2.7.3 w -Optimal Designs

Banerjee and Mukerjee (2007) proposed a design criterion they called w -optimality for multifactor experiments in which each factor has a designated baseline level. They assumed that the goal of the analysis is to estimate all interactions and to make all possible pairwise comparisons between the baseline level of a factor and other levels of that factor while holding all other factors at their baseline levels. Like Glonek and Solomon (2004), they placed special emphasis on two-factor experiments in which each factor has two levels.

To illustrate their ideas in the context of our simple example summarized in Table 2.1, suppose mock is the baseline level of the factor infection type and that drought is the baseline level of the factor watering protocol. Then the estimable quantities of interest are assumed to be $\tau_1 - \tau_2$, $\tau_1 - \tau_3$, and $\tau_1 - \tau_2 - \tau_3 + \tau_4$. Denote the best linear unbiased estimates of these quantities by $\hat{\theta}_{12}$, $\hat{\theta}_{13}$, and $\hat{\theta}_{1234}$, respectively. Banerjee and Mukerjee (2007) defined a design to be w -optimal if it minimizes

$$\text{var}(\hat{\theta}_{12}) + \text{var}(\hat{\theta}_{13}) + w \cdot \text{var}(\hat{\theta}_{1234}),$$

among designs in a candidate set, where w is a specified positive weight. They used approximate design theory (see e.g., Silvey 1980) to guide the choice of designs that are at least near optimal for given $w > 0$ and number of slides r .

Banerjee and Mukerjee (2007) also considered w -optimality for general factorial experiments. They provided optimal designs for the saturated case in which the number of slides is sufficient to estimate all linear combinations of interest but insufficient to estimate the error variance. They presented strategies for augmenting optimal saturated designs to obtain at least near optimal designs for nearly saturated cases.

2.7.4 e -Efficiency

Rather than considering admissibility or w -optimality, Landgrebe, Bretz, and Brunner (2006) proposed the use of the e -efficiency criterion for selecting the best design among a given set of candidate designs. To define the e -efficiency criterion, it is necessary to introduce some additional notation. First, let X_1, \dots, X_m denote the design matrices associated with a set of m candidate designs, and suppose that Equation (2.2) holds for a parameter vector β of the form $(\delta_2 - \delta_1, \tau_1, \dots, \tau_t)'$, where τ_1, \dots, τ_t are the effects associated with the t treatments formed by combinations of factor levels in a multifactor experiment. Next, suppose $\mathbf{C}_1'\beta, \dots, \mathbf{C}_p'\beta$ are vectors of scientific interest that are each estimable for each of the candidate designs. For $j = 1, \dots, p$, we will assume that the first row of \mathbf{C}_j is $\mathbf{0}$, and that $\mathbf{C}_j'\mathbf{1} = \mathbf{0}$, so that each column of \mathbf{C}_j is a contrast of treatment effects.

As an example, consider again our hypothetical two-factor experiment involving the factors infection type and watering protocol. Suppose that in addition to the simple effect of the factor infection type under drought conditions and the interaction between the factors watering protocol and infection type, we are also interested in testing for equality of all four treatment means. Then scientific interest rests on the vectors $\mathbf{C}_1'\beta$, $\mathbf{C}_2'\beta$, and $\mathbf{C}_3'\beta$, where

$$\mathbf{C}_1' = [0, 1, -1, 0, 0], \mathbf{C}_2' = [0, 1, -1, -1, 1], \text{ and } \mathbf{C}_3' = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \end{bmatrix}. \quad (2.3)$$

Now, for $i = 1, \dots, m$ and $j = 1, \dots, p$, let

$$e_j(\mathbf{X}_i) = \lambda_{\max}(\mathbf{C}_j'(\mathbf{X}_i)^{-1}\mathbf{C}_j)/\text{tr}(\mathbf{C}_j'\mathbf{C}_j),$$

where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue function, and $\text{tr}(\cdot)$ denotes the trace function. The quantity $e_j(\mathbf{X}_i)$ represents the e -efficiency of design i with respect to the j th estimable vector. Note that $\eta^2 e_j(\mathbf{X}_i)$ is simply the variance of the best linear unbiased estimator of $\mathbf{C}_j'\beta$ when \mathbf{C}_j is a contrast vector standardized to unit length. When \mathbf{C}_j is a matrix with multiple columns corresponding to uncorrelated contrasts, $\eta^2 e_j(\mathbf{X}_i)$ is the variance of the most variable contrast divided by the sum of all squared contrast coefficients. More generally, $\eta^2 e_j(\mathbf{X}_i)\text{tr}(\mathbf{C}_j'\mathbf{C}_j)$ measures the maximum variance of the best linear unbiased estimate of $\mathbf{a}'\mathbf{C}_j'\beta$ over all nonrandom, unit-length vectors \mathbf{a} . Division by the trace is included in the definition of $e_j(\mathbf{X}_i)$ so that $e_j(\mathbf{X}_i)$ remains unchanged if \mathbf{C}_j is multiplied by a nonzero scalar.

To permit the comparison of multiple designs with respect to their ability to estimate multiple vectors of interest, Landgrebe, Bretz, and Brunner (2006) recommended computing

$$e_j^{\text{rel}}(\mathbf{X}_i) = \frac{e_j^{-1}(\mathbf{X}_i)}{\max\{e_j^{-1}(\mathbf{X}_k) : k = 1, \dots, m\}},$$

for each candidate design $i = 1, \dots, m$ and each vector of interest $\mathbf{C}_j'\beta, j = 1, \dots, p$. Note that $e_j^{\text{rel}}(\mathbf{X}_i)$ is a measure of the efficiency of design i for estimation of $\mathbf{C}_j'\beta$ relative to the efficiency of the candidate design most efficient for estimating $\mathbf{C}_j'\beta$. The candidate design with the highest relative efficiency averaged across the p vectors of interest is recommended as the most efficient design in the candidate set for the p vectors of interest; that is, the most efficient design matrix is defined as \mathbf{X}_{i^*} , where

$$i^* \equiv \arg \max_{i=1, \dots, m} \frac{1}{p} \sum_{j=1}^p e_j^{\text{rel}}(\mathbf{X}_i).$$

Returning again to our hypothetical example, suppose designs 1 and 2 in Figure 2.5 comprise the set of candidate designs for estimating $\mathbf{C}_1'\beta$, $\mathbf{C}_2'\beta$, and $\mathbf{C}_3'\beta$, where \mathbf{C}_1 , \mathbf{C}_2 , and \mathbf{C}_3 are defined in model 3. Straightforward computations give approximate $e_j(\mathbf{X}_i)$ and $e_j^{\text{rel}}(\mathbf{X}_i)$ values as follows:

	Approximate $e_j(\mathbf{X}_i)$ Values			Approximate $e_j^{\text{rel}}(\mathbf{X}_i)$ Values			
	\mathbf{C}_1	\mathbf{C}_2	\mathbf{C}_3		\mathbf{C}_1	\mathbf{C}_2	\mathbf{C}_3
\mathbf{X}_1	0.2500	0.2500	0.1782	\mathbf{X}_1	0.875	0.750	1.000
\mathbf{X}_2	0.2188	0.1875	0.2214	\mathbf{X}_2	1.000	1.000	0.805.

Because the average of the $e_j^{\text{rel}}(\mathbf{X}_2)$ values is larger than the average of the $e_j^{\text{rel}}(\mathbf{X}_1)$ values, design 2 is more efficient than design 1 according to the criterion of Landgrebe, Bretz, and Brunner (2006).

2.8 PHASE 2 DESIGNS FOR COMPLEX PHASE 1 DESIGNS

In previous sections, we have assumed that the phase 1 experimental design is either a completely randomized design or a randomized complete or incomplete block design with blocks of size 2. Such phase 1 designs are natural choices for two-color microarray experiments where the block size is 2 in the second phase of the design. However, it is sometimes necessary to use blocks of size greater than 2 or some other more complex design during the first phase of a microarray experiment. We focus this section on a simple example to illustrate the design and analysis challenges that can arise when the phase 1 designs differs from those we have considered previously.

Suppose a randomized complete block design with four blocks and three treatments is used in the first phase of the microarray experiment. Suppose six two-color microarray slides are available for measuring the 12 experimental units. If the goal is to conduct all pairwise comparisons of the three treatments, how should we pair experimental units on slides and assign dyes? Figure 2.6 shows one possible design. If we focus only on the microarray slides depicted in Figure 2.6, we see that the design is an incomplete block design in

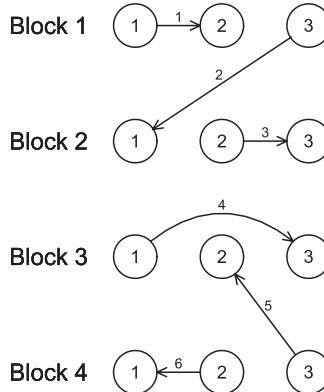


Figure 2.6 A candidate phase 2 design for a phase 1 RCBD.

which each pair of treatments is compared on two slides. The design is dye-balanced in that two experimental units are measured with the red dye and two with the green dye within each treatment. However, the presence of the phase 1 blocks makes the overall design and analysis more complicated than the designs discussed in the previous sections. Given the resource constraints, there is an unavoidable asymmetry in the way that pairs of treatments are compared. Although each pair of treatments is compared on exactly two slides, each of the comparisons of treatments 1 and 2 involve experimental units from within a single phase 1 block. In contrast, one comparison of treatments 1 and 3 and one comparison of treatments 2 and 3 involve experimental units from different phase 1 blocks. As a result, we will have greater precision for estimating the difference between treatments 1 and 2 than for estimating differences between the other two pairs of treatments.

How should we analyze the data from the design depicted in Figure 2.6? First, consider the case where both block and slide are assumed to be fixed factors. As discussed in Sections 2.5 and 2.7, it is customary in microarray data analysis to use the red–green log-scale expression difference from each slide as the response. Let d_i denote the slide i difference for $i = 1, \dots, 6$, where without loss of generality, slides are assumed to be numbered as in Figure 2.6. A model for these differences is given by

$$\begin{aligned}
 d_1 &= \delta_2 - \delta_1 + \tau_2 - \tau_1 + e_1 \\
 d_2 &= \delta_2 - \delta_1 + \tau_1 - \tau_3 + \beta_2 - \beta_1 + e_2 \\
 d_3 &= \delta_2 - \delta_1 + \tau_3 - \tau_2 + e_3 \\
 d_4 &= \delta_2 - \delta_1 + \tau_3 - \tau_1 + e_4 \\
 d_5 &= \delta_2 - \delta_1 + \tau_2 - \tau_3 + \beta_3 - \beta_4 + e_5 \\
 d_6 &= \delta_2 - \delta_1 + \tau_1 - \tau_2 + e_6,
 \end{aligned}$$

where δ_1 and δ_2 are effects for the green and red dyes, respectively; τ_1 , τ_2 , τ_3 are the treatment effects; β_1 , β_2 , β_3 , β_4 are the phase 1 block effects; and e_1, \dots, e_6 are independent and identically distributed mean-zero random variables with variance $\eta^2 > 0$.

It is straightforward to show that the best linear unbiased estimators of $\tau_1 - \tau_2$, $\tau_1 - \tau_3$, and $\tau_2 - \tau_3$ are

$$\widehat{\tau_1 - \tau_2} = (-2d_1 + 0d_2 + 1d_3 - 1d_4 + 0d_5 + 2d_6)/5 \quad (2.4)$$

$$\widehat{\tau_1 - \tau_3} = (3d_1 + 0d_2 - 4d_3 - 6d_4 + 0d_5 + 7d_6)/10 \quad (2.5)$$

$$\widehat{\tau_2 - \tau_3} = (7d_1 + 0d_2 - 6d_3 - 4d_4 + 0d_5 + 3d_6)/10, \quad (2.6)$$

with variances

$$\text{var}(\widehat{\tau_1 - \tau_2}) = 0.4\eta^2, \text{var}(\widehat{\tau_1 - \tau_3}) = 1.1\eta^2, \text{and } \text{var}(\widehat{\tau_2 - \tau_3}) = 1.1\eta^2. \quad (2.7)$$

If we assume normality of the model errors, the REML estimator of η^2 is given by

$$\hat{\eta}^2 = (1d_1 + 0d_2 + 2d_3 - 2d_4 + 0d_5 - 1d_6)^2 / 10, \quad (2.8)$$

and inferences about treatment differences could, in theory, be conducted by noting that

$$t_{ij} = \frac{\widehat{\tau_i - \tau_j} - (\widehat{\tau_i - \tau_j})}{\sqrt{\text{var}(\widehat{\tau_i - \tau_j})}},$$

has a t distribution for $i \neq j$, where

$$\widehat{\text{var}}(\widehat{\tau_i - \tau_j}) = (\eta^2 / \hat{\eta}^2) \text{var}(\widehat{\tau_i - \tau_j}).$$

Unfortunately, this t distribution has only a single degree of freedom. Thus, this analysis strategy cannot be recommended in practice.

Note that the presence of the block effects in d_2 and d_5 makes these observations useless for inferences regarding treatment effects if block effects are considered fixed. This is easily seen by examining the best linear unbiased estimates of $\tau_i - \tau_j$ ($i \neq j$) and the REML estimate of η^2 in Equations (2.4)–(2.6) and (2.8). To make use of all the data and to obtain additional degrees of freedom for error, it is desirable to model block or slide or both as random factors. The analysis strategy described above using the red–green log-scale expression difference from each slide as the response remains valid but is inefficient when block or slide or both are random factors. A linear mixed model analysis using the original normalized log-scale expression values as the response allows for recovery of interblock information (see

HK2, sections 1.7–1.11) and can provide more efficient estimates of treatment differences.

Consider the linear mixed effects model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_b\mathbf{b} + \mathbf{Z}_s\mathbf{s} + \mathbf{e}, \quad (2.9)$$

where \mathbf{y} is the response vector with entries ordered first by block and then by treatment, $\boldsymbol{\beta} = (\mu, \delta_1, \delta_2, \tau_1, \tau_2, \tau_3)'$ is a vector of fixed effects, $\mathbf{b} = (b_1, b_2, b_3, b_4)'$ is a vector of random block effects, $\mathbf{s} = (s_1, \dots, s_6)'$ is a vector of random slide effects, $\mathbf{e} = (e_1, \dots, e_{12})'$ is a vector of random errors, and \mathbf{X} , \mathbf{Z}_b , and \mathbf{Z}_s are the matrices

$$\begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \text{ and } \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix},$$

respectively. As is customary, we assume that $(\mathbf{b}, \mathbf{s}, \mathbf{e})'$ has a multivariate normal distribution with mean $\mathbf{0}$ and variance–covariance matrix

$$\begin{bmatrix} \sigma_b^2 \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_s^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_e^2 \mathbf{I} \end{bmatrix},$$

where $\mathbf{0}$ and \mathbf{I} denote zero and identity matrices, respectively, of appropriate dimensions, and σ_b^2 , σ_s^2 , and σ_e^2 denote unknown variance components.

Let $\hat{\sigma}_b^2$, $\hat{\sigma}_s^2$, and $\hat{\sigma}_e^2$ denote the REML estimators for σ_b^2 , σ_s^2 , and σ_e^2 , respectively. Denote the estimated variance of the response vector \mathbf{y} by

$$\widehat{\text{var}}(\mathbf{y}) = \hat{\mathbf{V}} = \hat{\sigma}_b^2 \mathbf{Z}_b \mathbf{Z}'_b + \hat{\sigma}_s^2 \mathbf{Z}_s \mathbf{Z}'_s + \hat{\sigma}_e^2 \mathbf{I}.$$

Although the best linear unbiased estimator of $\tau_i - \tau_j$ ($i \neq j$) depends on the unknown variance components, it can be approximated by

$$\mathbf{c}'_{ij} \hat{\beta}, \text{ where } \hat{\beta} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y},$$

$\mathbf{c}_{12} = (0, 0, 0, 1, -1, 0)'$, $\mathbf{c}_{13} = (0, 0, 0, 1, 0, -1)'$, and $\mathbf{c}_{23} = (0, 0, 0, 0, 1, -1)'$. Approximate inferences regarding treatment differences can be obtained using the method of Kenward and Roger (1997).

We illustrate the analysis described above with the following numerical example.

Example 2.1. SAS PROC MIXED code and output is provided in Table 2.2 for the analysis of hypothetical data from a single gene from the experiment depicted in Figure 2.6. The first set of PROC MIXED commands treats block and slide as fixed factors and yields inferences for treatment differences identical to those that would be obtained by the analysis of red–green log-scale expression differences described previously. The second set of PROC MIXED commands performs the analysis based on the linear mixed model described in Equation (2.9). In this case, the approximate t -tests associated with the linear mixed model analysis suggest significant differences between treatments 1 and 3 and treatments 2 and 3. In contrast, the analysis that treats both block and slide as fixed factors yields no compelling evidence of treatment differences.

Unfortunately, both sets of results must be interpreted with caution given the relatively small sample size available for fitting complex models. Note that the RCBD used in the phase 1 portion of the experimental design would provide 6 degrees of freedom for estimating error variance if each experimental unit were individually measured. However, the necessity of measuring pairs of experimental units together on slides with two dyes in the second phase of the experiment leads to a substantial loss in degrees of freedom. This illustrates one important lesson about the design of two-color microarray experiments and two-phase designs in general. What might normally be considered an adequately replicated design in phase 1 may prove to be inadequate once factors introduced at the second phase of the design are accounted for. Thus, the design used for the first phase of the experiment should be selected with the phase 2 design in mind.

As discussed previously, using an incomplete block design with blocks of size 2 in the phase 1 portion of the experiment is a natural strategy for two-color microarray experiments. However, restrictions on available experimental units sometimes dictate block sizes other than 2. Note that there is some advantage to an even block size over an odd block size. For example, as an alternative to the design in Figure 2.6, consider the experiment depicted in Figure 2.7. The design in Figure 2.7 uses three blocks of size 4—that is, an extended block design (see HK1, section 9.8.5)—rather than four blocks of size 3 in the phase 1 design. If we ignore the second phase of the experiment, we would certainly prefer the design in Figure 2.6 to the design in Figure 2.7.

Table 2.2 Analysis of Data from the Design Depicted in Figure 2.6

```

options nodate linesize=64 pageno=1;
data example;
input block trt slide dye y;
datalines;
1 1 1 1 12.35
1 2 1 2 16.47
1 3 2 1 6.96
2 1 2 2 8.71
2 2 3 1 6.31
2 3 3 2 4.46
3 1 4 1 11.10
3 2 5 2 14.20
3 3 4 2 9.17
4 1 6 2 7.60
4 2 6 1 5.51
4 3 5 1 2.42
;
run;

proc mixed data=example;
  class block trt slide dye;
  model y=dye trt block slide / e3;
  estimate '1 - 2' trt 1 -1 0;
  estimate '1 - 3' trt 1 0 -1;
  estimate '2 - 3' trt 0 1 -1;
run;

proc mixed data=example;
  class block trt slide dye;
  model y=dye trt / ddfm=kr;
  random block slide;
  estimate '1 - 2' trt 1 -1 0;
  estimate '1 - 3' trt 1 0 -1;
  estimate '2 - 3' trt 0 1 -1;
run;

```

The Mixed Procedure

Model Information

Data Set	WORK.EXAMPLE
Dependent Variable	Y
Covariance Structure	Diagonal
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Residual

(Continued)

Table 2.2 (Continued)

Class Level Information										
Class Levels Values										
block 4 1 2 3 4										
trt 3 1 2 3										
slide 6 1 2 3 4 5 6										
dye 2 1 2										
Dimensions										
Covariance Parameters 1										
Columns in X 16										
Columns in Z 0										
Subjects 1										
Max Obs Per Subject 12										
Number of Observations										
Number of Observations Read 12										
Number of Observations Used 12										
Number of Observations Not Used 0										
Covariance Parameter										
Estimates										
Cov Parm	Estimate									
Residual	0.2398									
Fit Statistics										
-2 Res Log Likelihood 4.4										
AIC (smaller is better) 6.4										
AICC (smaller is better) 10.4										
BIC (smaller is better) 4.4										
Type 3 Coefficients for dye										
Effect	block	trt	slide	dye	Row1					
Intercept										
dye				1	1					
dye				2	-1					
trt				1						
trt				2						

Table 2.2 (Continued)

trt	3
block	1
block	2
block	3
block	4
slide	1
slide	2
slide	3
slide	4
slide	5
slide	6
Type 3 Coefficients for trt	
Effect	block trt slide dye Row1 Row2
Intercept	
dye	1
dye	2
trt	1
trt	2
trt	3
block	1
block	2
block	3
block	4
slide	1
slide	2
slide	3
slide	4
slide	5
slide	6
Type 3 Coefficients for block	
Effect	block trt slide dye Row1 Row2
Intercept	
dye	1
dye	2
trt	1
trt	2
trt	3
block	1
block	2
block	3

(Continued)

Table 2.2 (Continued)

block	4		-1					
slide		1						
slide		2						
slide		3						
slide		4						
slide		5						
slide		6						
Type 3 Coefficients for slide								
Effect	block	trt	slide	dye	Row1	Row2	Row3	Row4
Intercept								
dye				1				
dye				2				
trt		1						
trt		2						
trt		3						
block	1							
block	2							
block	3							
block	4							
slide		1			1			
slide		2				1		
slide		3			-1	-1		
slide		4					1	
slide		5						1
slide		6					-1	-1
Type 3 Tests of Fixed Effects								
	Num	Den						
Effect	DF	DF	F	Value	Pr > F			
dye	1	1	40.20	0.0996				
trt	2	1	27.66	0.1332				
block	2	1	37.66	0.1145				
slide	4	1	4.01	0.3562				
Estimates								
	Standard							
Label	Estimate	Error	DF	t Value	Pr > t			
1 - 2	-0.7960	0.4380	1	-1.82	0.3202			
1 - 3	4.5970	0.7263	1	6.33	0.0998			
2 - 3	5.3930	0.7263	1	7.42	0.0852			

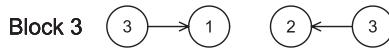
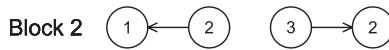
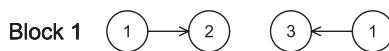
Table 2.2 (Continued)

The Mixed Procedure										
Model Information										
Data Set	WORK.EXAMPLE									
Dependent Variable	Y									
Covariance Structure	Variance Components									
Estimation Method	REML									
Residual Variance Method	Profile									
Fixed Effects SE Method	Kenward-Roger									
Degrees of Freedom Method	Kenward-Roger									
Class Level Information										
Class Levels Values										
block	4	1	2	3	4					
trt	3	1	2	3						
slide	6	1	2	3	4					
dye	2	1	2							
Dimensions										
Covariance Parameters	3									
Columns in X	6									
Columns in Z	10									
Subjects	1									
Max Obs Per Subject	12									
Number of Observations										
Number of Observations Read	12									
Number of Observations Used	12									
Number of Observations Not Used	0									
Iteration History										
Iteration	Evaluations	-2 Res Log Like	Criterion							
0	1	48.96915339								
1	2	37.22142990	0.06274607							
2	2	36.86472152	0.03825346							
3	1	36.29915729	0.02514148							
4	1	35.94717879	0.01146477							
5	1	35.79645289	0.00283385							
6	1	35.76199907	0.00022639							
7	1	35.75948284	0.00000183							
8	1	35.75946348	0.00000000							

(Continued)

Table 2.2 (Continued)

Convergence criteria met.								
Covariance Parameter Estimates								
Cov Parm Estimate								
block 8.5678								
slide 1.0303								
Residual 0.1773								
Fit Statistics								
-2 Res Log Likelihood 35.8								
AIC (smaller is better) 41.8								
AICC (smaller is better) 47.8								
BIC (smaller is better) 39.9								
Type 3 Tests of Fixed Effects								
Num Den								
Effect	DF	DF	F Value	Pr > F				
dye	1	3.12	55.82	0.0044				
trt	2	2.8	37.21	0.0096				
Estimates								
Standard								
Label	Estimate	Error	DF	t Value	Pr > t	t		
1 - 2	-0.7245	0.4047	1.99	-1.79	0.2158			
1 - 3	4.3281	0.5760	3.9	7.51	0.0019			
2 - 3	5.0525	0.5760	3.9	8.77	0.0010			

**Figure 2.7** A candidate phase 2 design when an extended block design is used in phase 1.

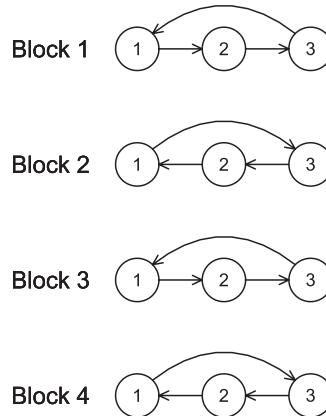


Figure 2.8 A candidate phase 2 design for a phase 1 RCBD when one slide is available for each experimental unit.

However, once we take the second phase of the design to account, the design in Figure 2.6 is inferior to the design in Figure 2.7 because, for the latter design, slide effects are nested within the phase 1 block effects and between-block comparisons are avoided. As a result, a simple analysis based on red–green differences provides exact inferences for treatment differences based on the t distribution with 3 degrees of freedom for error. Although 3 degrees of freedom is still quite small, it represents a substantial improvement over the exact single-degree-of-freedom analysis discussed previously for the design in Figure 2.6. Furthermore, the asymmetry in precision for treatment comparisons shown in Equation (2.7) for the design in Figure 2.6 is avoided.

If we are restricted to four blocks of size 3 but can afford to use 12 rather than six microarray slides to measure the 12 experimental units, the design depicted in Figure 2.8 is recommended. In this design, two RNA samples from each experimental unit are measured on different slides with different dyes and compared with samples from opposing treatment groups. This design is balanced in all relevant ways, and the resulting 24 observations can be analyzed using a linear mixed model with fixed effects for dyes and treatments and random effects for blocks, slides, and experimental units. Note that it is important to explicitly include a random effect for each experimental unit (in addition to the implicit observation-specific error terms) to allow for correlation among the two measurements of expression obtained for each experimental unit.

Jarrett and Ruggiero (2008) used the within-block looping strategy depicted in Figure 2.8 for the second phase design of a two-color microarray experiment where the phase 1 design is a balanced incomplete block design with seven treatments and seven blocks of size 3. They detailed analysis strategies for this complex combination of phase 1 and phase 2 designs and provided a more

general discussion of two-phase designs that is supported by additional microarray design examples.

Milliken, Garrett, and Travers (2007) studied second phase design for two-color microarray experiments that use a split-plot design in the first phase. They focused on the special case where the whole-plot treatment factor and the split-plot treatment factor each have two levels. Furthermore, they assumed that the whole-plot portion of the experiment is conducted as a randomized complete block design. They considered three strategies for pairing experimental units on two-color microarray slides. In strategy A, split-plot experimental units from the same split-plot treatment level are compared across differently treated whole-plot experimental units from the same block. In strategy B, split-plot experimental units from different split-plot treatment levels and different whole-plot experimental units are compared directly on slides. In strategy C, the two split-plot experimental units within each whole-plot experimental unit are compared directly. Milliken, Garrett, and Travers (2007) provided detailed information about modeling and analysis for data from each of these designs and concluded that any one of these three designs may be appropriate depending on the treatment comparisons of greatest interest to an investigator. For example, design C provides the most precision for comparing split-plot treatments within each whole-plot treatment, while designs A and B provide greater precision for comparing whole-plot treatments.

In general, it can be challenging to evaluate multiple candidate designs when phase 1 designs are necessarily complex. When linear mixed model analyses are required, the best candidate designs may be identified via simulations or approximations as described by Stroup (2002), Tempelman (2005), and Rosa, Steibel, and Tempelman (2005). However, the results of design comparisons may depend on the relative magnitudes of unknown quantities like block, slide, experimental unit, and error variance components. Thus, identifying a design that is guaranteed to be optimal will be impossible in many circumstances. However, use of simple strategies illustrated in this chapter can lead to effective designs that satisfy practical constraints. In particular, this section shows that researchers should pay careful attention to the interplay between the two phases of a microarray experimental design and attempt to avoid between-block comparisons by nesting slides within phase 1 blocks when possible.

REFERENCES

- Altman, N.S. and J. Hua (2006). Extending the loop design for two-channel microarray experiments. *Genetical Research*, **88**, 153–163.
- Banerjee, T. and R. Mukerjee (2007). Optimal factorial designs for cDNA microarray experiments. *The Annals of Applied Statistics*, **2**, 366–385.

- Barrett, T., D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, R.N. Muertter, and R. Edgar (2009). NCBI GEO: Archive for high-throughput functional genomic data. *Nucleic Acids Research*, **37**, D5–D15.
- Choe, S.E., M. Boutros, A.M. Michelson, G.M. Church, and M.S. Halfon (2005). Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, **6**, R16.
- Churchill, G.A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics Supplement*, **32**, 490–495.
- Cope, L.M., R.A. Irizarry, H.A. Jaffee, Z. Wu, and T.P. Speed (2004). A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**, 323–331.
- Dobbin, K., J.H. Shih, and R. Simon (2003). Statistical design of reverse dye microarrays. *Bioinformatics*, **19**, 803–810.
- Dudoit, S., Y.H. Yang, M.J. Callow, and T.P. Speed (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111–140.
- Edgar, R., M. Domrachev, and A.E. Lash (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, **30**, 207–210.
- Gadbury, G.L., G.P. Page, J. Edwards, T. Kayo, T.A. Prolla, R. Weindruch, P.A. Permana, J.D. Mountz, and D.B. Allison (2004). Power and sample size estimation in high dimensional biology. *Statistical Methods in Medical Research*, **13**, 325–338.
- Glonek, G.F.V. and P.J. Solomon (2004). Factorial and time course designs for cDNA microarray experiments. *Biostatistics*, **5**, 89–111.
- Irizarry, R.A., B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, and T.P. Speed (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
- Irizarry, R.A., Z. Wu, and H.A. Jaffee (2006). Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, **22**, 789–794.
- Jarrett, R.G. and K. Ruggiero (2008). Design and analysis of two-phase experiments for gene expression microarrays—Part I. *Biometrics*, **64**, 208–216.
- Kendziorski, C.M., Y. Zhang, H. Lan, and A.D. Attie (2003). The efficiency of pooling mRNA in microarray experiments. *Biostatistics*, **4**, 465–477.
- Kenward, M.G. and J.H. Roger (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **53**, 983–997.
- Kerr, M.K. (2003). Design considerations for efficient and effective microarray studies. *Biometrics*, **59**, 822–828.
- Kerr, M.K. and G.A. Churchill (2001). Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183–201.
- Landgrebe, J., F. Bretz, and E. Brunner (2006). Efficient design and analysis of two colour factorial microarray experiments. *Computational Statistics and Data Analysis*, **50**, 499–517.
- McIntyre, G.A. (1955). Design and analysis of two phase experiments. *Biometrics*, **11**, 324–334.

- Milliken, G.A., K.A. Garrett, and S.E. Travers (2007). Experimental design for two color microarrays applied in a pre-existing split plot experiment. *Statistical Applications in Genetics and Molecular Biology*, **6**, 1–21.
- Mosteller, F. and J.W. Tukey (1977). *Data Analysis and Regression: A Second Course in Statistics*. Reading, MA: Addison-Wesley.
- Nguyen, D.V., A.B. Arpat, N. Wang, and R.J. Carroll (2002). DNA microarray experiments: Biological and technological aspects. *Biometrics*, **58**, 701–717.
- Phimister, B. (1999). Going global. *Nature Genetics*, **21**, 1.
- Pukelsheim, F. (1993). *Optimal Design of Experiments*. New York: Wiley.
- Ritchie, M.E., J. Silver, A. Oshlack, M. Holmes, D. Diyagama, A. Holloway, and G.K. Smyth (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, **23**, 2700–2707.
- Rosa, G.J.M., J.P. Steibel, and R.J. Tempelman (2005). Reassessing design and analysis of two-colour microarray experiments using mixed effects models. *Comparative and Functional Genomics*, **6**, 123–131.
- Ruppert, D., D. Nettleton, and J.T.G. Hwang (2007). Exploring the information in p-values for the analysis and planning of multiple test experiments. *Biometrics*, **63**, 483–495.
- Schena, M., D. Shalon, R.W. Davis, and P.O. Brown (1995). Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science*, **270**, 467–470.
- Silver, J.D., M.E. Ritchie, and G.K. Smyth (2009). Microarray background correction: Maximum likelihood estimation for the normal—Exponential convolution. *Biostatistics*, **2**, 352–363.
- Silvey, S.D. (1980). *Optimal Design*. London: Chapman and Hall.
- Smyth, G.K. and T.P. Speed (2003). Normalization of cDNA microarray data. *Methods*, **31**, 265–273.
- Stroup, W.W. (2002). Power analysis based on spatial effects mixed models: A tool for comparing design and analysis strategies in the presence of spatial variability. *Journal of Agricultural, Biological, and Environmental Statistics*, **7**, 491–511.
- Tempelman, R.J. (2005). Assessing statistical precision, power, and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models. *Veterinary Immunology and Immunopathology*, **105**, 175–186.
- Wit, E., A. Nobile, and R. Khanin (2005). Near-optimal designs for dual channel microarray studies. *Applied Statistics*, **54**, 817–830.
- Wu, Z. and R.A. Irizarry (2007). A statistical framework for the analysis of microarray probe-level data. *The Annals of Applied Statistics*, **2**, 333–357.
- Yang, Y.H. and T. Speed (2002). Design issues for cDNA microarray experiments. *Nature Reviews Genetics*, **3**, 579–588.
- Yang, Y.H., S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, and T.P. Speed (2002). Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30**, e15.
- Zhang, W., A. Carriquiry, D. Nettleton, and J.C.M. Dekkers (2007). Pooling mRNA in microarray experiments and its effect on power. *Bioinformatics*, **23**, 1217–1224.

C H A P T E R 3

Spatial Analysis of Agricultural Field Experiments

Joanne K. Stringer, Alison B. Smith, and Brian R. Cullis

3.1 INTRODUCTION

In field trials, which contain a large number of treatments, growing conditions may be quite variable throughout the experimental site. This leads to the phenomenon known as *spatial trend or spatial variability*. Unless accounted for, spatial variability may seriously bias treatment estimates and inflate standard errors. The problem of spatial variability can be addressed by sound experimental design, careful trial management, and appropriate statistical analyses. Much research associated with spatial methods for agricultural field trials has been in the context of plant breeding variety trials where the experimental treatments are different varieties of a particular crop. A key component in plant breeding is to select varieties that are more productive than the established commercial varieties. This requires many years of testing through a number of selection stages. In the early stages of selection, planting material for field testing is typically limited, resulting in only partial replication of experimental varieties. Once the elite experimental varieties are identified from early stage trials, more intense testing is undertaken in advanced stage trials. In this stage, varieties are replicated within an experimental site and across several sites.

In this chapter, we describe the spatial analysis methodology in the plant breeding variety trial setting, but most of the methodology applies to field trials in general. We commence with a historical review of experimental designs and spatial analysis methods for variety trials. In Section 3.3, a spatial linear

Design and Analysis of Experiments: Special Designs and Applications, First Edition. Edited by Klaus Hinkelmann.

© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

mixed model is introduced followed by estimation, prediction, and testing. The spatial modelling process is then outlined and applied to two examples.

3.2 METHODS TO ACCOUNT FOR SPATIAL VARIATION

3.2.1 Design of Experiments

Optimal plot technique and sound experimental design is the cornerstone of field experimentation. This is irrespective of the role of modern based approaches, such as spatial analysis. The fundamental principle in design is the arrangement of experimental units (or plots) into groups, which are homogeneous within a group for the trait of interest. These groups are referred to as *blocks*, and the process of grouping experimental units is called *blocking*. Traditional designs, such as the randomized complete block design (RCBD), attempt to reduce spatial heterogeneity in field trials by maximizing the variation among different blocks and minimizing the variation among plots in the same block. In an RCBD, all treatments occur once in every block. As a consequence in large field trials, the RCBD may not be suitable as within-trial heterogeneity is likely to be present and blocks will not be homogeneous. Experimental error will increase, and the precision of the experiment, that is, its sensitivity to detect treatment differences, may be adversely affected.

To overcome this problem, a class of designs was developed in which each block was divided into smaller more homogenous subgroups that did not contain all treatments. These types of designs are incomplete-block (IB) designs. The first type of IB design to be developed were lattice and other balanced incomplete block BIB designs (Yates 1936). In these designs, every pair of treatments occurs together in a block the same number of times, and hence all treatments are compared with the same precision (HK2, section 2.8.3). Although lattices and other BIB designs were developed for large-scale field trials, they require too many replicates to be particularly useful. Partially balanced incomplete block (PBIB) designs allow for greater freedom of choice in the number of replicates (Bose and Nair 1939). Square and rectangular lattice designs are available for a low number of replicates, but require that the number of treatments, t , be related to the incomplete block size, k , in the form $t = k^2$ (simple lattice) or $t = k^3$ (triple lattice) or $t = k(k + 1)$ (rectangular lattice) (HK2, section 18.1). This limits their application in large field trials.

An improvement on PBIB designs would be to control the physical position of plots in the field to minimize bias to some treatments. This led to the development of resolvable incomplete block designs, such as alpha designs, constructed using cyclic methods (Patterson and Thompson 1971). In resolvable designs, the incomplete blocks can be grouped together to form a complete replicate of the treatments. Resolvability is advantageous for many reasons, including trial management; for example, replicate blocks can be managed, sprayed, or harvested on different days (Coombes 2002). If replicates are

contiguous, and depending on orientation of the replicates, an alpha design can be further optimized to ensure that treatments occur only once in each long column or long row. Such designs are called Latinized row–column designs (Williams and John 1989). Both alpha and Latinized row–column designs are very flexible and are available for most combinations of the number of treatments, replicates, and plots per block.

For early-stage field trials, which are often unreplicated, an augmented design (Federer 1956) is commonly used. In this design, replicated standard or *control plots* are allocated systematically on a diagonal grid in a complete or incomplete block design and are augmented with a number of unreplicated test varieties. The plots with the standard varieties are also called *check plots*. More sophisticated augmented designs allow for the adjustment of test varieties by rows and columns (Federer and Raghavarao 1975; Lin and Poushinsky 1983). In these designs, each test variety is surrounded by either two, three, or four check plots. Hence, in these designs, up to 50% of the total plot number may be used for check plots, which limits their application in the early stages of a plant improvement program. A similar design to the augmented row–column design was developed by Lin and Poushinsky (1983). In their modified augmented design (MAD), test varieties are arranged in a 3×3 array with check plots in the middle. Additionally, the check plots can be arranged in a Latin square design. This arrangement ensures that each test variety is equidistant from the check plot. The MAD is far more flexible than the augmented row–column design in that a larger number of test varieties can be assessed.

The development of nearest neighbor methods for analyzing agricultural field trials (see Section 3.2.2) has led to much interest into trial designs that consider how the response on a given plot is affected by neighboring plots. Such *neighbor-balanced* designs (NBD) have been developed for plots arranged linearly (one dimension) or in a row–column array (two dimensions). The earliest NBD were proposed by Williams (1949), and since then a large number of papers have been written on the subject. Williams (1952), Kiefer and Wynn (1981) and Cheng (2003) centered on one-dimensional designs. In these designs, each treatment had every other treatment as its neighbor in an adjacent plot an equal number of times (Chan and Eccleston 2003). These designs were extended to two dimensions for a range of designs by Freeman (1979), Street and Street (1985), Street (1986), Martin (1996) and Morgan and Uddin (1991). A two-dimensional design is said to be nearest-neighbor balanced if each treatment has every other treatment as its neighbor in an adjacent plot both in rows and columns an equal number of times (Chan and Eccleston 2003). Many of these NBD require too many replicates for large number of varieties, and this limits their usefulness in the early stages of a plant breeding program. In the approach developed by Chan and Eccleston (2003), resolvable spatial designs were developed for early generation plant breeding trials. As in augmented designs, the test varieties to be evaluated in resolvable spatial designs are unreplicated and checks plots are replicated. In an alternative approach, Cullis, Smith, and Coombes (2006) proposed the use

of partially replicated, *p*-rep, designs in which a subset, say 30%, of the test lines are replicated and these are arranged in a resolvable spatial design. The remaining unreplicated test lines are allocated at random to the remaining plots. The original motivation behind these designs was to maintain the same trial size as in an augmented designs but replace the standard varieties by replicated plots of test lines.

3.2.2 Spatial Analysis Methods

Plant breeding field trials are usually laid out in a rectangular array of plots. The size of the plot depends on the objective of the field trial, but usually plots are long and thin, and there are more rows than columns as indicated in Figure 3.1.

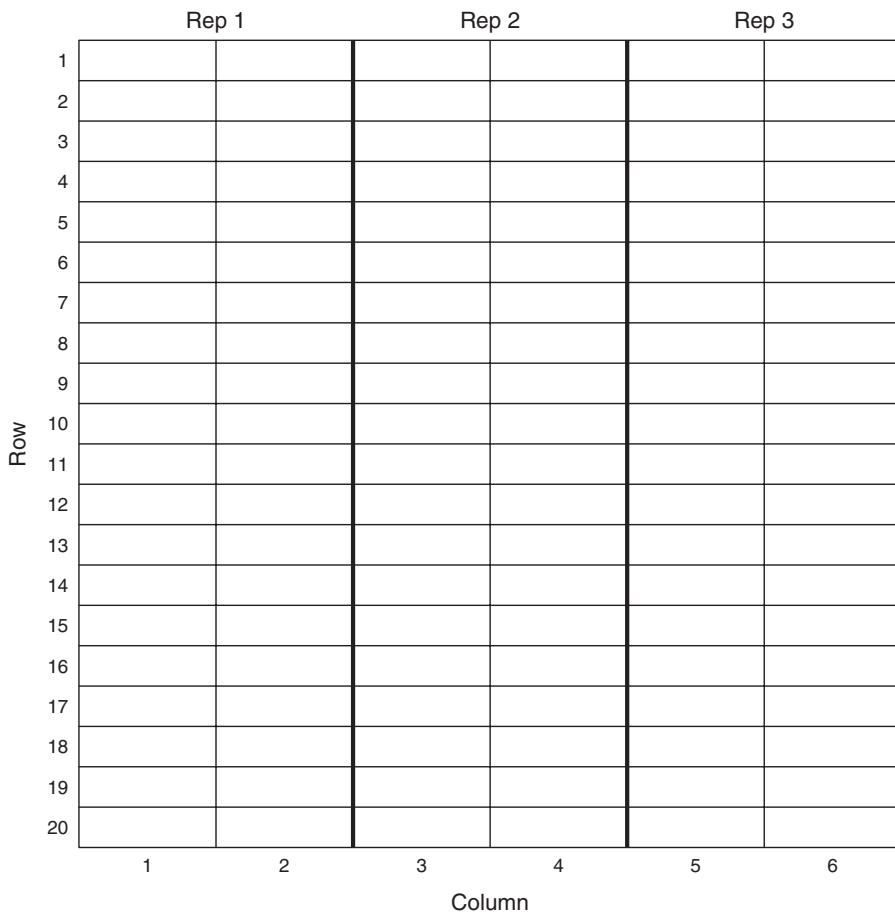


Figure 3.1 Field layout of a replicated trial.

One of the classical analysis-based approaches for local control of spatial variation is to use the method of check plots (Waincko 1914). The check plots are distributed over the trial area (see Section 3.2.1), and a fertility index is derived from their yields to provide a benchmark to assess the yields of test varieties. An alternative approach, which may be more useful for detecting small-scale variation, is *spatial or nearest neighbor* analysis in which the yield of a plot is adjusted by using information from immediate neighbors. Papadakis (1937) proposed the earliest nearest neighbor based method for plots in a single row. In this method, the yield of neighboring plots, corrected for treatment effects, are used as covariates to provide a measure of local control of soil fertility. The Papadakis method has been widely used, and substantial increases in accuracy of variety estimates were obtained (Pearce and Moore 1976). However, it lacks efficiency when large trends are present, as the covariates are computed from variety means, which are inaccurate as a result of within block trends (Wilkinson et al. 1983).

Bartlett (1978) extended the Papadakis method to two dimensions and suggested that the efficiency of the method may be improved by using an iterative version in which treatment estimates from the previous iteration are used to redefine the nearest neighbor covariate. In a study by Kempton and Howes (1981) of 118 wheat variety trials, the iterated Papadakis method resulted in approximately a 9% increase in precision for treatment estimates relative to the incomplete block analysis. However, Wilkinson et al. (1983) and Wu, Mather, and Dutilleul (1998) subsequently found that the iterated Papadakis method produced treatment *F*-tests with inflated type I errors, resulting in more significant differences being declared than actually existed.

Research into the area of nearest neighbor analysis was largely neglected until Wilkinson et al. (1983) proposed the *smooth trend plus independent error model* on which many spatial models have since been based. In this model, Wilkinson et al. (1983) assumed that trend effects were made small by taking second differences of the data. As the approach does not use a nearest neighbor covariate corrected for treatments effects, many of the inefficiencies of the Papadakis method were avoided. It can be thought of as the “moving block” extension of the classical fixed block analysis. The paper by Wilkinson et al. (1983) also proposed a radical change in how trials should be designed. They recommended that designs with some form of neighbor balance (see Section 3.2.1) be used and then analyzed using nearest neighbor analysis. Wilkinson’s paper, which was read before the Royal Statistical Society in 1983, was followed by a series of discussions from a range of eminent biometricalians (see pp. 177–211 of Wilkinson et al. 1983). It received both support for being highly innovative, which warranted further investigation, and some concerns about precision (p. 178) and loss of efficiency in the analysis (p. 188).

Since the paper by Wilkinson et al. (1983), there have been many alternative models that differ in the assumptions about trend, how to remove it, and the estimation methods used. One approach to accommodate spatial variability which has been widely adopted is to use *trend or polynomial regression*

analysis (Federer and Schlottfeldt 1954; Kirk, Haynes, and Monroe 1980; Warren and Mendez 1982; Tamura, Nelson, and Naderman 1988; Bowman 1990). One of the shortcomings of trend analysis is that there is often no biological justification for the degree of the polynomial function applied to the data, and this results in overfitting or fitting an incorrect model (Mead and Pike 1975; Brownie, Bowman, and Burton 1993).

The technique developed by Green, Jennison, and Seheult (1985) and used by Clarke and Baker (1996), Durbán Reguera (1998), and Durban, Currie, and Kempton (2001) applies *least squares smoothing* (LSS) to the smooth trend plus independent error model of Wilkinson et al. (1983). LSS uses a different estimation approach of the smoothing parameter in comparison to that used in Wilkinson et al. (1983).

The two spatial models of Besag and Kempton (1986) also built on the approach by Wilkinson et al. (1983). In the *first difference model*, the random part of the model contains a combination of a stochastic fertility process plus an error component. An extension to the first difference model is the *errors in variables model* where a random error term is included to allow a comparison between the experimental error and fertility.

In spatial modeling, a stochastic process whose joint probability distribution changes when shifted in time or space is said to be *nonstationary*. In a unified approach, Gleeson and Cullis (1987) suggested that many of the previous models for spatial variability in field trials can be modeled by the general class of autoregressive integrated moving average (ARIMA) (p,d,q) random processes where p and q refer to the order of the autoregressive (AR) and moving average (MA) processes, and d refers to the degree of differencing required to achieve stationarity. Gleeson and Cullis (1987) proposed the use of residual maximum likelihood (REML) for variance parameter estimation. Subsequently, Lill, Gleeson, and Cullis (1988), in a large simulation study, demonstrated that REML resulted in a relatively unbiased treatment F ratio for low-order ARIMA models. They also found more accurate and precise estimates of variety effects compared with the incomplete or complete block analysis.

Gleeson and Cullis (1987) fitted their model to the plot errors in one dimension. However, substantial spatial correlation may exist in two directions. Cullis and Gleeson (1991) extended their earlier model to two directions by using separable lattice processes as suggested by Martin (1990). Many authors, such as Zimmerman and Harville (1991), Kempton, Seraphin, and Sword (1994), Stroup, Baenziger, and Mulitz (1994), Grondona et al. (1996) and Gilmour, Cullis, and Verbyla (1997), have since demonstrated the need for two-dimensional models even when plots are long and narrow.

Many early spatial trend models that were proposed used some form of differencing (of adjacent plots) to remove nonstationarity in the data (Green, Jennison, and Seheult 1985; Besag and Kempton 1986; Gleeson and Cullis 1987; Cullis and Gleeson 1991). However, differencing can often lead to more complex modeling of the variance structure and hence Gilmour, Cullis, and

Verbyla (1997) and Zimmerman and Harville (1991) believed that differencing is unnecessary.

Zimmerman and Harville (1991) proposed to model spatial heterogeneity directly by using a geostatistical approach. They introduced a model termed *random field linear model—RFLM*, so called as the mean vector is assumed to be given by a linear model and the data vector, \mathbf{y} , is a realization of a random field. The RFLM approach is consistent with the models proposed by Ripley (1981) and Stein (1999) in which they consider \mathbf{y} as a realization of a Gaussian random field. In the model by Zimmerman and Harville (1991) spatial variation is partitioned into large-scale variation or global trend, which is modeled through the mean structure and small-scale variation or local trend, which is modeled through a random spatially correlated error structure. They used polynomials to account for the large-scale variation and isotropic covariance functions, such as linear, Gaussian, exponential, and spherical for the correlated error structure. Isotropic covariance functions have also been used by Brownie, Bowman, and Burton (1993), Bhatti et al. (1991), Ball, Mull, and Kodak (1993), and Stroup, Baenziger, and Mulitze (1994) to model spatial variation.

The geostatistical models used by Zimmerman and Harville (1991) often have two drawbacks. First, they often assume isotropy, which may be inappropriate for field trials, and, second, nonseparable processes often have computational difficulties in large data sets.

Gilmour, Cullis, and Verbyla (1997) building on earlier work by Cullis and Gleeson (1991), developed a more general approach to spatial modeling. They demonstrated that modeling plot errors alone, as a single process may not be appropriate in small plot field experimentation where variation from experimental procedures often arises. Hence, they partitioned spatial variation into smooth spatial trend and extraneous variation associated with trial management.

Gilmour, Cullis, and Verbyla (1997) partitioned trend within a field trial into the following three additive components:

- *Local trend (ζ)* which reflects small changes in fertility, soil moisture, and light. If trend is present within a field trial, then the errors on plots that are closer together will be positively related. If the correlation between the errors on pairs of plots, either in the row or column direction, decays toward zero as the distance increases, then this is characteristic of an autoregressive (AR) process. Gilmour, Cullis, and Verbyla (1997) usually modeled local trend using a first-order separable autoregressive process in the row (AR(1)) and column (AR(1)) directions.
- *Large-scale variation or global trend* is usually aligned with the rows and columns of a field trial. Global trend can be accommodated in the model by design factors, such as linear row and/or linear column effects or by fitting polynomial or spline functions (Verbyla et al. 1999) to the row and/or column coordinates.

- *Extraneous variation* arises from experimental procedures or management practices that have a recurrent pattern, such as direction of harvesting or method of planting. Such procedures may result in systematic and/or random row/column effects in the data, for example, serpentine harvesting up and down the rows causes plots in the “up” direction to be consistently higher/lower than in the “down” direction. Extraneous variation is often modeled by design factors such as a fixed “harvesting effect”.

In the following section, we examine a linear mixed model approach based on Gilmour, Cullis, and Verbyla (1997). Such an approach is used to analyze thousands of replicated and unreplicated cereal trials in Australia annually. Many authors, such as Gilmour, Cullis, and Verbyla (1997), Qiao et al. (2000), Sarker, Singh, and Erskine (2001), Silva, Dutkowski, and Gilmour (2001), Dutkowski et al. (2002), Singh et al. (2003), and Dutkowski et al. (2006), have used these techniques and report increased accuracy and precision in the estimates of variety effects in a range of crops.

3.3 A SPATIAL LINEAR MIXED MODEL

We assume that the field experiment comprises n plots laid out in a rectangular array of r rows and c columns (see Figure 3.1). It is further assumed that the plots form a single contiguous array but extensions to other layouts are possible. The vector of data, $\mathbf{y}^{(n \times 1)}$, is assumed to be ordered as rows within columns.

The linear mixed model for \mathbf{y} is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad (3.1)$$

where $\boldsymbol{\tau}^{(t \times 1)}$ and $\mathbf{u}^{(b \times 1)}$ are vectors of fixed and random effects, respectively and $\mathbf{X}^{(n \times t)}$ and $\mathbf{Z}^{(n \times b)}$ are the associated design-model matrices. We assume that \mathbf{X} is of full column rank. The vector of errors, \mathbf{e} is ordered as for the data vector. It is assumed that the joint distribution of (\mathbf{u}, \mathbf{e}) is the multivariate normal with zero mean and variance matrix

$$\sigma^2 \begin{bmatrix} \mathbf{G}(\gamma) & \mathbf{0} \\ \mathbf{0} & \mathbf{R}(\phi) \end{bmatrix},$$

where γ and ϕ are vectors of variance parameters. Hence, the distribution of the data is multivariate normal with mean $\mathbf{X}\boldsymbol{\tau}$ and variance matrix $\sigma^2\mathbf{H}$, where $\mathbf{H} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$.

In the variety trial context, the aim of an experiment is often the selection of superior entries, either for promotion to the next stage of testing or for commercialization. This aim is best achieved by assuming the variety effects to be random rather than fixed (Smith, Cullis, and Gilmour 2001). They would therefore be included in the vector \mathbf{u} . In the more general field trial setting,

the aim is often to test equality of treatment effects. In this case, treatment effects are regarded as fixed and are therefore included in the vector τ .

The model in Equation (3.1) may be used to conduct standard randomization-based analyses, such as RCB and IB. In the case of an RCB analysis, the vector of random effects, \mathbf{u} , includes the replicate effects. These are assumed independent with variance $\sigma^2\gamma_r$, say. The errors are also assumed independent with variance σ^2 . The IB analysis is similar except that \mathbf{u} also includes the block within replicate effects with variance $\sigma^2\gamma_b$, say.

A key feature that distinguishes spatial analysis from classical randomization-based analysis is the notion of correlation as a (smooth) function of the distance between plots. A randomization-based analysis arising from use of a block (either complete or incomplete) design does impose a correlation structure, but it is very restrictive, and there are discontinuities at block boundaries. For example, in the RCB analysis, all plots within a replicate are (equally) correlated (with correlation equal to $\gamma_r/(\gamma_r + 1)$), but plots in different blocks are not correlated (see also HK1, section 9.2). In an IB analysis, there are three levels of correlation, namely between plots in the same incomplete block (correlation equal to $(\gamma_r + \gamma_b)/(\gamma_r + \gamma_b + 1)$), between plots in different blocks but the same replicate (correlation equal to $\gamma_r/(\gamma_r + \gamma_b + 1)$), and between plots in different replicates (correlation equal to zero).

A spatial model provides for a much more realistic correlation structure between plots. In the linear mixed model, we partition the vector of errors, $\mathbf{e}^{(n \times 1)}$, into a vector $\zeta^{(n \times 1)}$ of spatially correlated effects, and a vector $\eta^{(n \times 1)}$ of independent technical or measurement errors. Accordingly, we assume

$$\text{var}[\zeta] = \sigma^2 \Sigma(\alpha) \quad \text{and} \quad \text{var}[\eta] = \sigma^2 \gamma_\eta \mathbf{I}_n,$$

where $\Sigma(\alpha)$ is an $n \times n$ correlation matrix that is a function of variance parameters (α) . Hence, the variance matrix \mathbf{R} , can be written as

$$\mathbf{R} = \sigma^2 (\Sigma(\alpha) + \gamma_\eta \mathbf{I}_n).$$

Any of the time series models described in Box and Jenkins (1976) and Zimmerman and Harville (1991) could be used to model the covariance structure in ζ . Separable lattice processes referred to in Cullis and Gleeson (1991) and Martin (1990) have been widely used in field experiments where there is spatial variability in two dimensions. The separability assumption allows the spatial correlation matrix, Σ , to be specified simply as the Kronecker product of two correlation matrices, one for each dimension. In particular, we consider the first-order separable autoregressive process, $AR(1) \times AR(1)$ (as used in Gilmour, Cullis, and Verbyla 1997), allows us to write the variance matrix for ζ as

$$\text{var}[\zeta] = \sigma^2 \Sigma(\alpha) = \sigma^2 (\Sigma_c(\alpha_c) \otimes \Sigma_r(\alpha_r)),$$

where $\Sigma_c^{(c \times c)}$ and $\Sigma_r^{(r \times r)}$ are the correlation matrices for columns and rows, respectively, and α_c and α_r are the associated autocorrelation parameters.

In our spatial linear mixed model, we include terms to respect the randomization employed in the experimental design. However, unlike in the full randomization-based analysis, we only include major blocking factors that are orthogonal to the treatments (e.g., replicate effects; main-plot effects in a split-plot design). We do not include design factors that are surrogates for local spatial trend (e.g., incomplete blocks) since these do not adequately accommodate trend (Cullis and Gleeson 1991).

3.3.1 Estimation, Prediction and Testing

Equation (3.1) is a linear mixed model. The first step in fitting the model is to obtain REML estimates of the variance parameters. In terms of the fixed and random effects in Equation (3.1), we obtain empirical best linear unbiased estimates (E-BLUEs) and empirical best linear unbiased predictions (E-BLUPs), respectively:

$$\hat{\tau} = (\mathbf{X}'\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{H}^{-1}\mathbf{y} \quad (3.2)$$

$$\tilde{\mathbf{u}} = \mathbf{GZ}'\mathbf{P}\mathbf{y}, \quad (3.3)$$

where $\mathbf{P} = \mathbf{H}^{-1} - \mathbf{H}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{H}^{-1}$. We use the term “empirical” to reflect the fact that the variance parameters in \mathbf{H} have been replaced with their REML estimates. Also note that the E-BLUPs of the errors are given by

$$\tilde{\mathbf{e}} = \mathbf{R}\mathbf{P}\mathbf{y}, \quad (3.4)$$

and are henceforth called the residuals.

In the spatial modeling process, effects may be added to the model to accommodate global trend and/or extraneous variation (see Section 3.2.2). The significance of these effects may be formally tested. Hypothesis tests for fixed effects may be made using Wald-type statistics. Traditional large-sample Wald statistics are assumed to follow chi-square distributions. Small sample statistics may be obtained in the form of F -statistics with denominator degrees of freedom calculated using the approach of Kenward and Roger (1997).

The significance of random terms may be assessed via hypothesis tests on the variance parameters. For example, to investigate the need for a measurement error term, we can test the null hypothesis that $\gamma_\eta = 0$. This is achieved using a residual maximum likelihood ratio test (REMLRT), with test statistic calculated as $D = 2(l_1 - l_0)$, where l_1 and l_0 are the residual log likelihoods for the two nested models, namely the models that include and exclude measurement error. Ordinarily, the statistic D has a chi-square distribution on d df, where d is the difference in number of variance parameters between the two nested models. For our test, $d = 1$, but our null hypothesis involves a boundary value for the variance parameter of interest (here, it is $\gamma_\eta = 0$, and we assume the parameter has been constrained to be non-negative). The statistic D is then

a mixture of chi-square variates on zero and 1 df. We refer the reader to Stram and Lee (1994) for further details. Tests for zero spatial correlations use the standard distribution for D .

REMLRTs may be used to compare two models in which the variance structures are nested and the fixed effects are the same. For non-nested models (still with the same fixed effects), information criteria, such as the Akaike information criterion, may be used.

It is important to note that random terms that are included to respect the randomization used in the experiment are always maintained in the model and so are not tested for significance.

3.3.2 The Spatial Modeling Process

The approach developed by Gilmour, Cullis, and Verbyla (1997) is a flexible and general approach. Model building is a difficult task and perhaps an inexact process. Indeed, as discussed by Grondona et al. (1996), the fact that partitioning of the spatial variation is not unique is a disadvantage. However, as Cressie (1991) points out, how one chooses to model spatial variation “depends on the underlying scientific problem,” and “we seek a model which is parsimonious and provides a reasonable description of the observed variation in our data.”

The modeling process outlined in Gilmour, Cullis, and Verbyla (1997) is a sequential approach with diagnostic checks on model adequacy at each stage. In Gilmour, Cullis, and Verbyla (1997), the baseline linear mixed model includes an $AR(1) \times AR(1)$ process for local trend, and measurement error is omitted. We note that in practice, if the number of rows and/or columns is small (e.g., less than four), we would not fit an autoregressive process for this dimension but assume independence. Also, our baseline model includes random effects for design factors that are orthogonal to the treatments as discussed in Section 3.3.

Having fitted the baseline model, diagnostic tools are used to assess the adequacy of the spatial model. Key features to investigate are the presence of global nonstationary trend, extraneous variation, and outliers. Residuals from the current model are used to explore these issues, mainly via the graphical tools described in the following sections.

3.3.2.1 Residual Plots

A plot of residuals against row (or column) number conditional on column (or row) is a useful tool to identify global trend and extraneous variation. It may also reveal potential outliers that should be checked for validity.

3.3.2.2 Sample Variogram

A particularly informative informal tool for selecting an appropriate spatial model is the sample or empirical variogram, or, more correctly, the *sample Cartesian semi-variogram* (Gilmour, Cullis, and Verbyla 1997). The variogram has been widely used in the geostatistics and repeated measures areas, and can

be adapted to field trials by considering that the errors in modelling field trial data arise from the realization of a stationary random field at a set of n spatial locations, identified by a vector, \mathbf{s}_i , $i = 1 \dots n$. In geostatistics, the data may relate to a single point or subset of points, say, \mathbf{s}_i , within a bounded subset of \mathfrak{R}^2 Euclidean space. In agricultural field trials, the data relate to a plot, so \mathbf{s}_i is taken to be centroid of the i^{th} plot. Let the spatial separation vector between two plots i and j be given by $\mathbf{h}_{ij} = \mathbf{s}_i - \mathbf{s}_j$, and we assume that the elements of \mathbf{R} are given by $\rho(\mathbf{h}^{ij}; \phi)$, where $\rho(\cdot)$ is a correlation function with a parameter vector ϕ and dependent on \mathbf{h}_{ij} .

Diggle et al. (2002) define the variogram of a random field to be the function $\Omega(\mathbf{s}, \mathbf{t}) = \Omega(\mathbf{h}) = \frac{1}{2} \text{var}((U(\mathbf{s}) - U(\mathbf{t})))$ for $\mathbf{h} = \mathbf{s} - \mathbf{t}$ and $\mathbf{s}, \mathbf{t} \in \mathfrak{R}^2$.

If the random field is second-order stationary, then it follows that:

$$\Omega(\mathbf{h}) = \sigma_H^2 (1 - \rho(\mathbf{h})).$$

In many geostatistical applications, it is often assumed that the second-order stationary random field is *isotropic*, which means that the dependence between any pair of observations depends only on the Euclidean distance between them, that is,

$$\rho(\mathbf{s}, \mathbf{t}) = \rho(\mathbf{h}) = \rho(d),$$

where d is the Euclidean distance and equals $\|\mathbf{h}\|$. Otherwise, the process is said to be *anisotropic*.

In field trials, the assumption of second-order stationarity is retained, but the assumption of isotropy is relaxed. It is assumed that field trials exhibit *geometric anisotropy* where the correlation function depends on the spatial separation vector $\mathbf{h} = (h_1, h_2)'$, where, in the case of field trials, h_1 and h_2 are the row and column displacements, that is, row and column differences between plots.

The *sample omnidirectional semi-variogram* is based on the empirical omnidirectional semi-variogram of the BLUP of a random field denoted $\tilde{\mathbf{e}} = (\tilde{e}_1 \dots \tilde{e}_n)'$, which is given by the set of points (d_{ij}, \tilde{v}_{ij}) : $j < i$ where $d_{ij} = \|\mathbf{h}_{ij}\|$ and

$$\tilde{v}_{ij} = \frac{1}{2} (\tilde{e}_i - \tilde{e}_j)^2.$$

To examine for geometric anisotropy, consideration is given to the graphical display of the semi-variogram in terms of the elements (or functions) of \mathbf{h} . The empirical semi-variogram cloud is the set of triples $(h_{ij1}, h_{ij2}, \tilde{v}_{ij})$. This is then represented graphically for field trials by binning based on the Cartesian coordinates of (h_1, h_2) . Formally, the Cartesian sample semi-variogram is then the set of points $(h_{k1}, h_{l2}, \bar{v}_{kl})$, $k = 1 \dots q_1$, $l = 1 \dots q_2$ and

$$\bar{v}_{kl} = \frac{1}{n_{kl}} \sum_{\substack{h_{ij1} \in S_{k1} \\ h_{ij2} \in S_{l2}}} \tilde{v}_{ij},$$

for prespecified lags h_{k1} and h_{l2} , S_{k1} is the set of values of h_1 closer to h_{k1} than any other $h_{k'1}$, S_{l2} is the set of values closer to h_2 than any other $h_{l'2}$, and n_{kl} is the number of points such that $h_{ij1} \in S_{k1}$ and $h_{ij2} \in S_{l2}$. For field trials, the h_{ij1} and h_{ij2} take a unique set of values given by $0, Q_1, 2Q_1, \dots, (r-1)Q_1$ and $0, Q_2, 2Q_2, \dots, (c-1)Q_2$, where Q_1 and Q_2 are the distances between plot centers.

It is also noted that we generally choose not to discriminate between (h_1, h_2) and $(h_1, -h_2)$ and consider binning (i.e., combining) the variogram of the first and fourth quadrants in Cartesian space.

Figure 3.2 illustrates the form of a theoretical variogram in which the only form of spatial variability present is local trend and is of the form of an $AR(1) \times AR(1)$ process with autocorrelations $\alpha_r = 0.8$ and $\alpha_c = 0.2$ for the row and column dimensions, respectively. Note that it has a smooth appearance and increases exponentially in the row and column directions to the variance of the process. When examining a sample variogram, departures from a smooth appearance may indicate the presence of extraneous variation. Similarly, failure of the sample variogram to reach a plateau in the row and/or column directions may indicate global trend. It is also important to note that in a sample variogram, the semi-variance at large displacements is based on few points so these should be excluded.

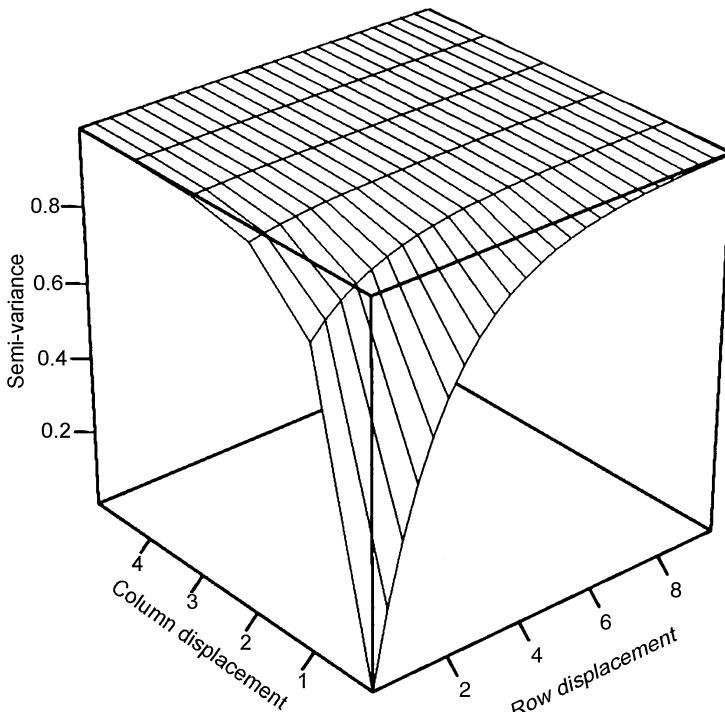


Figure 3.2 Theoretical variogram for an $AR(1) \times AR(1)$ process with $\alpha_r = 0.8$ and $\alpha_c = 0.2$.

The Cartesian sample semi-variogram has been used by Gilmour, Cullis, and Verbyla (1997) in an informal manner to identify possible problems with the current variance model during the model selection process. As discussed in Gilmour, Cullis, and Verbyla (1997), the presence of global and/or extraneous variation produces distinctive patterns in the Cartesian sample semi-variogram. However, if the covariance structure is complex the sample variogram can be difficult to interpret (Stefanova, Smith, and Cullis 2009). An enhanced diagnostic developed in Stefanova, Smith, and Cullis (2009) is to plot the two “faces” of the Cartesian sample semi-variogram; that is, the slices corresponding to zero row/column displacement augmented with approximate 95% pointwise coverage intervals. The *row face* is the plot of $\{(h_{kl}, \bar{v}_{kl} : k = 0, Q_1, 2Q_1, \dots, (r-1)Q_1), l = 0\}$, while the *column face* is the plot of $\{(h_{kl}, \bar{v}_{kl} : k = 0, Q_2, 2Q_2, \dots, (r-1)Q_2), l = 0\}$.

3.4 ANALYSIS OF EXAMPLES

In the following section, the spatial modelling process is applied to two examples. The analyses were undertaken using ASReml R (Butler et al. 2009). Details on ASReml R are given at <http://www.vsni.co.uk/software/asreml>.

3.4.1 Herbicide Tolerance Trial

Here we describe the analysis of a herbicide tolerance trial for six varieties of barley (data kindly supplied by Dr H.S. Dhammu, Department of Agriculture and Food Western Australia). The aim of this trial was to assess the tolerance of these varieties to the application of 30 different herbicides. Varieties are deemed tolerant to a herbicide if they suffer no significant yield loss compared with the variety grown in the absence of any herbicide and in a weed-free environment. Thus, there were a total of 31 treatments for each variety, namely the 30 herbicides and an untreated control treatment. There were three replicates of each herbicide for each variety and extra replicates of the control treatment. The full trial comprised 612 plots arranged as 102 rows by 6 columns. Replicate blocks were aligned with rows, with the first block corresponding to rows 1–34; the second block to rows 35–68; and the third block to rows 69–102. The design was a strip-plot (or split-block) design (see HK1, section 13.5) augmented with additional control plots. Within each block, varieties were randomized to columns, and treatments were randomized to rows with the exception that the control treatment was positioned systematically every 11 rows. The randomization of varieties and treatments is shown for the first block in Figure 3.3. Across the full design, the control treatment occurred in rows 1, 12, 23, 34, 35, 46, 57, 68, 69, 80, 91, and 102. The mean yield for the full trial was 2.63 tons per hectare, and there were 17 plots with missing yields.

For illustrative purposes, we conduct the randomization-based analysis for these data. In the underlying linear mixed model, the fixed effects comprise

	V4	C	V1	C	V3	C	V6	C	V5	C	V2	C
Row	V4	H12	V1	H12	V3	H12	V6	H12	V5	H12	V2	H12
1	V4	H13	V1	H13	V3	H13	V6	H13	V5	H13	V2	H13
2	V4	H19	V1	H19	V3	H19	V6	H19	V5	H19	V2	H19
3	V4	H18	V1	H18	V3	H18	V6	H18	V5	H18	V2	H18
4	V4	H3	V1	H3	V3	H3	V6	H3	V5	H3	V2	H3
5	V4	H20	V1	H20	V3	H20	V6	H20	V5	H20	V2	H20
6	V4	H25	V1	H25	V3	H25	V6	H25	V5	H25	V2	H25
7	V4	H1	V1	H1	V3	H1	V6	H1	V5	H1	V2	H1
8	V4	H27	V1	H27	V3	H27	V6	H27	V5	H27	V2	H27
9	V4	H2	V1	H2	V3	H2	V6	H2	V5	H2	V2	H2
10	V4	C	V1	C	V3	C	V6	C	V5	C	V2	C
11	V4	H5	V1	H5	V3	H5	V6	H5	V5	H5	V2	H5
12	V4	H11	V1	H11	V3	H11	V6	H11	V5	H11	V2	H11
13	V4	H16	V1	H16	V3	H16	V6	H16	V5	H16	V2	H16
14	V4	H14	V1	H14	V3	H14	V6	H14	V5	H14	V2	H14
15	V4	H26	V1	H26	V3	H26	V6	H26	V5	H26	V2	H26
16	V4	H9	V1	H9	V3	H9	V6	H9	V5	H9	V2	H9
17	V4	H10	V1	H10	V3	H10	V6	H10	V5	H10	V2	H10
18	V4	H30	V1	H30	V3	H30	V6	H30	V5	H30	V2	H30
19	V4	H7	V1	H7	V3	H7	V6	H7	V5	H7	V2	H7
20	V4	H21	V1	H21	V3	H21	V6	H21	V5	H21	V2	H21
21	V4	C	V1	C	V3	C	V6	C	V5	C	V2	C
22	V4	H17	V1	H17	V3	H17	V6	H17	V5	H17	V2	H17
23	V4	H28	V1	H28	V3	H28	V6	H28	V5	H28	V2	H28
24	V4	H23	V1	H23	V3	H23	V6	H23	V5	H23	V2	H23
25	V4	H22	V1	H22	V3	H22	V6	H22	V5	H22	V2	H22
26	V4	H4	V1	H4	V3	H4	V6	H4	V5	H4	V2	H4
27	V4	H8	V1	H8	V3	H8	V6	H8	V5	H8	V2	H8
28	V4	H29	V1	H29	V3	H29	V6	H29	V5	H29	V2	H29
29	V4	H6	V1	H6	V3	H6	V6	H6	V5	H6	V2	H6
30	V4	H15	V1	H15	V3	H15	V6	H15	V5	H15	V2	H15
31	V4	H24	V1	H24	V3	H24	V6	H24	V5	H24	V2	H24
32	V4	C	V1	C	V3	C	V6	C	V5	C	V2	C
33												
34												

Figure 3.3 Randomization of varieties (V1-V6) and herbicide treatments (C, H1-H30) for first replicate block in herbicide tolerance trial. Full trial comprises 102 rows by 6 columns.

an overall intercept, the variety and treatment effects, and the variety by treatment interaction. In order to respect the randomization used in the design, we must include random effects for replicate blocks, column effects within blocks, and row effects. The REML estimates of the variance parameters for this model are given in the first column of Table 3.1. The F -statistics, together with their probability levels (p -values), are given in Table 3.2. Note that the noninteger denominator degrees of freedom are due to the presence of missing values. In the randomization-based analysis, neither the effects of variety and treatment, nor the interaction were significant.

We now turn to the spatial analysis of these data. The baseline model is as for the randomization-based analysis except that we include an $AR(1) \times AR(1)$ model for local trend. The REML estimates of the variance parameters for this model are given in the second column of Table 3.1. The residual

Table 3.1 REML Estimates of Variance Parameters (Variance Components for Blocks, Columns within Blocks, Rows and Residual; Spatial Autocorrelations for Columns and Rows) for Three Models Fitted to Herbicide Tolerance Trial Data

Parameter	Linear Mixed Model		
	Randomization Based	Baseline Spatial	Final Spatial
Block variance	0.2500	0.0380	0.1735
Block:Column variance	0.0512	0.0000	0.0210
Row variance	0.1745	0.0005	0.0013
Residual variance	0.1183	0.3349	0.1805
Column autocorrelation		0.47	0.36
Row autocorrelation		0.79	0.64

Table 3.2 Wald F-Statistics (Given with Numerator and Denominator Degrees of Freedom and Probability Values) for Fixed Effects (Variety and Herbicide Main Effects; Variety by Herbicide Interactions; Orthogonal Polynomials over Row Number) in Two Models Fitted to Herbicide Tolerance Trial Data

Fixed Term	num df	Randomization-Based Model			Final Spatial Model		
		den df	F-Stat	p-Value	den df	F-Stat	p-Value
Variety	5	10.0	2.51	0.101	8.2	4.34	0.033
Herbicide	30	68.8	0.89	0.624	39.4	1.13	0.356
Variety:Herbicide	150	328.9	0.91	0.741	187.4	1.38	0.019
pol1 (linear)	1				56.5	20.42	0.000
pol2 (quadratic)	1				52.4	4.53	0.038
pol3 (cubic)	1				44.8	46.56	0.000

log-likelihood for this model was 116.71, whereas it was only 32.26 for the randomization-based analysis. The only difference between these models is the addition of two spatial autocorrelation parameters (estimated as 0.47 and 0.79 for columns and rows, respectively) so that a REMLRT can be conducted to assess the significance of the correlations. The test statistic is 168.90 on 2df, which is highly significant ($p < 0.001$). The residual plot and 2D plot of the sample variogram for the base-line spatial model are given in Figure 3.4. These plots suggest the existence of global trend across the rows of the trial. To clarify this, we also consider the two faces of the variogram, supplemented with the mean and 95% coverage intervals computed from 500 simulations (see Figure 3.5). An acceptable spatial model will have sample variogram faces that show

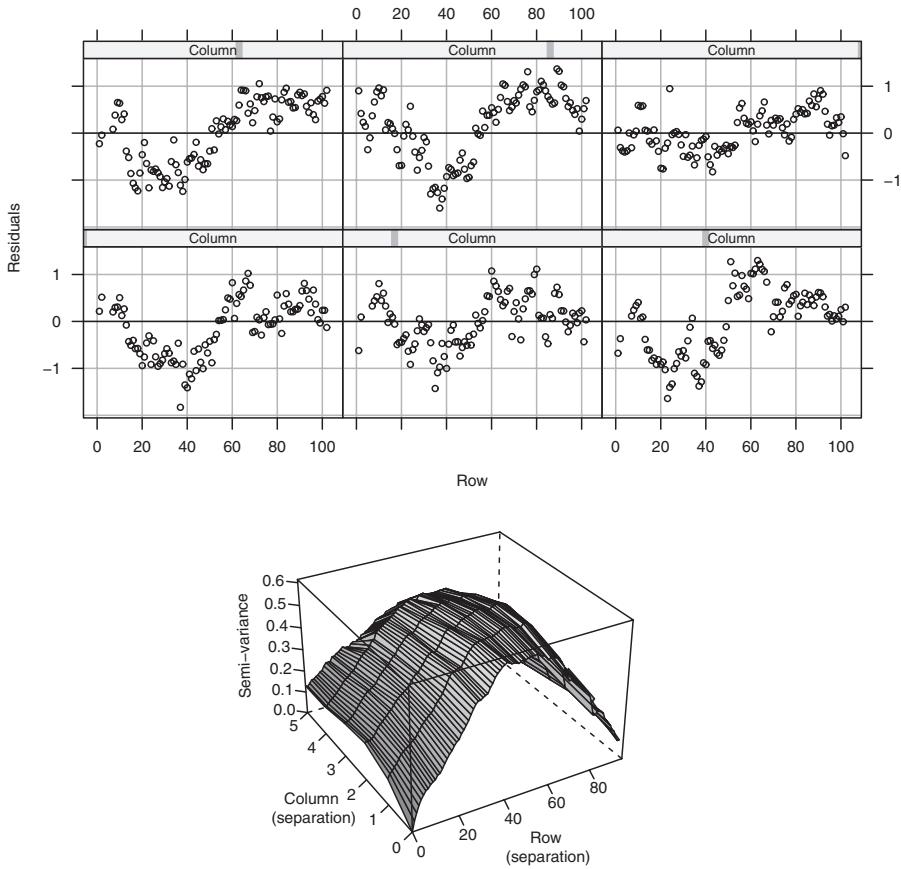


Figure 3.4 Residual plot and 2D plot of sample variogram for baseline spatial model fitted to herbicide tolerance trial data.

no systematic deviations from the simulation mean and remain within the 95% coverage intervals. The semi-variances for zero column separation (top graph in Figure 3.5) show a systematic nonlinear pattern and exceed the 95% coverage intervals for mid-range (between about 30 and 70) row separations. The shape of the variogram and the systematic up/down patterns observed in the residual plot (Figure 3.4) suggest the existence of nonlinear global trend across row number for each column. Such trend may be accommodated by including appropriate terms in the linear mixed model. Some authors suggest the use of cubic smoothing splines (see Verbyla et al. 1999); however, it appears that the trend here may be adequately modeled using the simpler approach of fitting orthogonal polynomials as fixed terms in the model. We found that a polynomial of order three was required. The Wald F -tests for assessing the significance of the polynomial terms (linear, quadratic, and cubic) are given in Table 3.2 under the heading of “Final spatial model.” The REML estimates of the

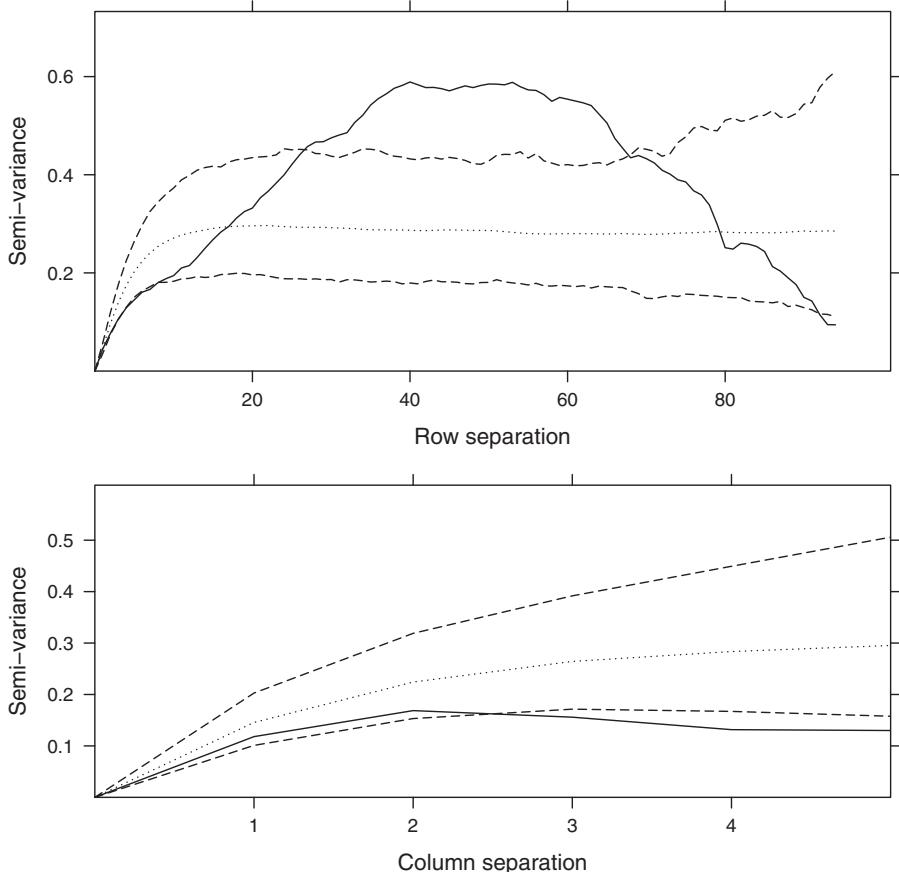


Figure 3.5 Plot of sample variogram faces for baseline spatial model fitted to herbicide tolerance trial data. Solid line on each plot is sample variogram; dashed lines are upper and lower 95% coverage intervals; dotted line is simulation mean.

variance parameters are given in Table 3.1 under the same heading. The variogram slices (Figure 3.6) now show the desired properties.

Importantly, the Wald F -tests for variety effects and for variety by herbicide interactions are now significant (see Table 3.2), whereas in the randomization-based analysis, they were not. The interaction is particularly important in terms of the aims of the experiment in that the researcher may now investigate comparisons of herbicide treatments with untreated control for each variety.

3.4.2 Variety Trial

Here we describe the analysis of an early stage canola breeding trial (data kindly provided by Prof. W. Cowling, Canola Breeders Western Australia, Pty

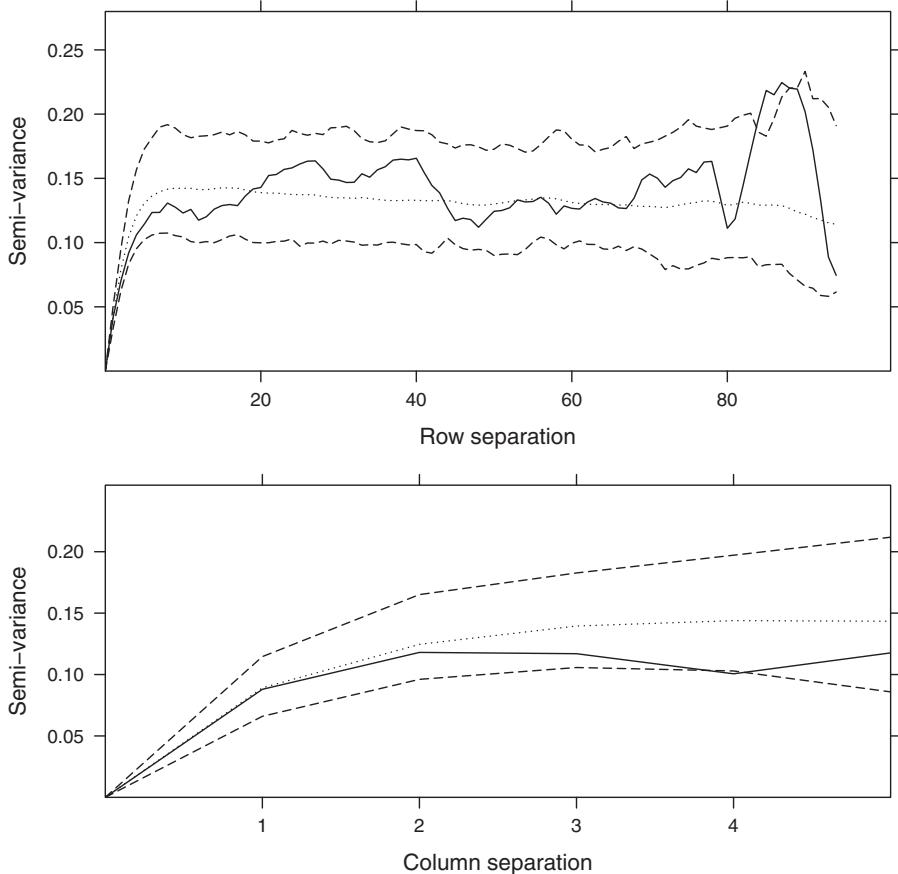


Figure 3.6 Plot of sample variogram faces for final spatial model fitted to herbicide tolerance trial data. Solid line on each plot is sample variogram; dashed lines are upper and lower 95% coverage intervals; dotted line is simulation mean.

Ltd.). A total of 232 entries were grown in a trial comprising 306 plots arranged as 51 rows by 6 columns. The aim was to select the best (highest yielding) entries for promotion to the next stage of testing. The design was a *p*-rep design (see Section 3.2.1). In these designs, a proportion, *p*, of the test lines is replicated with the remainder having only single plots. There may be multiple plots of check varieties. The designs were introduced by Cullis, Smith, and Coombes (2006) as a replacement for the unreplicated designs described previously in Section 3.2.1. The original idea was to maintain the same trial size as in an augmented design but replace (most of) the check plots with replicated plots of some test lines. In a simulation study conducted by Cullis, Smith, and Coombes (2006), the partially replicated designs resulted in better selection

decisions (thence higher genetic gains) compared with the augmented designs. In the trial under consideration here, 74 entries were replicated, and 158 had single plots only so that $p = 0.319$. The design was resolvable in the sense that all replicated entries occurred once in the first block (columns 1–3) and then again in the second block (columns 4–6).

The baseline spatial model for this trial comprised a single fixed effect (the overall intercept), random effects for blocks (in accordance with the randomization employed in the design), random effects for varieties (in accordance with the aim of the experiment), and an $AR(1) \times AR(1)$ model for local trend. The REML estimates of the spatial autocorrelations from this model were 0.18 and 0.64 for rows and columns, respectively. The residual plot and 2D plot of the sample variogram are shown in Figure 3.7. There is evidence of both global trend across row numbers and random column effects. Once again, we consider

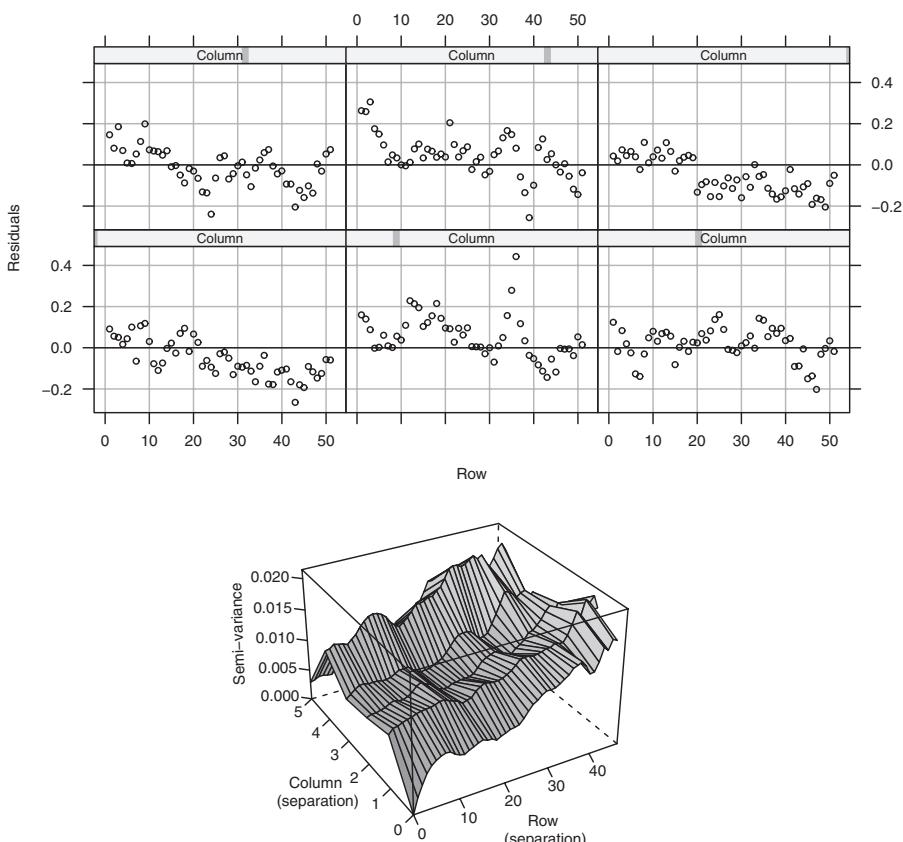


Figure 3.7 Residual plot and 2D plot of sample variogram for base-line spatial model fitted to variety trial data.

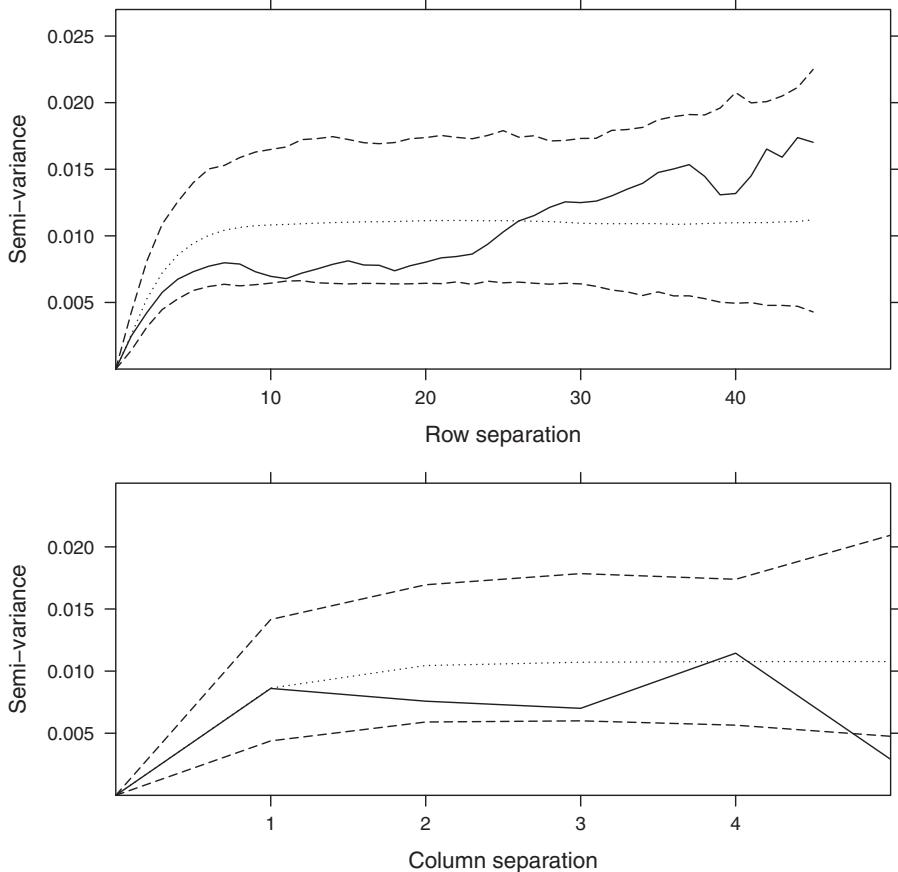


Figure 3.8 Plot of sample variogram faces for base-line spatial model fitted to variety trial data. Solid line on each plot is sample variogram; dashed lines are upper and lower 95% coverage intervals; dotted line is simulation mean.

the faces of the variogram supplemented with coverage intervals (Figure 3.8). Although the semi-variances largely remain within the 95% coverage intervals, there are clear systematic departures from the mean. The face at zero column displacement (top graph in Figure 3.8) shows an increasing trend as row separation increases. This, together with the residual plot in Figure 3.7, suggests the existence of a global linear trend across row number for each column. We accommodate this in the model by fitting a (fixed) linear regression on row number. For computational reasons, we use an orthogonal polynomial of order 1. When added to the linear mixed model, a Wald test revealed this to be significant ($p < 0.001$). The variogram faces for this model are shown in Figure 3.9. The semi-variances for zero column displacement no longer

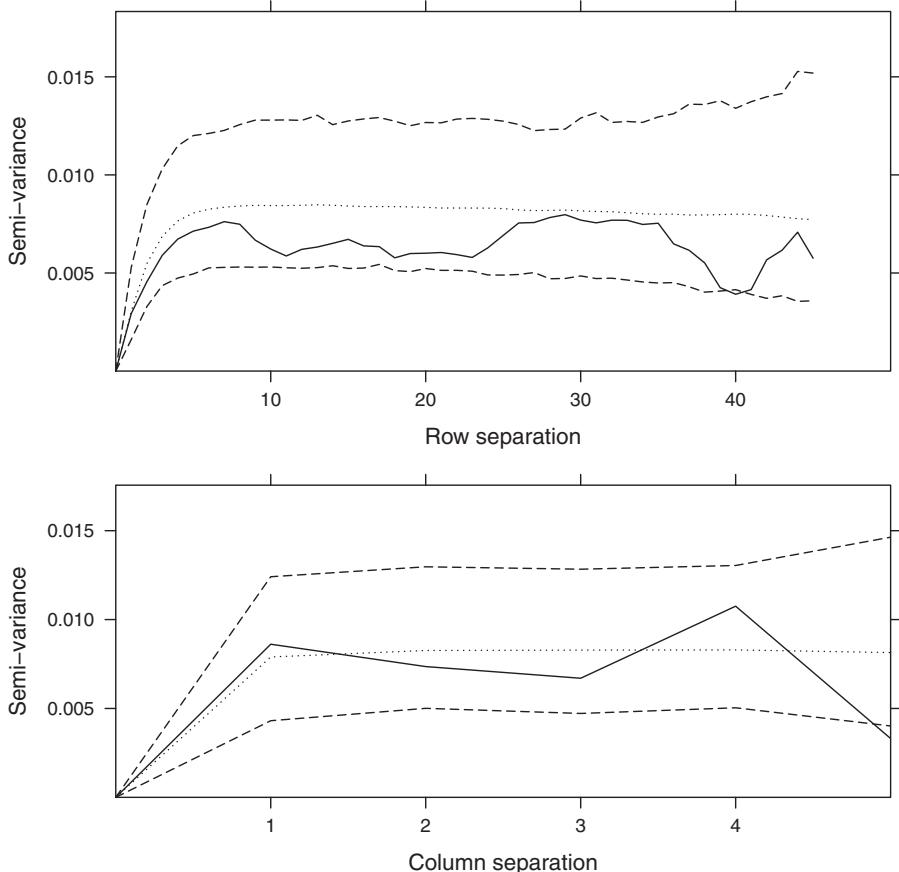


Figure 3.9 Plot of sample variogram faces for second spatial model fitted to variety trial data. Solid line on each plot is sample variogram; dashed lines are upper and lower 95% coverage intervals; dotted line is simulation mean.

display a linear trend but the sill (plateau variance) is substantially lower than the mean. This suggests that the variance is lower for plots in the same column compared with plots in different columns. Thus we should include random column effects in the model. When added to the linear mixed model, the associated variance component was significant ($p < 0.001$) based on a REMLRT. The variogram faces for this model are shown in Figure 3.10. Both graphs in this figure now show the desired form so we may be satisfied with the final spatial model that includes a fixed effect for the linear regression on row number, random effects for entries, blocks and columns, and an $AR(1) \times AR(1)$ model for local trend.

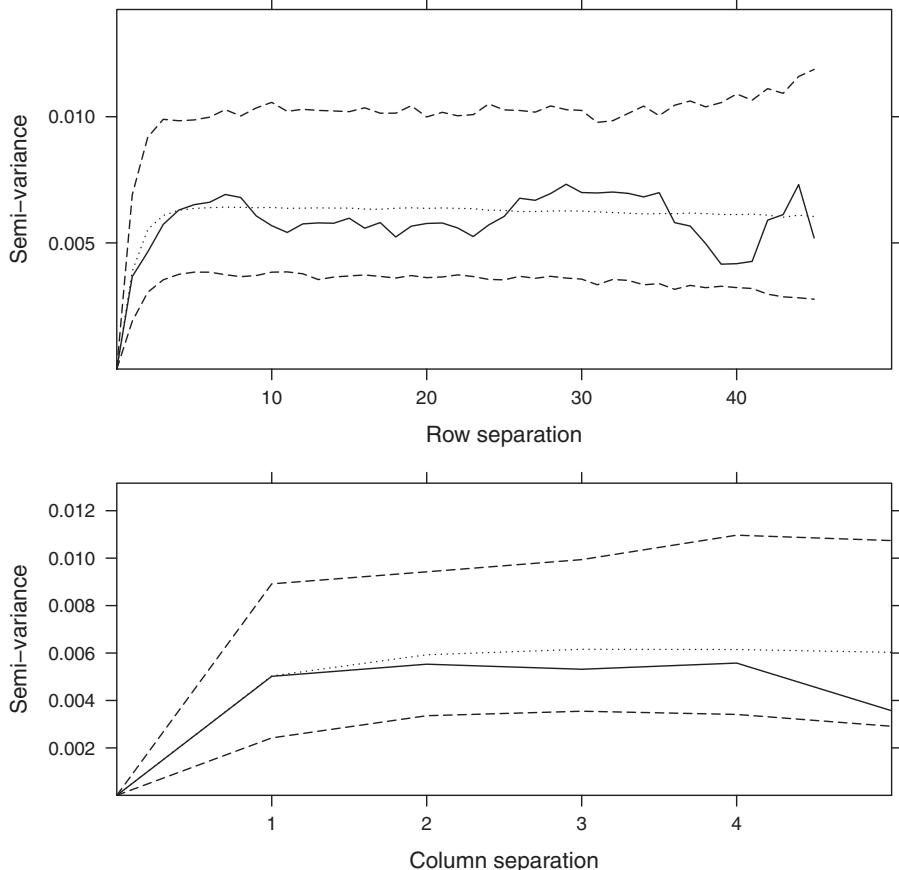


Figure 3.10 Plot of sample variogram faces for final spatial model fitted to variety trial data. Solid line on each plot is sample variogram; dashed lines are upper and lower 95% coverage intervals; dotted line is simulation mean.

Finally, since the aim of the experiment was the selection of the best entries, we obtain the E-BLUPs of the entry effects from the final spatial model. For comparative purposes, we also compute E-BLUPs from the randomization-based model (model with random effects for entries and blocks and independent errors). These are plotted against each other in Figure 3.11. Although there is general agreement, there would be some differences in the selected entries with the two models. The cutoff lines for selection of the top 20 entries for each model are shown on Figure 3.11. Using these cutoffs, there are 13 entries that would be selected using either model but 7 that would be selected under one model and not the other.

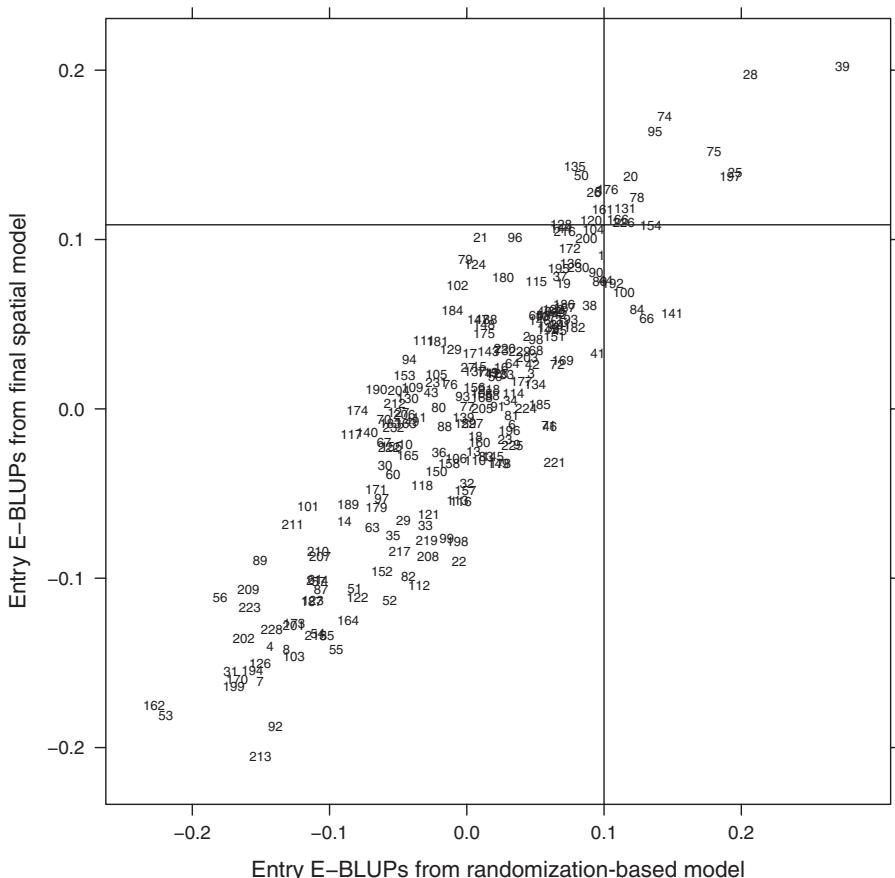


Figure 3.11 Plot of entry E-BLUPs from randomization-based and final spatial model fitted to variety trial data. Solid lines represent cutoff for selection of the best 20 entries for each model. Points are labeled with their entry number (1–232).

REFERENCES

- Ball, S.T., D.J. Mull, and C.F. Kodak (1993). Spatial heterogeneity affects variety trial interpretation. *Crop Science*, **33**, 931–935.
- Bartlett, M.S. (1978). Nearest neighbour models in the analysis of field experiments. *Journal of the Royal Statistical Society. Series B*, **40**, 147–158.
- Besag, J. and R. Kempton (1986). Statistical analysis of field experiments using neighbouring plots. *Biometrics*, **42**, 231–251.
- Bhatti, A.U., D.J. Mull, F.E. Koechler, and A.H. Gurmani (1991). Identifying and removing spatial correlation from yield experiments. *Soil Science Society of America Journal*, **55**, 1523–1528.

- Bose, R.C. and K.R. Nair (1939). Partially balanced incomplete block designs. *Sankhya*, **4**, 337–372.
- Bowman, D.T. (1990). Trend analysis to improve efficiency of agronomic trials in flue-cured tobacco. *Agronomy Journal*, **82**, 499–501.
- Box, G.E.P. and G.M. Jenkins (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.
- Brownie, C., D.T. Bowman, and J.W. Burton (1993). Estimating spatial variation in analysis of data from yield trials: A comparison of methods. *Agronomy Journal*, **85**, 1244–1253.
- Butler, D.G., B.R. Cullis, A.R. Gilmour, and B.J. Gogel (2009). ASReml-R reference manual, release 3. Technical report, Brisbane, Queensland.
- Chan, B.S.P. and J.A. Eccleston (2003). On the construction of nearest-neighbour balanced row-column designs. *Australian and New Zealand Journal of Statistics*, **45**, 97–106.
- Cheng, C.S. (2003). Construction of optimal balanced incomplete block designs for correlated observations. *The Annals of Statistics*, **11**, 240–246.
- Clarke, F.R. and R.J. Baker (1996). Spatial analysis improved precision of seed lot comparisons. *Crop Science*, **36**, 1180–1184.
- Coombes, N.E. (2002). The reactive tabu search for efficient correlated experimental designs. PhD thesis, Liverpool John Moores University, Liverpool.
- Cressie, N.A.C. (1991). *Statistics for Spatial Data*. New York: John Wiley and Sons Inc.
- Cullis, B.R. and A.C. Gleeson (1991). Spatial analysis of field experiments: An extension to two dimensions. *Biometrics*, **47**, 1449–1460.
- Cullis, B.R., A.B. Smith, and N.E. Coombes (2006). On the design of early generation variety trials with correlated errors. *Journal of Agricultural, Biological and Environmental Statistics*, **11**, 381–393.
- Diggle, P.J., P.J. Heagerty, K.Y. Liang, and S.L. Zeger (2002). *Analysis of Longitudinal Data*. New York: Oxford University Press.
- Durban, M., I.D. Currie, and R.A. Kempton (2001). Adjusting for fertility and competition in variety trials. *Journal of Agricultural Science*, **136**, 129–140.
- Durbán Reguera, M.L. (1998). Modelling spatial trends and local competition effects using semiparametric additive models. PhD thesis, Department of Actuarial Mathematics and Statistics, Heriot-Watt University.
- Dutkowski, G.W., J.C.E. Silva, A.R. Gilmour, and G.A. Lopez (2002). Spatial analysis methods for forest genetic trials. *Canadian Journal of Forest Research*, **32**, 2201–2214.
- Dutkowski, G.W., J.C.E. Silva, A.R. Gilmour, H. Wellendorf, and A. Aguiar (2006). Spatial analysis enhances modelling of a wide variety of traits in forest genetic trials. *Canadian Journal of Forest Research*, **36**, 1851–1870.
- Federer, W.T. (1956). Augmented (or Hoonuiaku) designs. *The Hawaiian Planters' Record*, **55**, 191–208.
- Federer, W.T. and D. Raghavarao (1975). On augmented designs. *Biometrics*, **31**, 29–35.
- Federer, W.T. and C.S. Schlottfeldt (1954). The use of covariance to control gradients in experiments. *Biometrics*, **10**, 282–290.

- Freeman, G.H. (1979). Some two-dimensional designs balanced for nearest neighbours. *Journal of the Royal Statistical Society. Series B*, **41**, 88–95.
- Gilmour, A.R., B.R. Cullis, and A.P. Verbyla (1997). Accounting for natural and extra-neous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics*, **2**, 269–293.
- Gleeson, A.C. and B.R. Cullis (1987). Residual maximum likelihood (REML) estimation of a neighbour model for field experiments. *Biometrics*, **43**, 277–288.
- Green, P., C. Jennison, and A. Seheult (1985). Analysis of field experiments by least squares smoothing. *Journal of the Royal Statistical Society. Series B*, **47**, 299–315.
- Grondona, M.O., J. Crossa, P.N. Fox, and W.H. Pfeiffer (1996). Analysis of variety yield trials using two-dimensional separable ARIMA processes. *Biometrics*, **52**, 763–770.
- Kempton, R.A. and C.W. Howes (1981). The use of neighbouring plot values in the analysis of variety trials. *Applied Statistics*, **30**(1), 59–70.
- Kempton, R.A., J.C. Seraphin, and A.M. Sword (1994). Statistical analysis of two-dimensional variation in variety yield trials. *Journal of Agricultural Science*, **122**, 335–342.
- Kenward, M.G. and J.H. Roger (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **53**, 983–997.
- Kiefer, J. and H.P. Wynn (1981). Optimum balanced block and latin square designs for correlated observations. *The Annals of Statistics*, **9**, 737–757.
- Kirk, H.J., F.L. Haynes, and R.J. Monroe (1980). Application of trend analysis to horticultural field trials. *Journal of the American Society of Horticultural Science*, **105**, 189–193.
- Lill, W.J., A.C. Gleeson, and B.R. Cullis (1988). Relative accuracy of a neighbour method for field trials. *Journal of Agricultural Science*, **111**, 339–346.
- Lin, C.S. and G. Poushinsky (1983). A modified augmented design for an early stage of plant selection involving a large number of test lines without replication. *Biometrics*, **39**, 553–561.
- Martin, R.J. (1990). The use of time-series models and methods in the analysis of agricultural field trials. *Communications in Statistics: Theory and Methods*, **19**, 55–81.
- Martin, R. (1996). *Handbook of Statistics 13: Design and Analysis of Experiments, Chapter Spatial Experimental Designs*. p. 477–514. Amsterdam: Elsevier Science.
- Mead, R. and D.J. Pike (1975). A review of response surface methodology from a biometric viewpoint. *Biometrics*, **31**, 803–851.
- Morgan, J.P. and N. Uddin (1991). Two-dimensional design for correlated errors. *The Annals of Statistics*, **19**, 2160–2182.
- Papadakis, J.S. (1937). Méthode statistique pour des expériences sur champ. *Bulletin Institut d'Amélioration des Plantes à Salonique*, **23**, 1–30.
- Patterson, H. and R. Thompson (1971). Recovery of interblock information when block sizes are unequal. *Biometrika*, **31**, 100–109.
- Pearce, S.C. and C.S. Moore (1976). Reduction of experimental error in perennial crops, using adjustment by neighbouring plots. *Experimental Agriculture*, **12**, 267–272.
- Qiao, C.G., K.E. Basford, I.H. Delacy, and M. Cooper (2000). Evaluation of experimental designs and analyses in wheat breeding trials. *Theoretical and Applied Genetics*, **100**, 9–16.

- Ripley, B.D. (1981). *Spatial Statistics*. New York: John Wiley and Sons Inc.
- Sarker, A., M. Singh, and W. Erskine (2001). Efficiency of spatial methods in yield trials in lentil *Lens culinaris* ssp. *culinaris*. *Journal of Agricultural Science*, **137**, 427–438.
- Silva, J.C.E., G.W. Dutkowski, and A.R. Gilmour (2001). Analysis of early tree height in forest genetic trials is enhanced by including a spatially correlated residual. *Canadian Journal of Forest Research*, **31**, 1887–1893.
- Singh, M., R.S. Malhotra, S. Ceccarelli, A. Sarker, S. Grando, and W. Erskine (2003). Spatial variability models to improve dryland field trials. *Experimental Agriculture*, **39**, 151–160.
- Smith, A.B., B.R. Cullis, and A.R. Gilmour (2001). The analysis of crop variety evaluation data in Australia. *Australian and New Zealand Journal of Statistics*, **43**, 129–145.
- Stefanova, K.T., A.B. Smith, and B.R. Cullis (2009). Enhanced diagnostics for the spatial analysis of field trials. *Journal of Agricultural, Biological, and Environmental Statistics*, **14**, 392–410.
- Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging (Springer Series in Statistics)*. New York: Springer-Verlag.
- Stram, D.O. and J.W. Lee (1994). Variance components testing in the longitudinal mixed effects setting. *Biometrics*, **50**, 1171–1177.
- Street, D.J. (1986). Unbordered two dimensional nearest neighbour designs. *Ars Combinatoria*, **22**, 51–57.
- Street, D.J. and A.J. Street (1985). Designs with partial neighbour balance. *Journal of Statistical Planning and Inference*, **120**, 47–59.
- Stroup, W.W., P.S. Baenziger, and D.K. Mulitze (1994). Removing spatial variation from wheat yield trials: A comparison of methods. *Crop Science*, **34**, 62–66.
- Tamura, R.N., L.A. Nelson, and G.C. Naderman (1988). An investigation of the validity and usefulness of trend analysis for field plot data. *Agronomy Journal*, **80**, 712–718.
- Verbyla, A.P., B.R. Cullis, M.G. Kenward, and S.J. Welham (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Applied Statistics*, **48**, 269–311.
- Waincko, A.T. (1914). Use and management of check plots in soil fertility investigations. *Journal of the American Society of Agronomy*, **6**, 122–124.
- Warren, J.A. and I. Mendez (1982). Methods for estimating background variation in field experiments. *Agronomy Journal*, **74**, 1004–1009.
- Wilkinson, G.N., S.R. Eckert, T.W. Hancock, and O. Mayo (1983). Nearest neighbour (NN) analysis of field experiments. *Journal of the Royal Statistical Society. Series B*, **45**, 151–178.
- Williams, E.J. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Scientific Research*, **2(A)**, 149–168.
- Williams, E.J. (1952). Experimental designs for serially correlated observations. *Biometrika*, **39**, 152–167.
- Williams, E.R. and J.A. John (1989). Construction of row and column designs with contiguous replicates. *Applied Statistics*, **38**, 149–154.

- Wu, T., D.E. Mather, and P. Dutilleul (1998). Application of geostatistical and neighbor analyses to data from plant breeding trials. *Crop Science*, **38**, 1545–1553.
- Yates, F. (1936). A new method of arranging variety trials involving a large number of varieties. *Journal of Agricultural Science*, **26**, 424–455.
- Zimmerman, D.L. and D.A. Harville (1991). A random field approach to the analysis of field-plot experiments and other spatial experiments. *Biometrics*, **47**, 223–239.

C H A P T E R 4

Optimal Designs for Generalized Linear Models

John Stufken and Min Yang

4.1 INTRODUCTION

Both HK1 and HK2 deal with experiments in which the planned analysis is based on a linear model. Selecting designs for such experiments remains a critically important problem. However, there are many problems for which a linear model may not be a great choice. For example, if the response is a binary variable or a count variable rather than a continuous measurement, a linear model may be quite inappropriate. Experiments with such response variables are quite common. For example, in an experiment to compare different dosages of a drug, the outcome may be success (the dosage worked for a particular patient) or failure (it did not work). A design would consist of selecting the different dosages to be used in the experiment and the number of patients to be assigned to the selected dosages. How can one identify a good design for such a problem in which a linear model for the binary response is simply inadequate?

Another feature of the example in the previous paragraph is that the purpose of the experiment is not to compare different treatments, but to understand the relationship between the response and the dosage of the drug. Most of the chapters in HK1 and HK2 are devoted to comparative experiments in which various treatments are to be compared with each other. Typically, each treatment is associated with a treatment effect, and the purpose of the experiment is some form of comparison of these effects. An exception to this is chapter 12 in HK1, in which the purpose of the experiment is to

Design and Analysis of Experiments: Special Designs and Applications, First Edition. Edited by Klaus Hinkelmann.

© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

understand the relationship between the response and one or more regression variables. This latter scenario will also be the focus of the current chapter.

As noted in HK1, section 2.10, there are many considerations that can and should go into the selection of an appropriate design for a particular experiment. These could be of a general scientific nature, of a purely statistical nature, or of an entirely practical nature, and are almost always a combination of considerations of all types. The single consideration in this chapter is that of identifying a design that is *optimal* according to a specified statistical optimality criterion. Some criteria of this type were briefly presented in HK2, section 1.13 in the context of block designs. Thus, as a result of additional experiment-specific considerations, the designs identified in this chapter may not be the designs used in experiments. But even if one does not use a design that is optimal under a specific optimality criterion, the optimal design will still provide a benchmark under this criterion for all other designs. Thus we may be willing to sacrifice optimality for other desirable properties of a design provided that the design that we do use has still a reasonable performance with respect to the optimality criterion (or possibly with respect to multiple optimality criteria). Without knowing how to identify optimal designs, we will have no basis to assess whether a given design performs well under the criterion of interest.

The three previous paragraphs set the stage for this chapter. As a concise summary, we will provide a brief introduction to some recent results on finding optimal designs for experiments in which the focus is on studying the relationship between a response and one or more regression variables through a special class of nonlinear models, namely *generalized linear models* (GLMs).

GLMs have found wide applicability in many fields, including drug discovery, clinical trials, social sciences, marketing, and many more. Methods of analysis and inference for these models are well established (see e.g., McCullagh and Nelder 1989; McCulloch and Searle 2001; Agresti 2002). The study of optimal designs for experiments that plan to use a GLM is, however, not nearly as well developed (see also Khuri et al. 2006), and tends to be much more difficult than the corresponding and better studied problem for the special case of linear models. (While linear models are a special case of GLMs, this chapter focuses on GLMs that correspond to nonlinear models.)

One of the challenges is that for a GLM, an optimal design typically depends on the unknown parameters. This leads to the concept of locally optimal designs, which are optimal for *a priori* chosen values of the parameters. The designs may be poor if the choice of values is far from the true parameter values. Where feasible, a multistage approach could help with this, in which a small initial design is used to obtain some information about the parameters. We will briefly return to this later, but simply state for now that results presented in this chapter are also applicable for a multistage approach (see also Yang and Stufken 2009).

To illustrate some of the basic ideas, we present a small example. Many of these ideas will be revisited in more detail in later sections.

Example 4.1. In dose–response studies and growth studies, a subject receives a stimulus at a certain level x . The binary response indicates whether the subject does or does not respond to the stimulus. The purpose of such an experiment is often to study the relationship between the probability that a subject responds and the level of the stimulus. The level can be controlled by the experimenter, and a judicious selection of the levels must be made prior to the experiment.

With Y_i and x_i denoting the response and stimulus level for the i th subject, we could use a logistic regression model of the form

$$\text{logit}[P(Y_i = 1)] = \alpha + \beta x_i, \quad (4.1)$$

where the logit function represents the log of the odds ratio. We consider two designs (without claiming that either is a good design). Design I uses level 0.067 for 53% of the subjects and level 0.523 for the other 47%. Design II uses each of the levels 0.1, 0.2, 0.3, 0.4, 0.5, and 0.6 equally often, namely for 16.7% of the subjects. Thus, we will think of a design as a probability measure on the possible levels. Such a design is also known as an *approximate design*. Whether it corresponds to a design that can be used in practice depends in part on the number of subjects in the experiment. For example, design I can be used with 100 subjects (53 of them at level 0.523 and 47 at level 0.067), but could not be used exactly with, for example, only 30 subjects. Thus, for an experiment with 30 subjects, there is no *exact design* corresponding to design I. An exact design corresponding to design II will only exist if the number of subjects is a multiple of 6.

While we can only use an exact design in practice, the discreteness of the design space for a fixed number of subjects makes it difficult to identify optimal designs in such a space. Working with approximate designs circumvents this difficulty, but at the expense that we might not be able to use an exact design that corresponds precisely to an optimal approximate design.

Continuing with the example, which of the two designs is best?

As stated, this question cannot be answered. First, we need to specify what we mean by “best.” We will do this by using an optimality criterion. For example, we could compare the designs in terms of $\text{Var}(\hat{\alpha}) + \text{Var}(\hat{\beta})$, where $\hat{\alpha}$ and $\hat{\beta}$ denote the maximum likelihood estimators (MLEs) for the unknown parameters α and β . We might then want to select a design that minimizes this criterion. But it is not quite that simple. For GLMs considered here, these variances depend on the unknown parameters. How the two designs compare in terms of $\text{Var}(\hat{\alpha}) + \text{Var}(\hat{\beta})$, or other commonly used criteria, will thus depend on the true, but unknown values of the parameters. This leads to the concept of *locally optimal designs*. The experimenter offers one or more “guesses” for the true values of the parameters, and at each of these we can compare the two designs. The first may be better for some guessed values, the second for

others. For any guess of α and β , the best design among all possible designs is said to be locally optimal for that guess.

Continuing with our example, suppose that we have guessed $\alpha = -3$ and $\beta = 10$. We will therefore compare the two designs under the assumption that these are the true parameter values. Under this assumption, it can be shown that $\text{Var}(\hat{\alpha}) + \text{Var}(\hat{\beta})$ for design I is 40% smaller than for design II. This means that with design I, we need only 60% of the number of subjects that would be needed for design II in order to achieve the same efficiency (in terms of $\text{Var}(\hat{\alpha}) + \text{Var}(\hat{\beta})$).

Moreover, it can be shown that design I remains more efficient than design II under this optimality criterion if our guess for the true parameter values is slightly off.

This example shows that judicious selection of a design can make a big difference, but also shows that the problem of selecting a good or optimal design is a fairly complicated one.

One of the difficulties compared with linear models is that GLMs present a much broader class of models. While linear models are all of the form $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, often accompanied by assumptions of independence and normality, in GLMs, the form of the relationship between the response and the regression variables depends on a *link function* and on distributional assumptions for \mathbf{Y} , which vary from one model to the next. This means that it is very difficult to establish unifying and overarching results, and that the mathematics becomes more difficult and messier. Until recently, many of the available results to identify optimal designs were obtained via the so-called *geometric approach*. This method is inspired by the pioneering work of Elfving (1952) for linear models. But for a long time, it meant that results could only be obtained on a case-by-case basis. Each combination of model, optimality criterion, and objective required its own proof. A number of very fine papers have appeared over the years that address optimal designs for several such combinations. But there is a limit to what can be handled with this approach.

Another difficulty, as already alluded to in Example 4.1, is that design comparisons depend on the unknown parameters. The principal cause for this is that the Fisher information matrix for the parameters in a GLM tends to depend on the parameters. Thus the challenge in designing an experiment for such a model is that one would like to find the best design for estimating the unknown parameters, and yet one has to know the parameters to find the best design. As explained in the context of the example, one way to solve this problem is to use locally optimal designs, which are based on the best guess of the parameters. Even when a good guess is not available, knowing locally optimal designs is valuable because they provide benchmarks for all other designs. An alternative that will not be pursued in this chapter is to adopt a Bayesian approach, which naturally facilitates the inclusion of some uncertainty about a guess for the unknown parameters. This also requires some prior knowledge about the parameters, and may only lead to a computational solution for specific problems without providing broader insights.

Another way to deal with the dependence on unknown parameters is, whenever feasible, through a multistage approach (see Silvey 1980; Sitter and Forbes 1997). In a multistage approach, we would start with a small initial design to obtain some information about the unknown parameters. This information is then used at the next stage to estimate the true parameter values and to augment the initial design in an optimal way. The resulting augmented design could be the final design, or there could be additional stages at which more design points are added. This is also a mathematically difficult problem, but as we will see it is not more difficult than using a single-stage approach. Since the multistage approach uses information obtained in earlier stages, one may hope that it will lead in the end to a better overall design than can be obtained with the single-stage approach.

Khuri et al. (2006) surveyed design issues for GLMs and noted (p. 395) that “The research on designs for generalized linear models is still very much in its developmental stage. Not much work has been accomplished either in terms of theory or in terms of computational methods to evaluate the optimal design when the dimension of the design space is high. The situations where one has several covariates (control variables) or multiple responses corresponding to each subject demand extensive work to evaluate ‘optimal’ or at least efficient designs.”

There are, however, a number of recent papers that have made major advances in this area. These include Biedermann, Dette, and Zhu (2006), Yang and Stufken (2009), and (on nonlinear models in general) Yang (2010). These papers tackle some of the aforementioned difficulties, the first by further exploring the geometric approach and the other two by a new analytic approach. These papers convincingly demonstrate that unifying results for multiple models, multiple optimality criteria, and multiple objectives can be obtained in the context of nonlinear models. While these papers do provide many answers, they also leave many open questions, especially with regard to slightly more complicated models.

This chapter will provide an introduction to the general problem and a peek at available results. The emphasis will be on the analytic approach. In Section 4.2, we introduce notation and basic concepts, such as the information matrix and optimality criteria. Some tools, including Kiefer’s equivalence theorem, Elfving’s geometric approach, and the new analytic approach are presented in Section 4.3. Some optimality results for the simplest GLMs are introduced in Section 4.4. In Section 4.5, we study GLMs with multiple covariates and with block effects. We conclude with some brief remarks in Section 4.6.

4.2 NOTATION AND BASIC CONCEPTS

While GLMs can be appropriate for many types of data, the focus in this chapter is on binary and count data. Let Y denote the response variable. In a

GLM, a *link function* G relates $E(Y)$ to a linear combination of the regression variables, that is, $G(E(Y)) = \mathbf{X}^T \boldsymbol{\theta}$.

4.2.1 Binary Data

If Y_1, \dots, Y_n denote the binary response variables for n subjects, and x_{i1}, \dots, x_{ip} denote the values of p regression variables for subject i , then a class of GLMs can be written as

$$\text{Prob}(Y_i = 1) = P(\mathbf{X}_i^T \boldsymbol{\theta}). \quad (4.2)$$

Here, the superscript T denotes matrix transposition, $\mathbf{X}_i = (1, x_{i1}, \dots, x_{ip})^T$, the vector $\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_p)^T$ contains the unknown regression parameters, and $P(x)$ is a cumulative distribution function (cdf). Popular choices for the latter include $P(x) = e^x/(1 + e^x)$ for the logistic model, as in Equation (4.1) and $P(x) = \Phi(x)$, the cdf of the standard normal distribution, for the probit model. Other choices include the double exponential and double reciprocal models. The inverse of $P(x)$ is the *link function* for these GLMs, that is, $P^{-1}(\text{Prob}(Y_i = 1)) = \mathbf{X}_i^T \boldsymbol{\theta}$. For the logistic model, this link function corresponds to the *logit* function of Example 4.1.

The likelihood function for $\boldsymbol{\theta}$ under Equation (4.2) can be written as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n P(\mathbf{X}_i^T \boldsymbol{\theta})^{Y_i} (1 - P(\mathbf{X}_i^T \boldsymbol{\theta}))^{(1-Y_i)}. \quad (4.3)$$

The resulting likelihood equations are

$$\sum_{i=1}^n \mathbf{X}_i \frac{[Y_i - P(\mathbf{X}_i^T \boldsymbol{\theta})]P'(\mathbf{X}_i^T \boldsymbol{\theta})}{P(\mathbf{X}_i^T \boldsymbol{\theta})(1 - P(\mathbf{X}_i^T \boldsymbol{\theta}))} = \mathbf{0}. \quad (4.4)$$

The MLE of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, is obtained by numerically solving these nonlinear equations, such as in statistical software packages as SAS and SPSS. The likelihood function can also be used to obtain asymptotic covariance matrices for functions of $\hat{\boldsymbol{\theta}}$ that are of interest. To do this, we will need a generalized inverse of the Fisher information matrix. The information matrix can be computed as

$$E\left(-\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) = \sum_{i=1}^n \mathbf{I}_{\mathbf{X}_i} = \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T \frac{[P'(\mathbf{X}_i^T \boldsymbol{\theta})]^2}{P(\mathbf{X}_i^T \boldsymbol{\theta})(1 - P(\mathbf{X}_i^T \boldsymbol{\theta}))}. \quad (4.5)$$

$\mathbf{I}_{\mathbf{X}_i}$ is the information matrix for $\boldsymbol{\theta}$ at a single design point \mathbf{X}_i .

4.2.2 Count Data

For count data Y_1, \dots, Y_n , we will assume the model for Y_i to be a Poisson regression model with mean λ_i . Using the same notation as in Section 4.2.1 and using the logarithm as the link function, we have

$$\log(\lambda_i) = \mathbf{X}_i^T \boldsymbol{\theta}. \quad (4.6)$$

The likelihood function for $\boldsymbol{\theta}$ under Equation (4.6) can be written as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n \frac{\exp(-\lambda_i) \lambda_i^{Y_i}}{Y_i!}, \quad (4.7)$$

which results in the likelihood equations

$$\sum_{i=1}^n \mathbf{X}_i (Y_i - \exp(\mathbf{X}_i^T \boldsymbol{\theta})) = \mathbf{0}. \quad (4.8)$$

The MLE for $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, can again be obtained numerically by solving these nonlinear equations. The information matrix for $\boldsymbol{\theta}$ under Equation (4.6) can be written as

$$E\left(-\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) = \sum_{i=1}^n \mathbf{I}_{\mathbf{X}_i} = \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T \exp(\mathbf{X}_i^T \boldsymbol{\theta}). \quad (4.9)$$

4.2.3 Optimality Criteria

For an exact design with a total of n subjects, we must select (1) distinct vectors $\mathbf{X}_1, \dots, \mathbf{X}_k$, as defined in Section 4.2.1, in a design space, say \mathcal{X} ; and (2) the number of subjects, n_i , to be assigned to \mathbf{X}_i so that $n = \sum_{i=1}^k n_i$. The \mathbf{X}_i 's are also called the *support points* for the design. The optimal exact design problem is to make such selections so that the resulting design is best with respect to a certain optimality criterion. As already alluded to in Example 4.1, instead of this typically intractable exact design problem, the corresponding approximate design problem is considered. Thus, we would like to find a design $\xi = \{(\mathbf{X}_i, \omega_i), i = 1, \dots, k\}$, where the ω_i 's are nonnegative weights that sum to 1. Thus, ω_i represents the proportion of subjects that are to be assigned to \mathbf{X}_i . The corresponding information matrix for $\boldsymbol{\theta}$ can be written as

$$\mathbf{I}_\xi = \sum_{i=1}^k \omega_i \mathbf{I}_{\mathbf{X}_i}, \quad (4.10)$$

where $\mathbf{I}_{\mathbf{X}_i}$ is again the information matrix for $\boldsymbol{\theta}$ for a one-point design that only uses \mathbf{X}_i . If there is an exact design for n subjects corresponding to ξ (which requires $n\omega_i$ to be integral for all i), then the information matrix for this exact design is n times the matrix shown in Equation (4.10).

The interest of the experimenter may not always be in $\boldsymbol{\theta}$, but could be in a vector function of $\boldsymbol{\theta}$, say $g(\boldsymbol{\theta})$. By the Delta method, the approximate covariance matrix of $g(\hat{\boldsymbol{\theta}})$ under design ξ is equal to

$$\boldsymbol{\Sigma}_\xi = \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right) \mathbf{I}_\xi \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right)^T, \quad (4.11)$$

where \mathbf{I}_ξ^- is a generalized inverse of the information matrix in Equation (4.10). In some situations, we could be interested in a function $g(\boldsymbol{\theta})$ for which the matrix $\boldsymbol{\Sigma}_\xi$ is singular for any design ξ . For example, this would happen if the elements of $g(\boldsymbol{\theta})$ are linearly dependent, such as for $g(\boldsymbol{\theta}) = (\theta_1, \theta_2, \theta_1 + \theta_2)^T$. We will, however, limit our interest in this chapter to functions $g(\boldsymbol{\theta})$ for which there are designs that make $\boldsymbol{\Sigma}_\xi$ nonsingular. In particular, if we are interested in $g(\boldsymbol{\theta}) = \boldsymbol{\theta}$ or if g is a one-to-one function, the parametrization should be such that there are designs for which \mathbf{I}_ξ^- is nonsingular.

For designs for which the inverse of the matrix in Equation (4.11) exists, this inverse is the information matrix for $g(\boldsymbol{\theta})$ under design ξ . We would like to select a design that in some sense, maximizes this information matrix, or minimizes the covariance matrix. The following are some of the more prominent optimality criteria that suggest how we might want to do this. For the first three of these, the minimization is over those designs ξ for which $\boldsymbol{\Sigma}_\xi$ is nonsingular. For the fourth criterion, ξ should be a design so that the vector \mathbf{c} is not in the null space of $\boldsymbol{\Sigma}_\xi$.

- *D-Optimality.* A design is *D-optimal* for $g(\boldsymbol{\theta})$ if it minimizes $|\boldsymbol{\Sigma}_\xi|$ over all possible designs. Such a design minimizes the expected volume of the asymptotic $100(1 - \alpha)\%$ joint confidence ellipsoid for the elements of $g(\boldsymbol{\theta})$. For a one-to-one transformation $h(\boldsymbol{\theta})$ of $g(\boldsymbol{\theta})$, if ξ is *D-optimal* for $g(\boldsymbol{\theta})$, then the same holds for $h(\boldsymbol{\theta})$. Thus *D-optimality* is invariant under such transformations. Many other optimality criteria do not have this property.
- *A-Optimality.* A design is *A-optimal* for $g(\boldsymbol{\theta})$ if it minimizes $Tr(\boldsymbol{\Sigma}_\xi)$ over all possible designs. Such a design minimizes the sum of the asymptotic variances of the estimators of the elements of $g(\boldsymbol{\theta})$.
- *E-Optimality.* A design is *E-optimal* for $g(\boldsymbol{\theta})$ if it minimizes the largest eigenvalue of $\boldsymbol{\Sigma}_\xi$ over all possible designs. Such a design minimizes the expected length of the longest semi-axis of the asymptotic $100(1 - \alpha)\%$ joint confidence ellipsoid for the elements of $g(\boldsymbol{\theta})$.
- *c-Optimality.* A design is *c-optimal* for $g(\boldsymbol{\theta})$ if it minimizes $\mathbf{c}^T \boldsymbol{\Sigma}_\xi \mathbf{c}$ over all possible designs, where \mathbf{c} is a vector of the same length as $g(\boldsymbol{\theta})$. Such a design minimizes the asymptotic variance of $\mathbf{c}^T g(\boldsymbol{\theta})$.

To facilitate simultaneous study of some of the above and additional criteria, Kiefer (1974) introduced, among others, the class of functions

$$\Phi_p(\boldsymbol{\Sigma}_\xi) = \left[\frac{1}{v} Tr((\boldsymbol{\Sigma}_\xi)^p) \right]^{1/p}, \quad 0 < p < \infty. \quad (4.12)$$

Here, v is the dimension of $\boldsymbol{\Sigma}_\xi$. A design is Φ_p -optimal for $g(\boldsymbol{\theta})$ if it minimizes $\Phi_p(\boldsymbol{\Sigma}_\xi)$ over all possible designs. In addition, we define $\Phi_0(\boldsymbol{\Sigma}_\xi) = \lim_{p \downarrow 0} \Phi_p(\boldsymbol{\Sigma}_\xi)$ and $\Phi_\infty(\boldsymbol{\Sigma}_\xi) = \lim_{p \rightarrow \infty} \Phi_p(\boldsymbol{\Sigma}_\xi)$. Obviously, Φ_1 -optimality is equivalent to *A-*

optimality. It can be shown that Φ_0 -optimality corresponds to D -optimality and Φ_∞ -optimality to E -optimality.

Which optimality criterion one should use may depend on the objective of the experiment, but also on personal preference (see also HK2, section 1.13.4). One compromise could be to find a design that performs well under multiple criteria. In order to assess whether a design performs well under different criteria, one would however first need to know optimal designs under these criteria. In this chapter, we will only focus on tools for identifying optimal designs.

To emphasize an earlier observation, the information matrix \mathbf{I}_ξ and the covariance matrix Σ_ξ depend, for the GLMs considered here, on $\boldsymbol{\theta}$. Thus, none of the above minimizations can be carried out, unless we have a “guess” for $\boldsymbol{\theta}$. This approach, resulting in locally optimal designs, is taken here.

4.3 TOOLS FOR FINDING LOCALLY OPTIMAL DESIGNS

Once we have decided on a model, a function $g(\boldsymbol{\theta})$ of interest, an optimality criterion, and a guess for the parameter values, we are ready to identify a locally optimal design by optimizing the objective function corresponding to the selected criterion. (We will often drop the adjective “locally” from hereon, but the reader should remember that the discussion in this chapter is always about locally optimal designs.) However, this is a very challenging problem. How many support points do we need in an optimal design? What are those points? What are the corresponding weights? Directly optimizing the objective function is generally not feasible because the objective function is too complicated and there are too many variables. Moreover, even if a purely numerical optimization were feasible, it might not provide enough insight into the structure and general features of optimal designs. There are two standard traditional tools for studying optimal designs that have inspired many researchers: Elfving’s geometric approach and Kiefer’s equivalence theorem. The emphasis in this chapter is, however, on a new analytical approach, although we will give the reader also a flavor of the traditional tools.

4.3.1 Traditional Approaches

The geometric approach proposed by Elfving (1952) for linear models has had a profound impact on optimal design theory. Whereas Elfving was interested in c - and A -optimal experimental designs for linear models in two dimensions, his work has proven to be inspirational for the development of optimal design theory in a much broader framework.

To describe the basic idea, write the information matrix for $\boldsymbol{\theta}$ as $\Sigma_i \omega_i f(X_i, \boldsymbol{\theta}) f(X_i, \boldsymbol{\theta})^T$, where $f(X_i, \boldsymbol{\theta})$ is a column vector that depends on the design point X_i and the parameter vector $\boldsymbol{\theta}$. Note that the information matrices in Equations (4.5) and (4.9), or, more precisely, those for the corresponding approximate

designs, are of that form with $f(\mathbf{X}, \boldsymbol{\theta})$ being a multiple of \mathbf{X} . Let \mathcal{G} (which depends on $\boldsymbol{\theta}$) be the space that is generated through $f(\mathbf{X}, \boldsymbol{\theta})$ by letting \mathbf{X} take all possible values in the design space X . Assume that \mathcal{G} , which is called the *induced design space*, is closed and bounded. The support points for an optimal design must lie on the “smallest ellipsoid” centered at the origin that contains \mathcal{G} . The definition of “smallest ellipsoid” depends on the optimality criterion but can be stated explicitly. Studying these ellipsoids and their intersections with the induced design space provides, therefore, a method for determining possible support points for an optimal design. Many optimality results in the literature are based on variations of this approach. Some seminal contributions include Ford, Torsney, and Wu (1992) for c - and D -optimal designs; Dette and Haines (1994) for E -optimal designs; and Biedermann, Dette, and Zhu (2006) for Φ_p -optimal designs. We refer the reader to these articles and their references to learn more about this approach.

A second result that has had a profound impact on research in optimal design is the equivalence theorem. In a seminal contribution, Kiefer and Wolfowitz (1960) derived an equivalence theorem for D -optimality. Later, Kiefer (1974) presented a more general result that applies for Φ_p optimality. Pukelsheim (2006) contains very general results on equivalence theorems. All of these studies focus on optimality for linear functions of the parameters under linear models. However, once a value of $\boldsymbol{\theta}$ is fixed under the local optimality approach, then the results extend under mild conditions. The following result is formulated in the spirit of Pukelsheim (2006).

Theorem 4.1 (Equivalence Theorem). Suppose that the information matrix for $\boldsymbol{\theta}$ under a design $\xi = \{(\mathbf{X}_i, \omega_i), i = 1, \dots, k\}$ is given by $\Sigma_i \omega_i f(\mathbf{X}_i, \boldsymbol{\theta}) f(\mathbf{X}_i, \boldsymbol{\theta})^T$. A design ξ^* is Φ_p -optimal for $g(\boldsymbol{\theta})$, $0 \leq p < \infty$, if and only if there exists a generalized inverse $\mathbf{I}_{\xi^*}^-$ of \mathbf{I}_{ξ^*} so that

$$f(\mathbf{X}, \boldsymbol{\theta})^T \mathbf{I}_{\xi^*}^- \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right)^T (\Sigma_{\xi^*})^{p-1} \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right) \mathbf{I}_{\xi^*}^- f(\mathbf{X}, \boldsymbol{\theta}) \leq \text{Tr}((\Sigma_{\xi^*})^p), \text{ for all } \mathbf{X} \in \mathcal{X}, \quad (4.13)$$

where Σ_ξ is as defined in Equation (4.11) and \mathcal{X} is the design space. Equality in Equation (4.13) holds if and only if \mathbf{X} is a support point for a Φ_p -optimal design.

The equivalence theorem is very useful to verify whether a candidate design is indeed optimal. For the case that $p = 0$, corresponding to D -optimality, the right-hand side of Equation (4.13) reduces simply to the dimension of $g(\boldsymbol{\theta})$.

4.3.2 An Analytical Approach

The equivalence theorem and the geometric approach are very powerful tools for studying optimal designs for GLMs. Nonetheless, there are many problems

of great practical interest for which neither of these tools has, as of yet, helped to provide a solution. Yang and Stufken (2009) propose a new strategy for studying optimal designs for GLMs. The remainder of this chapter will discuss this strategy and will present some of the results that have been obtained by using it.

For a given model and design space, suppose that we can identify a subclass of designs, say Ξ , so that for any design $\xi \notin \Xi$, there is a design $\tilde{\xi} \in \Xi$, so that $I_{\tilde{\xi}}(\boldsymbol{\theta}) \geq I_{\xi}(\boldsymbol{\theta})$, that is, the information matrix for $\boldsymbol{\theta}$ under $\tilde{\xi}$ dominates that under ξ in the Loewner ordering. Then $\tilde{\xi}$ is locally at least as good as ξ under the commonly used optimality criteria, such as D -, A -, and E -optimality, and, more generally, Φ_p -optimality. Moreover, if interest is in $g(\boldsymbol{\theta})$, design $\tilde{\xi}$ is also better than design ξ in the Loewner ordering since

$$\Sigma_{\xi} = \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right) I_{\tilde{\xi}}^{-1} \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right)^T \geq \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right) I_{\xi}^{-1} \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right)^T = \Sigma_{\xi}. \quad (4.14)$$

This would mean that the search for optimal designs could be restricted to a search in Ξ . We will refer to Ξ as a *complete class* for this problem. One trivial choice for Ξ is the class of all possible designs. This choice is not useful. In order for Ξ to be helpful, it needs to be a small class. For example, for some models, Ξ might perhaps consist of “all designs with at most three support points.” That would be an enormous reduction from having to consider all possible designs with any number of support points. In order to be useful, the approach must also work for any “guess” of $\boldsymbol{\theta}$ under the local optimality approach. Note also from our formulation that the choice of Ξ is not based on any of the common optimality criteria and does not depend on a particular choice for $g(\boldsymbol{\theta})$. So we would like to use the same complete class Ξ no matter what the guess for $\boldsymbol{\theta}$ is, no matter which function $g(\boldsymbol{\theta})$ we are interested in, and no matter which information-based optimality criterion we have in mind. Ξ will depend on the model and the design space.

Thus, for a given model and design space, the approach consists of identifying small complete classes Ξ .

Before we continue, we observe that this approach is also helpful for multistage experiments, where an initial experiment may be used to get a better idea about the unknown parameters. At a second or later stage, the question then becomes how to add more design points in an optimal fashion. If ξ_1 denotes the design used so far for n_1 design points, and we want to augment this design optimally by n_2 additional design points, then we are looking for an approximate design ξ_2 that maximizes the combined information matrix $n_1 I_{\xi_1} + n_2 I_{\xi_2}$. Because the first part of this matrix is fixed, if we have a complete class Ξ for the single-stage approach, it is immediately clear that we can restrict our choice for ξ_2 to Ξ in order to obtain a combined information matrix that can not be beaten in the Loewner ordering by a choice of ξ_2 that is not in Ξ . Thus Ξ is also a complete class for the multistage approach.

The strategy described in the previous paragraphs was used in Yang and Stufken (2009). They characterized complete classes Ξ for nonlinear models (including GLMs) with two parameters. We summarize their results here for GLMs only. For proofs and results for other nonlinear models with two parameters, we refer the reader to Yang and Stufken (2009).

In the context of GLMs, and using the notation of Section 4.2, results are for two-parameter models with $\mathbf{X}_i^T \boldsymbol{\theta} = \theta_0 + \theta_1 x_{i1}$. We define $c_i = \theta_0 + \theta_1 x_{i1}$, and call this the *induced design point*. Note that under local optimality, using guessed values for θ_0 and θ_1 , a design can be expressed in terms of choices for x_{i1} , or, equivalently, in terms of the induced design points c_i .

For a design $\xi = \{(c_i, \omega_i), i = 1, \dots, k\}$, write the information matrix for $\boldsymbol{\theta}$ as $\mathbf{I}_\xi = \sum_{i=1}^k \omega_i \mathbf{I}_{c_i}$, and write \mathbf{I}_{c_i} as $\mathbf{A}^T \mathbf{C}(c_i) \mathbf{A}$. Here matrix \mathbf{A} may depend on $\boldsymbol{\theta}$, but not on the induced design point c_i . We write the matrix $\mathbf{C}(c)$, which can depend on $\boldsymbol{\theta}$ and the induced design points, as

$$\mathbf{C}(c) = \begin{pmatrix} \Psi_1(c) & \Psi_2(c) \\ \Psi_2(c) & \Psi_3(c) \end{pmatrix}. \quad (4.15)$$

Define

$$F(c) = \Psi'_1(c) \left(\frac{\Psi'_2(c)}{\Psi'_1(c)} \right)' \left(\left(\frac{\Psi'_3(c)}{\Psi'_1(c)} \right)' \Big/ \left(\frac{\Psi'_2(c)}{\Psi'_1(c)} \right)' \right). \quad (4.16)$$

We will assume that the design space, in terms of the induced design points, is an interval $[D_1, D_2]$ (where the end points can for some applications be $-\infty$ or ∞ , respectively).

Theorem 4.2. Assume that $F(c)$ is well-defined for $c \notin [D_1, D_2]$. If $F(c) \leq 0$ for all $c \notin [D_1, D_2]$, then a complete class Ξ is obtained by taking all designs with at most two support points, with D_1 being one of them. If $F(c) \geq 0$ for all $c \notin [D_1, D_2]$, then the class of all designs with at most two support points and D_2 being one of them forms a complete class.

Theorem 4.2 requires $F(c)$ to be well defined and either positive or negative in the entire interval $[D_1, D_2]$. These requirements can be relaxed as stated in the following result.

Theorem 4.3. For the matrix $\mathbf{C}(c)$ in Equation (4.15), suppose that $\Psi_1(c) = \Psi_1(-c)$, $\Psi_2(c) = -\Psi_2(-c)$, and $\Psi_3(c) = \Psi_3(-c)$. Suppose further that $D_1 < 0 < D_2$, and that $F(c)$ is well-defined for $c \notin (0, D_2]$. Let $F(c) < 0$ and $\Psi'_1(c)(\Psi'_3(c)/\Psi'_1(c)) < 0$ for $c \notin (0, D_2]$. Then a complete class Ξ is obtained by taking all designs with at most two support points with the additional following restrictions: If $|D_1| = D_2$, the two points can be taken to be symmetric in the induced design space; if $|D_1| < D_2$, then the two points can be taken to be either

symmetric or one of the points is taken as D_1 and the other in $(-D_1, D_2]$; if $|D_1| > D_2$, then the two points can be taken to be either symmetric or one of the points is taken as D_2 and the other in $[D_1, -D_2]$.

For the proofs of these results, we refer to Yang and Stufken (2009). The next sections will provide some applications of these results.

4.4 GLMs WITH TWO PARAMETERS

In this section, we will focus on GLMs with $\mathbf{X}_i^T \boldsymbol{\theta} = \theta_0 + \theta_1 x_i$ for the binary data model Equation (4.2) and the count data model Equation (4.6). For simplicity of notation, we replace the parameters by α and β , that is we replace $\boldsymbol{\theta} = (\theta_0, \theta_1)^T$ by $\boldsymbol{\theta} = (\alpha, \beta)^T$. By the expressions for the information matrix provided in Equations (4.5) and (4.9), the information matrix for $\boldsymbol{\theta}$ can be written as

$$\mathbf{I}_{\xi} = \sum_{i=1}^k \omega_i \Psi(\alpha + \beta x_i) \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}, \quad (4.17)$$

where, $\xi = \{(x_i, \omega_i), i = 1, \dots, k\}$. The function $\Psi(x)$ depends on the model and takes the following forms for the three models that we will focus on:

$$\Psi(x) = \begin{cases} \frac{e^x}{(1+e^x)^2} & \text{for the logistic model} \\ \frac{\phi^2(x)}{\Phi(x)(1-\Phi(x))} & \text{for the probit model} \\ e^x & \text{for the Poisson regression model,} \end{cases} \quad (4.18)$$

where $\phi(x)$ is the density function for the standard normal distribution.

These simple models have been studied extensively in the optimal design literature. For binary data, under the restriction of symmetry for the induced design space, Abdelbasit and Plackett (1983) identify a two-point D -optimal design for the logistic model. Minkin (1987) strengthens this result by removing the restriction on the design space. Using the geometric approach (see Section 4.3.1), Ford, Torsney, and Wu (1992) study c -optimal and D -optimal designs for this model. Using the same approach and model, Sitter and Wu (1993a, 1993b) study A - and F -optimal designs, while Dette and Haines (1994) investigate E -optimal designs. Mathew and Sinha (2001) obtain a series of optimality results for the logistic model by using an analytic approach, while Biedermann, Dette, and Zhu (2006) obtained Φ_p -optimal designs for a restricted design space using the geometric approach. Chaloner and Larntz (1989) and Agin and Chaloner (1999) study Bayesian optimal designs for the logistic model. For count data, Ford, Torsney, and Wu (1992) identified c - and D -optimal designs. Minkin (1993) studied optimal designs for $1/\beta$.

In this section, we will present optimal designs for the above models by using the analytic approach presented in Section 4.3. As we will see, this approach generalizes and extends most of the available results. Perhaps even more importantly, in the next section, we will see that this approach also facilitates handling GLMs with more than two parameters.

From Equation (4.17), the information matrix \mathbf{I}_x for $(\alpha, \beta)^T$ under a one-point design with all weight at x can be written as

$$\mathbf{I}_x = \underbrace{\begin{pmatrix} 1 & 0 \\ -\alpha/\beta & 1/\beta \end{pmatrix}}_{A^T} \begin{pmatrix} \Psi(c) & c\Psi(c) \\ c\Psi(c) & c^2\Psi(c) \end{pmatrix} \underbrace{\begin{pmatrix} 1 & -\alpha/\beta \\ 0 & 1/\beta \end{pmatrix}}_A, \quad (4.19)$$

where $c = \alpha + \beta x$. Note that we are assuming that neither α nor β is equal to 0, as we will do throughout. Clearly, matrix A depends only on the parameters α and β , and not on x . Thus we can apply the analytic approach in Section 4.3 for these simple models.

We will first consider the Poisson regression model, where $\Psi_1(c) = e^c$, $\Psi_2(c) = ce^c$, and $\Psi_3(c) = c^2e^c$. We find that $\Psi'_1(c) = e^c$, $(\Psi'_2(c)/\Psi'_1(c))' = 1$, and $((\Psi'_3(c)/\Psi'_1(c))' / (\Psi'_2(c)/\Psi'_1(c))')' = 2$. Hence, the function $F(c)$ defined in Equation (4.16) is positive in any interval. Applying Theorem 4.2, we reach the following conclusion.

Theorem 4.4. For the Poisson regression model in Equation (4.6) with $c_i = \mathbf{X}_i^T \boldsymbol{\theta} = \alpha + \beta x_i$, let $c_i \notin [D_1, D_2]$ be the induced design space, $D_1 < D_2 < \infty$. Then, the class of designs with at most two support points and D_2 being one of them forms a complete class.

Thus Theorem 4.4 tells us that when searching for an optimal design, we need to look no further than the complete class described in the theorem. Given how simple the designs in this class are, it is easy to use a search algorithm to do this for a given function $g(\boldsymbol{\theta})$, a given optimality criterion, and a guess for the parameters. The algorithm would merely have to optimize a function of interest over two unknowns: the second support point and the weight for that point. Moreover, depending on the problem, it may actually be possible to derive an explicit form for the optimal design by using Theorem 4.4.

Example 4.2. By Theorem 4.4, for the Poisson regression model, a D -optimal design for $(\alpha, \beta)^T$ can be based on D_2 and $c < D_2$, where c needs to be determined. The weight for each of these support points must be equal to $\frac{1}{2}$ (see Silvey 1980). Computing the determinant of the information matrix in Equation (4.17), it is easily seen that the D -optimal design must maximize $e^{c/2}(D_2 - c)$. As a function of c , this is an increasing function on $(-\infty, D_2 - 2]$ and a decreasing function on $[D_2 - 2, D_2]$. Thus, if $D_1 < D_2 - 2$, then a D -optimal design is

given by $\xi = \{(D_2 - 2, 1/2), (D_2, 1/2)\}$. Otherwise, a D -optimal design is given by $\xi = \{(D_1, 1/2), (D_2, 1/2)\}$.

Theorem 4.4 unifies and extends a number of results that had already appeared in the literature. For example, Ford, Torsney, and Wu (1992) identified c - and D -optimal designs using the geometric approach. They showed that an optimal design has two support points and that one of them is D_2 . Minkin (1993) studied optimal designs for $1/\beta$ for this model. He assumed $\beta < 0$ and used the induced design space $(-\infty, \alpha]$. He concluded that the optimal design has two support points, and that one of them is α .

Turning now to the logistic and probit models, both have the properties that $\Psi_1(c) = \Psi_1(-c)$, $\Psi_2(c) = -\Psi_2(-c)$, and $\Psi_3(c) = \Psi_3(-c)$. Theorems 4.2 and 4.3 can thus be applied as long as other conditions in the theorems are satisfied. For the logistic model, it is easily seen that

$$\begin{aligned}\Psi'_1(c) &= -\frac{e^c(e^c - 1)}{(e^c + 1)^3}, \\ \left(\frac{\Psi'_2(c)}{\Psi'_1(c)}\right)' &= \frac{e^{2c} + 1}{(e^c - 1)^2}, \\ \left(\frac{\Psi'_3(c)}{\Psi'_1(c)}\right)' &= \frac{2((c-1)e^{2c} + c + 1)}{(e^c - 1)^2}, \\ \left(\left(\frac{\Psi'_3(c)}{\Psi'_1(c)}\right)' \Big/ \left(\frac{\Psi'_2(c)}{\Psi'_1(c)}\right)'\right)' &= \frac{2(e^{2c} - 1)^2}{(e^{2c} + 1)^2}.\end{aligned}\tag{4.20}$$

Hence, if the induced design space $[D_1, D_2] \subset (-\infty, 0]$, then $F(c) > 0$; if the induced design space $[D_1, D_2] \subset [0, \infty)$, then $F(c) < 0$; further, $\Psi'_1(c)(\Psi'_3(c)/\Psi'_1(c))' < 0$ for $c > 0$. These same conclusions are also valid for the probit model, but we will skip the much more tedious details. Applying Theorems 4.2 and 4.3, we reach the following conclusion.

Theorem 4.5. For the logistic or probit model as in Equation (4.2) with $c_i = \mathbf{X}_i^T \boldsymbol{\theta} = \alpha + \beta x_i$, let $c_i \notin [D_1, D_2]$ be the induced design space. Then the following complete class results hold:

1. if $D_2 \leq 0$, then the class of designs with at most two support points and D_2 being one of them forms a complete class;
2. if $D_1 \geq 0$, then the class of designs with at most two support points and D_1 being one of them forms a complete class;
3. if $D_2 = -D_1$, then the class of designs with at most two support points that are symmetric forms a complete class;

4. if $0 < -D_1 < D_2$, then the class of designs with at most two support points, where either one of the points is D_1 and the other point is larger than $-D_1$, or the two points are symmetric, forms a complete class; and
5. if $0 < D_2 < -D_1$, then the class of designs with at most two support points, where either one of the points is D_2 and the other point is smaller than $-D_2$ or the two points are symmetric, forms a complete class.

Example 4.3. By Theorem 4.5, for the logistic and probit models, D -optimal designs for $(\alpha, \beta)^T$ can be based on two support points, say c_1 and c_2 . The weights for these support points must be equal to $\frac{1}{2}$ (see Silvey 1980). Using Equation (4.17), it is easily seen that the D -optimal design must maximize $(c_1 - c_2)^2 \Psi(c_1) \Psi(c_2)$. Because of the relationship between c_1 and c_2 provided in Theorem 4.5, there is essentially one unknown to be decided. For example, if $D_2 = -D_1$, so that $c_1 = -c_2 = c$, say, we need to maximize $4c^2 \Psi^2(c)$. For the logistic model, this is an increasing function for $c \in [0, 1.5434]$ and a decreasing function for $c > 1.5434$. (For the probit model, the critical value for c is 1.1381.) Thus, if $D_2 > 1.5434$, then a D -optimal design is given by $\xi = \{(-1.5434, 1/2), (1.5434, 1/2)\}$. Otherwise, a D -optimal design is given by $\xi = \{(-D_2, 1/2), (D_2, 1/2)\}$.

Example 4.4. Again using the logistic or probit model, suppose that $D_1 > 0$. By Theorem 4.5, we can take D_1 as one of the two support points, and with c , $D_1 < c \leq D_2$, denoting the other support point we need to maximize $(c - D_1)^2 \Psi(c)$ for a D -optimal design. This is an increasing function for $c \in [D_1, c^*]$ and a decreasing function when $c > c^*$. For the logistic model, c^* is the solution of $e^c(2 - c + D_1) + c - D_1 + 2 = 0$. If $c^* < D_2$, then a D -optimal design is given by $\xi = \{(D_1, 1/2), (c^*, 1/2)\}$. Otherwise, a D -optimal design is given by $\xi = \{(D_1, 1/2), (D_2, 1/2)\}$.

Example 4.5. Consider the logistic regression model in Equation (4.2), with $c_i = \alpha + \beta x_i$. Suppose that we are interested in a locally D -optimal design for $\alpha = 2$ and $\beta = -1$, with the design space restricted to $x \in [0, 1]$. Then $[D_1, D_2] = [1, 2]$. We can apply the conclusion in Example 4.4, since $D_1 > 0$. Simple computation shows that $c^* = 3.1745$. Thus, a D -optimal design in terms of the induced design space is given by $\xi = \{(1, 1/2), (2, 1/2)\}$, which is $\{(1, 1/2), (0, 1/2)\}$ in the original x -space.

In Section 4.5.2, we will return to finding D -optimal designs in a slightly more complicated setting. Here, we show how Theorem 4.5 can be used to identify A -optimal designs for $(\alpha, \beta)^T$ under the logistic and probit models when there is no constraint on the design space, that is, $D_1 = -\infty$ and $D_2 = \infty$. This question was posed by Mathew and Sinha (2001) and was studied by Yang (2008). By Theorem 4.5, we can focus on designs with two symmetric induced design points, that is, $\xi = \{(x_1, \omega_1), (x_2, \omega_2)\}$, where $\alpha + \beta x_1 = -\alpha - \beta x_2$. With $c = \alpha + \beta x_1$, and using Equation (4.17) and the observation that Ψ is an even

function for these two models, it follows that the information matrix for $(\alpha, \beta)^T$ under ξ can be written as

$$\mathbf{I}_\xi = \Psi(c) \begin{pmatrix} 1 & \omega_1 \left(\frac{c-\alpha}{\beta} \right) + \omega_2 \left(\frac{-c-\alpha}{\beta} \right) \\ \omega_1 \left(\frac{c-\alpha}{\beta} \right) + \omega_2 \left(\frac{-c-\alpha}{\beta} \right) & \omega_1 \left(\frac{c-\alpha}{\beta} \right)^2 + \omega_2 \left(\frac{-c-\alpha}{\beta} \right)^2 \end{pmatrix}. \quad (4.21)$$

For a locally A -optimal design, we need to minimize the following trace as a function of c and ω_1 (with $\omega_2 = 1 - \omega_1$):

$$\begin{aligned} Tr(\mathbf{I}_\xi^{-1}) &= \frac{1 + \omega_1 \left(\frac{c-\alpha}{\beta} \right)^2 + \omega_2 \left(\frac{-c-\alpha}{\beta} \right)^2}{\left[\omega_1 \left(\frac{c-\alpha}{\beta} \right)^2 + \omega_2 \left(\frac{-c-\alpha}{\beta} \right)^2 - \left(\omega_1 \left(\frac{c-\alpha}{\beta} \right) + \omega_2 \left(\frac{-c-\alpha}{\beta} \right) \right)^2 \right] \Psi(c)} \\ &= \frac{[\beta^2 + (c+\alpha)^2]/\omega_1 + [\beta^2 + (c-\alpha)^2]/\omega_2}{4c^2 \Psi(c)} \\ &\geq T^2(c). \end{aligned} \quad (4.22)$$

Here,

$$T(c) = \frac{\sqrt{\beta^2 + (c+\alpha)^2} + \sqrt{\beta^2 + (c-\alpha)^2}}{2c\sqrt{\Psi(c)}}. \quad (4.23)$$

Note that we do not have to worry about $c = 0$ since that corresponds to a 1-point design, which cannot be optimal for $(\alpha, \beta)^T$. The last inequality of Equation (4.22) gives equality for

$$\omega_1 = 1 - \omega_2 = \frac{\sqrt{\beta^2 + (c+\alpha)^2}}{\sqrt{\beta^2 + (c+\alpha)^2} + \sqrt{\beta^2 + (c-\alpha)^2}} =: \omega_1(c). \quad (4.24)$$

We are now ready to present locally A -optimal designs under the logistic and probit models.

Theorem 4.6. For the logistic or probit model as in Equation (4.2) with $\mathbf{X}_i^T \boldsymbol{\theta} = \alpha + \beta x_i$ and an unrestricted design space, the design $\xi^* = \{(x_1^*, \omega_1^*), (x_2^*, \omega_2^*)\}$ is A -optimal for $(\alpha, \beta)^T$, where $x_1^* = (c^* - \alpha)/\beta$, $x_2^* = (-c^* - \alpha)/\beta$, $\omega_1^* = \omega_1(c^*)$ as defined in Equation (4.24), $\omega_2^* = 1 - \omega_1^*$, and c^* is the only positive solution of the equation

$$\frac{c^2 - \alpha^2 - \beta^2}{\sqrt{\beta^2 + (c + \alpha)^2} \sqrt{\beta^2 + (c - \alpha)^2}} = 1 + \frac{c\Psi'(c)}{\Psi(c)}. \quad (4.25)$$

Proof. Starting from Equations (4.22) and (4.24), we will show that c^* is the unique positive design point that minimizes $T^2(c)$. The restriction to positive design points is justified because $T^2(c)$ is an even function. Since $T(c)$ in Equation (4.23) is positive, we can instead focus on minimizing $T(c)$ for $c > 0$. By straightforward computations, we obtain that

$$T'(c) = \frac{T_1(c)}{2c^2},$$

where

$$\begin{aligned} T_1(c) = & \left(c(\Psi^{-1/2}(c))' - \Psi^{-1/2}(c) \right) \left(\sqrt{\beta^2 + (c + \alpha)^2} + \sqrt{\beta^2 + (c - \alpha)^2} \right) \\ & + \Psi^{-1/2}(c) \left(\frac{c^2 + \alpha c}{\sqrt{\beta^2 + (c + \alpha)^2}} + \frac{c^2 - \alpha c}{\sqrt{\beta^2 + (c - \alpha)^2}} \right). \end{aligned} \quad (4.26)$$

Taking the derivative of $T_1(c)$, we find that

$$\begin{aligned} T_1'(c) = & c(\Psi^{-1/2}(c))'' \left(\sqrt{\beta^2 + (c + \alpha)^2} + \sqrt{\beta^2 + (c - \alpha)^2} \right) \\ & + 2c(\Psi^{-1/2}(c))' \left(\frac{c + \alpha}{\sqrt{\beta^2 + (c + \alpha)^2}} + \frac{c - \alpha}{\sqrt{\beta^2 + (c - \alpha)^2}} \right) \\ & + c\Psi^{-1/2}(c) \left(\frac{\beta^2}{\left(\sqrt{\beta^2 + (c + \alpha)^2} \right)^3} + \frac{\beta^2}{\left(\sqrt{\beta^2 + (c - \alpha)^2} \right)^3} \right). \end{aligned} \quad (4.27)$$

By studying the function $f(x) = x/\sqrt{\beta^2 + x^2}$, it is easily seen that for $c > 0$, it holds that

$$f(c + \alpha) + f(c - \alpha) > 0.$$

Combined with the observation that for both the logistic and probit model $\Psi^{-1/2}(c)$ has positive first- and second-order derivatives for $c > 0$, we can conclude that $T_1'(c) > 0$. Thus $T_1(c)$ is an increasing function for $c > 0$ and has at most one positive root. It follows that $T'(c)$ also has at most one positive root. But since $T(c)$ goes to $+\infty$ for both $c \downarrow 0$ and $c \rightarrow \infty$, $T'(c)$ must also have at least one positive root, which must thus be unique and must minimize $T(c)$. It is fairly straightforward to show that $T_1(c) = 0$ is equivalent to Equation (4.25). The result follows.

Example 4.6. Consider the probit model in Equation (4.2) with $c_i = \alpha + \beta x_i$. Suppose that we are interested in a locally A -optimal design for $\alpha = 1$, $\beta = 2$, and that there is no restriction on the design space. Using Equation (4.25) with these values for α and β and with the choice of Ψ for the probit model as in Equation (4.18), we obtain that c^* in Theorem 4.6 is equal to 1.3744. It follows that an A -optimal design is given by $\{(0.1872, 0.6041), (-1.1872, 0.3959)\}$ in the x -space.

4.5 GLMs WITH MULTIPLE PARAMETERS

Whereas a model with only two parameters is adequate for some applications, other situations call for models with more parameters. For example, the subjects in an experiment may have different characteristics that can be captured by one or more qualitative variables, such as race, gender, or age category. To allow for differences in the relationship between the response variable and a covariate for subjects belonging to different groups, effects associated with these qualitative variables could be included in the model in addition to a covariate, such as stimulus level (see e.g. Tighiouart and Rogatko 2006). As another example, an experimenter may be able to control more than one covariate that has a relationship with the response variable. Optimal design results for GLMs with more than two parameters are relatively rare. Selected results include Sitter and Torsney (1995a, 1995b), who study D -optimal designs for binary data with two and more covariates. Under certain constraints, Haines et al. (2007) study D -optimal designs for logistic regression in two variables. Russell et al. (2009) consider multivariate GLMs for count data. Computationally oriented contributions include Woods et al. (2006) and Dror and Steinberg (2006) for studying robust designs, as well as Dror and Steinberg (2008) for studying sequential designs. They all use D -optimality and provide algorithms for finding desirable designs. For example, Dror and Steinberg (2006) provide computer programs for deriving D -optimal designs for general models. This works best for smaller values of p .

In this section, we will systematically study (1) GLMs with multiple independently chosen covariates; and (2) GLMs with group effects.

4.5.1 GLMs with Multiple Covariates

We will focus on the model in Equation (4.2) for a binary response. We will again take $\mathbf{X}_i^T = (1, x_{i1}, \dots, x_{ip})$, where x_{i1}, \dots, x_{ip} denote the values of p regression variables for subject i that can be selected by the experimenter. We will assume that these values can be selected independently. This implies in particular that there cannot be any functional relationships (such as $x_2 = x_1^2$ or $x_3 = x_1 x_2$) between these covariates. The design space \mathcal{X} is thus a subset of \mathbb{R}^p .

An approximate design can be presented as $\xi = \{(\mathbf{X}_i, \omega_i), i = 1, \dots, k\}$, where ω_i is the weight for the vector \mathbf{X}_i . For the parameter vector $\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_p)^T$,

we will assume that $\theta_p \neq 0$. With $\mathbf{C}_i^T = (1, x_{i1}, \dots, x_{ip-1}, c_i)$ and $c_i = \mathbf{X}_i^T \boldsymbol{\theta}$, there is then a one-to-one relationship between X_i and \mathbf{C}_i . A design may thus also be written as $\xi = \{(\mathbf{C}_i, \omega_i), i = 1, \dots, k\}$.

Using Equation (4.5), and writing $\mathbf{A}^T = \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{A}_1^T & 1/\theta_p \end{pmatrix}$, where $\mathbf{A}_1^T = \frac{1}{\theta_p}(-\theta_0, -\theta_1, \dots, -\theta_{p-1})$, the information matrix for $\boldsymbol{\theta}$ under Equation (4.2) can be written as

$$\mathbf{I}_\xi = \sum_{i=1}^k \omega_i \mathbf{X}_i \Psi(c_i) \mathbf{X}_i^T = \mathbf{A}^T \left(\sum_{i=1}^k \omega_i \mathbf{C}_i \Psi(c_i) \mathbf{C}_i^T \right) \mathbf{A}, \quad (4.28)$$

where $\Psi(x) = [P'(x)]^2/P(x)(1 - P(x))$. Note that for the logistic and probit models, $\Psi(x)$ was presented explicitly in Equation (4.18). Also note that the case of $p = 1$ corresponds to the models in Section 4.4. The above matrix \mathbf{A}^T reduces in that case to the corresponding matrix in Equation (4.19).

As noted in Sitter and Torsney (1995a), unless appropriate constraints are placed on the design space, the information matrix can become arbitrarily large for the case that $p \geq 2$. In applications it may indeed be quite reasonable that each of the covariates can only take values in a bounded interval. In this chapter, we will assume that there are constraints on the first $p - 1$ covariates, but not on x_p . Specifically, we will assume that, for all $i, x_{ij} \in [U_j, V_j], j = 1, \dots, p - 1$, but x_{ip} can take any value in $(-\infty, \infty)$. The main reason for not placing a constraint on x_p is of a technical nature.

We will now show that there are, just as for the two-parameter models in Section 4.4, relatively simple complete classes for the multi-parameter models in this section. Consider a design $\xi = \{(\mathbf{C}_i, \omega_i), i = 1, \dots, k\}$, with $\mathbf{C}_i^T = (1, x_{i1}, \dots, x_{ip-1}, c_i)$. We focus for the moment on x_{ij} for a fixed i and j with $j \leq p - 1$. Define $r_{ij} = V_j - x_{ij}/V_j - U_j$. Note that $r_{ij} \in [0, 1]$. Then, using the convexity of the function x^2 , it is easy to see that

$$\begin{aligned} r_{ij}U_j + (1 - r_{ij})V_j &= x_{ij}, \\ r_{ij}U_j^2 + (1 - r_{ij})V_j^2 &\geq x_{ij}^2. \end{aligned} \quad (4.29)$$

It is now easy to see that if we replace \mathbf{C}_i in ξ by $\mathbf{C}_{ij1} = (1, x_{i1}, \dots, U_j, \dots, x_{ip-1}, c_i)^T$ and $\mathbf{C}_{ij2} = (1, x_{i1}, \dots, V_j, \dots, x_{ip-1}, c_i)^T$ with weights $\omega_{ij1} = r_{ij}\omega_i$ and $\omega_{ij2} = (1 - r_{ij})\omega_i$, respectively, then the matrices $\omega_{ij1} \mathbf{C}_{ij1} \Psi(c_i) \mathbf{C}_{ij1}^T + \omega_{ij2} \mathbf{C}_{ij2} \Psi(c_i) \mathbf{C}_{ij2}^T$ and $\omega_i \mathbf{C}_i \Psi(c_i) \mathbf{C}_i^T$ have the same elements except for the $(j+1)$ th diagonal element corresponding to covariate x_j . That diagonal element is larger in the former matrix than in the latter based on Equation (4.29). Repeating this argument for other values of j , $1 \leq j \leq p - 1$, we conclude that for any design point \mathbf{C}_i , there exist weights $\omega_i^\ell, \ell = 1, \dots, 2^{p-1}$, so that

$$\omega_i \mathbf{C}_i \Psi(c_i) \mathbf{C}_i^T \leq \sum_{\ell=1}^{2^{p-1}} \omega_i^\ell \mathbf{C}_i^\ell \Psi(c_i) (\mathbf{C}_i^\ell)^T. \quad (4.30)$$

The design points \mathbf{C}_i^ℓ are of the form $(\mathbf{C}_i^\ell)^T = (1, a_{\ell 1}, \dots, a_{\ell, p-1}, c_i)$, where $a_{\ell j}$ is either U_j or V_j and $\sum_{\ell=1}^{2^{p-1}} \omega_i^\ell = \omega_i$. Now we are ready to present a complete class result.

Theorem 4.7. For the logistic and probit model as in Equation (4.2) with $c_i = \mathbf{X}_i^T \boldsymbol{\theta}$, let $[U_j, V_j]$ be the bounded interval for the j th covariate, $1 \leq j \leq p - 1$. Then a complete class is formed by all designs with at most 2^p support points of the form $\{(\mathbf{C}_{\ell 1}, \omega_{\ell 1}) \text{ and } (\mathbf{C}_{\ell 2}, \omega_{\ell 2}), \ell = 1, \dots, 2^{p-1}\}$, where $\mathbf{C}_{\ell 1}^T = (1, a_{\ell 1}, \dots, a_{\ell, p-1}, c_\ell)$, $\mathbf{C}_{\ell 2}^T = (1, a_{\ell 1}, \dots, a_{\ell, p-1}, -c_\ell)$, and $c_\ell > 0$. Here, $a_{\ell j}$ is either U_j or V_j , and $(a_{\ell 1}, \dots, a_{\ell, p-1}), \ell = 1, \dots, 2^{p-1}$ cover all such combinations.

Proof. Let $\xi = \{(\mathbf{C}_i, \omega_i), i = 1, \dots, k\}$ be an arbitrary design with $x_{ij} \in [U_j, V_j]$, $i = 1, \dots, k, j = 1, \dots, p - 1$. By Equations (4.28) and (4.30), we have

$$\mathbf{I}_\xi \leq \mathbf{A}^T \left(\sum_{i=1}^k \sum_{\ell=1}^{2^{p-1}} \omega_i^\ell \mathbf{C}_i^\ell \Psi(c_i) (\mathbf{C}_i^\ell)^T \right) \mathbf{A}, \quad (4.31)$$

where \mathbf{C}_i^ℓ and ω_i^ℓ are as defined prior to the statement of the theorem. Notice that

$$\mathbf{C}_i^\ell \Psi(c_i) (\mathbf{C}_i^\ell)^T = \mathbf{B}_\ell^T \begin{pmatrix} \Psi(c_i) & c_i \Psi(c_i) \\ c_i \Psi(c_i) & c_i^2 \Psi(c_i) \end{pmatrix} \mathbf{B}_\ell, \quad (4.32)$$

where $\mathbf{B}_\ell = \begin{pmatrix} 1 & a_{\ell 1} & \cdots & a_{\ell, p-1} & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$. By Equations (4.31) and (4.32), we have

$$\begin{aligned} \mathbf{I}_\xi &\leq \mathbf{A}^T \left(\sum_{\ell=1}^{2^{p-1}} \mathbf{B}_\ell^T \left(\sum_{i=1}^k \omega_i^\ell \begin{pmatrix} \Psi(c_i) & c_i \Psi(c_i) \\ c_i \Psi(c_i) & c_i^2 \Psi(c_i) \end{pmatrix} \right) \mathbf{B}_\ell \right) \mathbf{A} \\ &\leq \mathbf{A}^T \left(\sum_{\ell=1}^{2^{p-1}} \sum_{i=1}^2 \omega_{\ell i} \tilde{\mathbf{C}}_{\ell i} \Psi(\tilde{c}_\ell) \tilde{\mathbf{C}}_{\ell i}^T \right) \mathbf{A} \\ &= \mathbf{I}_\xi \end{aligned} \quad (4.33)$$

for a design $\tilde{\xi}$ that belongs to the complete class in the theorem with support points $\tilde{\mathbf{C}}_{\ell 1}$ and $\tilde{\mathbf{C}}_{\ell 2}$ and weights $\omega_{\ell 1}$ and $\omega_{\ell 2}$, $\ell = 1, \dots, 2^{p-1}$. The existence of these support points so that the second inequality in Equation (4.33) holds is a consequence of Theorem 4.5.

The complete class in Theorem 4.7 is simple, but finding an optimal design for a specific problem still requires the determination of 2^{p-1} values for the c_ℓ 's and of the weights for the up to 2^p support points. This can easily be done by computer for small p , but is still challenging for larger p . For some problems,

it is again possible to find more explicit solutions for optimal designs. Below is a simple example without displaying all the computational details. We refer to Yang, Zhang, and Huang (2011) for more detailed examples of this kind.

Example 4.7. Consider a logistic model of the form of Equation (4.2) with $\mathbf{X}_i^T \boldsymbol{\theta} = c_i = \beta_0 + \sum_{j=1}^3 \beta_j x_{ij}$. Suppose that we are interested in a locally D -optimal design for $(\beta_0, \beta_1, \beta_2, \beta_3) = (1, -0.5, 0.5, 1)$, and that the first two covariates are contained in the intervals $[-2, 2]$ and $[-1, 1]$, respectively. Suppose further that there is no restriction on the third covariate. Theorem 4.7 assures us that we can find an optimal design based on at most eight design points for which the first two coordinates take all possible combinations of the limits of the respective intervals twice. Further computation gives a D -optimal design ξ^* for $(\beta_0, \beta_1, \beta_2, \beta_3)$ with equal weight $1/8$ for each of the following eight points:

$$(-2, -1, -0.4564), \quad (-2, -1, -2.5436), \quad (-2, 1, -1.4564), \quad (-2, 1, -3.5436), \\ (2, -1, 1.5436), \quad (2, -1, -0.5436), \quad (2, 1, 0.5436), \quad (2, 1, -1.5436).$$

We can now compare this optimal design to any other design, for example, to a full factorial design with three levels for each of the covariates, say $-2, 0$ and 2 for x_{i1} ; $-1, 0$ and 1 for x_{i2} ; and $-3, -1$ and 1 for x_{i3} , with equal weight for each of the 27 points. Simple computations show that the efficiency of ξ , defined as $(|I_\xi|/|I_{\xi^*}|)^{1/4}$, is only 70%.

We also point out that the support size for an optimal design need not be 2^p . This number is generally much larger than the number of parameters in $\boldsymbol{\theta}$, which is $p + 1$, so that it may be possible to find optimal designs with a (much) smaller support size. The result in Theorem 4.7 does not exclude this possibility. Indeed, the statement of the theorem refers to “at most 2^p support points.” This means that some of the weights can perhaps be taken as 0. It is our experience that this is often possible, but we do not have a general recipe for obtaining such smaller designs.

4.5.2 GLMs with Group Effects

In this section, we return to the problem of a single covariate, but now in the presence of group effects. If the subjects in the study have different characteristics with respect to one or more classificatory variables, such as race, gender, age group, and so on, and if the relationship between the response variable and the covariate can be different for subjects with different characteristics, then this should be reflected in the model. This is the main problem studied in Stufken and Yang (2012). In this section, we will present their main results, but refer for the proofs to the original paper. We will also restrict attention to a binary response variable and refer to Stufken and Yang (2012) for Poisson regression models.

To present the model, it is convenient to change the notation slightly from that in previous sections. We will use double subscripts to denote the subjects. For example, Y_{ij} is now used to denote the response from the j th subject in the i th group. We assume that there is a single covariate for this subject, x_{ij} , the value of which can be selected by the experimenter. In addition, the relationship between Y_{ij} and x_{ij} may depend on the group, i.e., it may depend on i . As before, we want to model $\text{Prob}(Y_{ij} = 1)$.

We consider two different models, one with a common slope and one with a group-dependent slope. The first of these models can be written as

$$\text{Prob}(Y_{ij} = 1) = P(\alpha_i + \beta x_{ij}), \quad (4.34)$$

where β is a common slope parameter, the α_i 's are group effects, and P is a cdf as in Equation (4.2). We could have used the notation of Section 4.2.1 for this model with $\mathbf{X}_{ij}^T = (0, \dots, 1, \dots, 0, x_{ij})$ and $\boldsymbol{\theta}^T = (\alpha_1, \dots, \alpha_k, \beta)$. Here, k denotes the number of groups, and \mathbf{X}_{ij} has a 1 in the i th position, with all other entries among the first k being 0. For the second model we can write

$$\text{Prob}(Y_{ij} = 1) = P(\alpha_i + \beta_i x_{ij}). \quad (4.35)$$

Now \mathbf{X}_{ij}^T is of length $2k$ with a 1 in the i th position and with x_{ij} in position $k + i$, i.e., $\mathbf{X}_{ij}^T = (0, \dots, 1, \dots, 0, 0, \dots, x_{ij}, \dots, 0)$, and $\boldsymbol{\theta}^T = (\alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_k)$.

While these models include group effects, they do not attempt to model these effects. If appropriate, simpler models could be used. For example, for the models in Equations (4.34) and (4.35), the α_i 's could be modeled as a sum of an overall mean and main effects for the different variables that induce the groups. This could also be done for the β_i 's in Equation (4.35). Such assumptions would reduce the number of parameters. For example, with two factors that have k_1 and k_2 levels, respectively, we have $k = k_1 \times k_2$. But if we can assume for modeling the α_i 's, say, that the two factors do not interact, then we can use main effects only and reduce the number of parameters from k to $k_1 + k_2 - 1$. We will not do this here, but merely note that the complete class results for the models in Equations (4.34) and (4.35) that we will formulate in Theorem 4.8 also hold for these reduced models.

We will formulate the complete class results in terms of the induced design space, using $c_{ij} = \alpha_i + \beta x_{ij}$ for the model in Equation (4.34) and $c_{ij} = \alpha_i + \beta_i x_{ij}$ for the model in Equation (4.35). The complete class result can be formulated succinctly, with some loss of precision, by stating that for each group, we should use a design of the form that we would have used if that group had been the only group (in which case we would have invoked Theorem 4.5). Here is the more precise statement.

Theorem 4.8. For the logistic or probit models of the form Equation (4.34) or (4.35), suppose that the design space is of the form $c_{ij} \in [D_{il}, D_{i2}]$. A complete class is formed by all designs with at most two support points per group, with

the additional restriction that if there are two support points in group i , then the five conditions in Theorem 4.5 apply with D_1 and D_2 replaced by D_{i1} and D_{i2} .

The proof of this result is fairly simple and can be found in Stufken and Yang (2012). The result vastly reduces the search for optimal designs. Nonetheless, there are still up to $2k$ support points and corresponding weights, so that an optimization algorithm may still have difficulties solving this problem if k is large. Depending on the model (and the use of possible further simplifying assumptions based on main-effects and lower order interactions), it may be possible to find optimal designs with a smaller support size than $2k$. We would still search in the complete class specified in Theorem 4.8, but some of the weights can be taken as zero.

Example 4.8. Consider a logistic model of the form as in Equation (4.34), with a common slope, for two factors that have three levels each. Assume that the two-factor interaction is negligible. Let $\alpha_1, \alpha_2, \dots, \alpha_9$ correspond to the groups $(1, 1), (1, 2), \dots, (3, 3)$, respectively. For finding a locally A -optimal design, suppose that $\boldsymbol{\theta}^T (-0.95, -1, -0.9, -0.85, -0.9, -0.8, -1.05, -1.1, -1, 1)$, where the first nine entries correspond to $\alpha_1, \alpha_2, \dots, \alpha_9$, and the last entry is for β . Note that these values are consistent with the assumption of no interaction between the two factors. We assume that there is no restriction on the design space. Suppose that the vector of interest, $\boldsymbol{\eta}$, consists of (1) the contrast of the average of the mean responses at levels 1 and 2 for the first factor versus the average at its third level (i.e., $\frac{1}{6}(\sum_{i=1}^6 \alpha_i - 2\sum_{i=7}^9 \alpha_i)$); (2) the contrast of the average of the mean responses at levels 1 and 2 for the second factor versus the average at its third level (i.e., $\frac{1}{6}(\sum_{i=1,2,4,5,7,8} \alpha_i - 2\sum_{i=3,6,9} \alpha_i)$); and (3) the slope parameter (i.e., β). An A -optimal design for $\boldsymbol{\eta}$ that uses fewer than 18 support points is shown in Table 4.1.

Table 4.1 Support Points and Weights for a Locally A-Optimal Design

Group	A -Optimal Design
$(1, 1)$	$(1.7691, 0.0550); (0.1309, 0.0700)$
$(1, 2)$	No support points (weights are 0)
$(1, 3)$	$(1.7191, 0.0783); (0.0809, 0.0466)$
$(2, 1)$	No support points (weights are 0)
$(2, 2)$	$(1.7191, 0.0482); (0.0809, 0.0769)$
$(2, 3)$	$(1.6191, 0.0852); (-0.0191, 0.0398)$
$(3, 1)$	$(1.8691, 0.0658); (0.2309, 0.0591)$
$(3, 2)$	$(1.9191, 0.0727); (0.2809, 0.0523)$
$(3, 3)$	$(1.8191, 0.0949); (0.1809, 0.1552)$

It is also possible to use Theorem 4.8 for some problems to find explicit optimal designs. Stufken and Yang (2012) prove such a result. To present it, let $\boldsymbol{\eta}_1 = (\alpha_1/\beta, \dots, \alpha_k/\beta, \beta)^T$ and $\boldsymbol{\eta}_2 = ((\alpha_1 - \alpha_k)/\beta, \dots, (\alpha_{k-1} - \alpha_k)/\beta, \beta)^T$.

Theorem 4.9. For the model in Equation (4.34), suppose that we have a single factor with k levels and no constraint on the design space. Let design $\xi^* = \{(c_{i1} = c^*, \omega_{i1} = \frac{1}{2k}), (c_{i2} = -c^*, \omega_{i2} = \frac{1}{2k}), i = 1, \dots, k\}$ for some c^* . Then, the following results hold:

1. ξ^* is D -optimal for $\boldsymbol{\eta}_1$ if c^* maximizes $c^2\Psi^{k+1}(c)$; and
2. ξ^* is D -optimal for $\boldsymbol{\eta}_2$ if c^* maximizes $c^2\Psi^k(c)$.

Computing the values of c^* for the logistic and probit models is easy to do with software like MATLAB and the expressions for $\Psi(x)$ in Equation (4.18). The results are shown in Table 4.2.

4.6 SUMMARY AND CONCLUDING COMMENTS

Selecting a good design is a complex problem that typically involves many different considerations. In the context of GLMs, this chapter focuses on one of these, namely design optimality with respect to some criterion based on information matrices. Rather than focusing on a single criterion, we focus on identifying complete classes of designs. For any optimality criterion that obeys the Loewner ordering of information matrices, we can always find a locally optimal design in these complete classes. Therefore, we can restrict searches for optimal designs to these classes. This is tremendously helpful if the complete classes are sufficiently small.

Table 4.2 The Value of c^* that Maximizes $c^2\Psi^q(c)$

q	c^*	
	Logistic	Probit
1	2.3994	1.5750
2	1.5434	1.1381
3	1.2229	0.9376
4	1.0436	0.8159
5	0.9254	0.7320
6	0.8399	0.6696
7	0.7744	0.6209
8	0.7222	0.5815
9	0.6793	0.5487
10	0.6432	0.5209

Identifying sufficiently small complete classes for GLMs is not a simple problem, but this chapter shows that the problem can be handled for a wide variety of models. By doing so, most results on optimal designs for GLMs obtained by other methods can be covered and extended. We have also shown how the determination of complete classes can at times be used to obtain explicit forms of optimal designs for specific criteria and objectives.

We reiterate that in practice, one may wind up not using a locally optimal design. Practical considerations, considerations of robustness of a design to uncertainty in the “local values” of the parameters, robustness to model uncertainty, and other considerations may play a role in selecting the design that is eventually used. Nonetheless, whatever the considerations are, in the end, one would hope to have a design that is efficient under reasonable criteria and assumptions. In order to assess the efficiency of a proposed design under optimality criteria, one has to compare it with an optimal design. That can only be done if one has the tools to identify an optimal design. Being able to identify optimal designs is thus important, irrespective of whether one plans to use an optimal design in an experiment or not.

While the analytic approach presented here, and in greater technical detail in some of the references given throughout this chapter, is very successful, there remain many open problems in the general area of identifying optimal designs for GLMs and other nonlinear models. The quote from Khuri, Mukherjee, Sinha et al. (2006) presented in Section 4.1 remains valid. There are still many dissertations to be written in this area.

ACKNOWLEDGMENTS

The research of John Stufken is supported by NSF grants DMS-0706917 and DMS-1007507. The research of Min Yang is supported by NSF grants DMS-0707013 and DMS-0748409.

REFERENCES

- Abdelbasit, K.M. and R.L. Plackett (1983). Experimental design for binary data. *J. Am. Stat. Assoc.*, **78**, 90–98.
- Agin, M. and K. Chaloner (1999). Optimal Bayesian design for a logistic regression model: Geometric and algebraic approaches. In: *Multivariate Analysis, Design of Experiments and Survey Sampling*, S. Ghosh (ed.). New York: Dekker, pp. 609–624.
- Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). New York: Wiley.
- Biedermann, S., H. Dette, and W. Zhu (2006). Geometric construction of optimal designs for dose-response models with two parameters. *J. Am. Stat. Assoc.*, **101**, 747–759.

- Chaloner, K. and K. Larntz (1989). Optimal Bayesian design applied to logistic regression experiments. *J. Stat. Plann. Inference*, **21**, 191–208.
- Dette, H. and L.M. Haines (1994). E-optimal designs for linear and nonlinear models with two parameters. *Biometrika*, **81**, 739–754.
- Dror, H.A. and D.M. Steinberg (2006). Robust experimental design for multivariate generalized linear models. *Technometrics*, **48**, 520–529.
- Dror, H.A. and D.M. Steinberg (2008). Sequential experimental designs for generalized linear models. *J. Am. Stat. Assoc.*, **103**, 288–298.
- Elfving, G. (1952). Optimum allocation in linear regression theory. *Ann. Math. Stat.*, **23**, 255–262.
- Ford, I., B. Torsney, and C.F.J. Wu (1992). The use of a canonical form in the construction of locally optimal designs for non-linear problems. *J. R. Stat. Soc. B*, **54**, 569–583.
- Haines, L.M., M.G. Kabera, P. Ndlovu, and T.E.O. O'Brien (2007). D-optimal designs for logistic regression in two variables. In: *MODA 8—Advances in Model-Oriented Design and Analysis*, J.L. López-Fidalgo, J.M. Rodríguez-Díaz, and B. Torsney (eds.). Heidelberg, Germany: Physica Verlag, pp. 91–98.
- Khuri, A.I., B. Mukherjee, B.K. Sinha, and M. Ghosh (2006). Design issues for generalized linear models: A review. *Stat. Sci.*, **21**, 376–399.
- Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *Ann. Stat.*, **2**, 849–879.
- Kiefer, J. and J. Wolfowitz (1960). The equivalence of two extremum problems. *Can. J. Math.*, **12**, 363–366.
- Mathew, T. and B.K. Sinha (2001). Optimal designs for binary data under logistic regression. *J. Stat. Plann. Inference*, **93**, 295–307.
- McCullagh, P. and J.A. Nelder (1989). *Generalized Linear Models* (2nd ed.). London: Chapman & Hall.
- McCulloch, C.E. and S.R. Searle (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley.
- Minkin, S. (1987). Optimal designs for binary data. *J. Am. Stat. Assoc.*, **82**, 1098–1103.
- Minkin, S. (1993). Experimental design for clonogenic assays in chemotherapy. *J. Am. Stat. Assoc.*, **88**, 410–420.
- Pukelsheim, F. (2006). *Optimal Design of Experiments*. Philadelphia: SIAM.
- Russell, K.G., D.C. Woods, S.M. Lewis, and J.A. Eccleston (2009). D-optimal designs for Poisson regression models. *Stat. Sin.*, **19**, 721–730.
- Silvey, S.D. (1980). *Optimal Design*. London: Chapman & Hall.
- Sitter, R.R. and B.E. Forbes (1997). Optimal two-stage designs for binary response experiments. *Stat. Sin.*, **7**, 941–955.
- Sitter, R.R. and B. Torsney (1995a). Optimal designs for binary response experiments with two design variables. *Stat. Sin.*, **5**, 405–419.
- Sitter, R.R. and B. Torsney (1995b). Optimal designs for generalised linear models. In: *MODA 4—Advances in Model-Oriented Design and Analysis*, C.P. Kitsos and W.G. Muller (eds.). Heidelberg, Germany: Physica-Verlag, pp. 87–102.
- Sitter, R.R. and C.F.J. Wu (1993a). Optimal designs for binary response experiments: Fieller, D, and A criteria. *Scand. J. Stat.*, **20**, 329–341.

- Sitter, R.R. and C.F.J. Wu (1993b). On the accuracy of Fieller intervals for binary response data. *J. Am. Stat. Assoc.*, **88**, 1021–1025.
- Stufken, J. and M. Yang (2012). On locally optimal designs for generalized linear models with group effects. *Stat. Sinica*, **22** (in press).
- Tighiouart, M. and A. Rogatko (2006). Dose escalation with overdose control. In: *Statistical Methods for Dose-Finding Experiments*, S. Chevret (ed.). Chichester: Wiley, pp. 173–188.
- Woods, D.C., S.M. Lewis, J.A. Eccleston, and K.G. Russell (2006). Designs for generalized linear models with several variables and model uncertainty. *Technometrics*, **48**, 284–292.
- Yang, M. (2008). A-optimal designs for generalized linear model with two parameters. *J. Stat. Plann. Inference*, **138**, 624–641.
- Yang, M. (2010). On the de la Garza phenomenon. *Ann. Stat.*, **38**, 2499–2524.
- Yang, M. and J. Stufken (2009). Support points of locally optimal designs for nonlinear models with two parameters. *Ann. Stat.*, **37**, 518–541.
- Yang, M., B. Zhang, and S. Huang (2011). Optimal designs for binary response experiments with multiple variables. *Stat. Sin.*, **21**, 1415–1430.

CHAPTER 5

Design and Analysis of Randomized Clinical Trials

Janet Wittes and Zi-Fan Yu

5.1 OVERVIEW

A randomized clinical trial in medicine differs from many other experiments because the experimental units are people who have consented to participate. This requirement that participants consent to the experiment is not unique to medicine; randomized clinical trials share this feature with other types of experiments, for example, experiments in psychology or sociology. Many of the statistical principles and scientific features of design applicable to randomized trials in medicine are also relevant to trials of veterinary medicine where the subjects have not given informed consent. Nonetheless, in thinking about how to design a randomized clinical trial of an intervention in medicine, one should always be aware of the logistical constraints imposed by the need for participants to agree to join the trial and the complexities attendant on their continued participation for the life of the trial.

The experiments we discuss in this chapter differ from other experiments in another way: many clinical trials testing medical interventions operate within a regulatory framework. If an experiment is testing a drug or device that the designers or sponsors of the trial hope will be sold in the marketplace, then the design of the experiment and the statistical analysis must be acceptable not only to scientists but to the regulatory agencies in the countries where the product will become available. Furthermore, the various regulatory agencies (e.g., the Food and Drug Administration in the United States and the European Medicines Agency, which operates in many European countries) may have different standards that lead to different designs.

Design and Analysis of Experiments: Special Designs and Applications, First Edition. Edited by Klaus Hinkelmann.

© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

Rather than beginning with a definition of the type of experiment this chapter discusses, we have started this chapter emphasizing informed consent and regulatory supervision because these two considerations should lurk behind nearly all aspects of the design of these experiments. The experiments in this chapter are randomized clinical trials (sometimes called randomized controlled trials, and often referred to simply as RCTs). We focus on trials of medical interventions in humans, but many of the features of such designs, as mentioned earlier, are applicable to other areas as well. Most of this chapter deals with what are sometimes called *confirmatory* or *phase 3* trials, that is, those experiments that aim to test one formal hypothesis concerning the effect of an intervention on one outcome of interest. (Some trials have more than one formal primary hypothesis or more than one primary outcome of interest. In such cases, as discussed below, the trial must adopt methods to control the type I error rate to avoid problems of multiplicity.) The goal of such a trial may be to change medical practice or to confirm the utility of current medical practice. One goal of trials conducted for regulatory purposes may be to gain approval for marketing the product or for changing the label of an already marketed product.

Investigators embark on a phase 3 trial only after preliminary trials have been completed. If the substance or device being tested is a new one, studies in animal models have shown evidence that the intervention is likely to be safe and effective in humans. Such trials are called *preclinical* studies. Some earlier trials in humans have shown that people can tolerate the product and that the experience thus far has indicated that the profile of adverse events is not unacceptable. Trials that aim to show tolerability and to give a preliminary indication of safety are called phase 1 trials. Phase 2 trials are those that have explored doses to establish a range likely to be beneficial and not too toxic. Other phase 1 and phase 2 trials have studied various aspects of the mechanism of the drug, for instance, the half-life of the drug, how food affects its metabolism, which organs in the body the drug enters, and its modes of excretion.

Even phase 3 trials come in a wide variety of flavors. Trials may require short-term (hours, days, or weeks) or long-term (months or years) participation. A trial of a drug for postoperative pain, for example, may study each patient for a few hours; a trial studying the effect of a pain medication on arthritis may require the participant to continue in the trial for several days; a trial of a pain medication for chronic pain may last weeks or months for each patient. Trials of antipsychotic medications often study each participant for one to 2 months even though the product is to be used for the lifetime of the patient. On the other hand, a trial to see whether an intervention to extend the lifetime of breast cancer patients or a trial investigating whether a lipid-lowering medication reduces the risk of cardiovascular outcomes may require each participant to agree to remain on the study for years. Phase 3 trials vary in other ways as well, for their designs depend on the questions being asked, and, as Figure 5.1 shows, the nature of the control group.

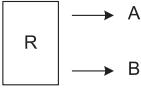
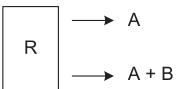
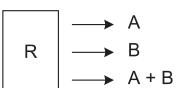
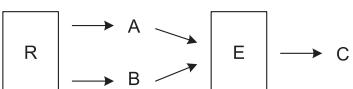
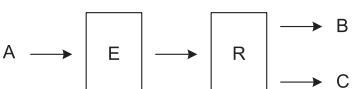
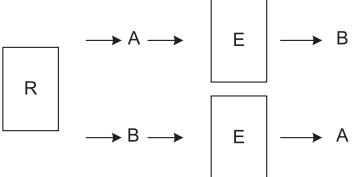
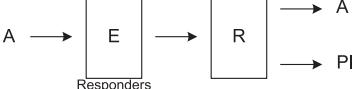
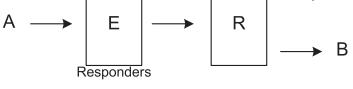
Design name	Design schematic	Comparison
Simple two treatment		Compare A vs. B
Adjuvant treatment		Does additional of B to A result in greater benefit?
Combination		Is A + B superior to each alone?
New treatment after event Event may be failure, response, or a fixed period of time		Compare treatment programs A or B followed by C
Common initial treatment		Compare treatment program A followed by B or C
Cross-over		Compare A vs. B both before and after event
Two randomizations		Compare A vs. B; Compare C vs. D after event
Randomized withdrawal		For responders to A, compare continued A or placebo
Enrichment		For responders to A, compare continued A vs. B

Figure 5.1 Types of phase 3 randomized clinical trial designs. Note: this figure is a variation and extension of a figure presented in Zelen (1993).

Although many phase 1 and phase 2 trials are conducted in a single clinical center, phase 3 trials often take place in many centers. Some take place in a single country or geographic region; others take place in a wide geographic area. The degree to which geographic variability affects such features as ethnic composition of the population and standards of medical care is relevant to how one makes inferences about the effect of treatment.

Many trials are *double-masked* (also called *blind*); that is, neither the investigator nor the participant knows what treatment the participant is receiving. Some trials are *single masked*—the investigator knows but the participant does not; and some—for example, most trials comparing surgery to medicine—are not masked at all. Many trials, especially those conducted for regulatory approval, are *triple masked*: neither the investigator nor the participant nor the statistician involved in analyzing the data knows the assignments of treatment until the trial is over and the database is cleaned and locked.

Trials differ from each other in the type of primary outcome they study—binary, ordinal, continuous, or time-to-event. The outcome, and its type, will determine sample size, follow-up time, and statistical methodology. This chapter touches on all these features and their consequences to design. For more detail, the reader should consult texts on methodology of clinical trials: for example, Pocock (1983); Friedman, Furberg, and DeMets (1998); or Piantadosi (2005).

The remainder of this chapter deals with important statistical aspects of the design of phase 3 trials. The various components of a randomized clinical trial are introduced in Section 5.2. Because of the importance of avoiding bias in randomized clinical trial, Section 5.3 is devoted to a discussion of bias and the various ways in which bias can arise. As pointed out in Section 5.4, which deals with general considerations related to the analysis of randomized trials, the methods of the remainder of this chapter are relevant to the trial when the primary outcome studied in a trial is a standard binary, ordinal, or continuous variable. On the other hand, when the outcome is timed to an event of interest, the relevant methodologies are techniques applicable to survival analysis, which is introduced in Section 5.5. Finally, Section 5.6 briefly mentions special topics in the design of randomized clinical trials.

5.2 COMPONENTS OF A RANDOMIZED CLINICAL TRIAL

Design of a randomized clinical trial requires specification of many elements, some of which are directly relevant to the statistical aspects of the design. First, one must specify clearly the question being asked. Understanding the purpose of the trial will drive the choice of population to be recruited, the variables to be measured, and the time course of the trial. This section describes the most important of these elements.

5.2.1 Target, or Reference, Population

Ultimately, the goal of the trial is to make inferences concerning a target population. As discussed in HK1, from a formal statistical view, inference from a randomized trial is limited to the sample studied because the study population is in no sense a random sample of the target population. Therefore, in designing the trial, the investigators need to define the population to which

the results apply. Suppose, for example, the trial is studying an intervention aimed at preventing myocardial infarction (heart attack). The target population may be *all persons at high risk for cardiovascular disease* where *high risk* is defined in some quantitative way. In past decades, specifically the 1960s through the early 1980s, many trials studying interventions for prevention of heart attack were limited to men and often excluded diabetics. From the point of view of the public health, however, the target population included women and diabetics. The difference between the study population and the target led to great uncertainty about the applicability of the results to the populations explicitly excluded from the trials. Similarly, in the current era, many trials of treatment of breast cancer exclude men, even though breast cancer—while rare in men—does occur in both men and women. Here again, the discordance between target and study populations leads to uncertainty about the applicability of results.

In the examples above, the exclusions clearly identify subsets of a population—men, women, diabetics—to which physicians might wish to generalize. A physician can make a judgment based not on direct data but on belief about the likely similarity of the excluded and included populations and therefore the likelihood that the treatment will work in those not studied. Sometimes, however, the exclusions are less clear. Consider age, for example. Many trials specify an age range. A trial studying a drug for malignant melanoma may specify that all people must be between the ages of 35 and 80 at the time of randomization. Suppose the trial shows that the experimental treatment is effective and the drug is approved. Then imagine that a woman of age 30 comes to her doctor with the disease. Should she be treated? Most people would say that 30 is close to 35 and that therefore, yes, the results can be generalized to those who are between 30 and 35. But what if the patient presenting is 18? Or 12? The farther one moves from the study population, the more uncertainty arises in generalizing the results. At the other end of the age range, the questions are somewhat different. Often a trial will specify an upper age limit or even say that anyone above a certain age is eligible. A careful reader of the results of a trial may note that actual age distribution includes very few people above the age of, say, 70. While the frail elderly were formally eligible to enter the trial, in fact very few of them actually were studied. Thus, the applicability of the trial to the target population is a function not only of the entry criteria but of the actual population entered.

Another example of the lack of connection between the target population and the study population is the frequent exclusion of women of childbearing potential. Many trials exclude such women because of the unknown, but potentially deleterious, effect of an experimental agent on a fetus. The consequence of such exclusion, however, is that many studies include very few premenopausal women.

The choice of clinical centers in which each trial is performed poses yet another problem in making inferences from a trial, or set of trials, to a target population . The clinical centers participating in the trials are often referral

centers, that is, clinics in academic institutions where the investigators are highly experienced clinicians and the patients are those who self-select to be treated at centers known for medical excellence. Such trials raise the question of generalization of the results to community settings. Conversely, some studies are performed in countries where few patients have access to expensive drugs. This lack of access may allow randomization of patients to a drug or placebo, both on top of a standard of care not considered optimal in wealthier countries. If under these conditions the trial shows the drug to be effective, the treating physician in countries with a more expensive standard of care will not know how to use the drug. The trial will be relevant neither to the comparison of the new drug to the already available expensive drugs nor to the comparison of the new drug in addition to previous drugs to the previous drug alone.

The inclusion and exclusion criteria may carve out a specific subset of the disease in question to study. For example, heart failure patients are categorized into one of four classes by the New York Heart Association (NYHA) classification. Many trials limit their study to class III and class IV patients, who are the most severely ill. If a treatment is effective for these patients, the question arises about its applicability to class II or even class I patients. Another example comes from studies of pain. If a study shows that a pain medication is effective in patients after knee surgery, can one generalize to other kinds of surgery? In cancer, if a drug is effective in reducing progression or mortality in prostate, breast, and colorectal cancer, are the results applicable to other solid tumors? Again, the ability to generalize will vary considerably depending on the biological closeness of the study population to a wider target population.

5.2.2 Study Population

Having defined the question the trial is asking, one selects the study population. The previous section already described some problems when the population actually studied differs from the target population. The study population should be broad enough to represent the population targeted to use the product should the trial show benefit, but not so broad that an effect, even if real, cannot be discerned. For example, if one wants to show that an intervention will lower the risk of stroke, one would choose a population at high risk for stroke because (1) that is the population who would use the drug and (2) a trial that recruits a population with a very low risk of stroke will yield very few strokes in the low-risk subpopulation, thus reducing the power relative to the power in the high-risk subpopulation. In clinical trials, the population is usually defined by a series of *inclusion* and *exclusion* criteria—people are eligible to enter a trial if they satisfy all of the inclusion and none of the exclusion criteria.

As described above, the typical study population in a clinical trial may differ in important ways from the target population. In selecting a study population, investigators need to randomize people who are likely to complete the trial

without violating the protocol even if that means the study and target populations do not match. A trial that recruits participants who fail to follow the protocol, who do not complete the trial, whose competing risks lead them to stop study medication, or even to die before the trial is complete, will remove from the study population groups of people who are in fact members of the target population. Inclusion of such patients in the trial, however, will render the results of the trial difficult to interpret. Thus, in thinking about inference from clinical trials, one must make a clear distinction between internal validity and external validity. Internal validity here means that the results of the trial hold for the population under study (or, more narrowly, if one is thinking in terms of permutation tests, for the actual participants in the study). A trial with internal validity must have been conducted without bias and without those confounding factors that would lead to ambiguous inference.

The very criteria that maximize the chance of strict internal validity may compromise the ability to generalize to the target population, that is, they may compromise the external validity of the trial. In general, the smaller the trial and the more exquisite the protocol, the larger will be the difference between the internal and external validity. Large simple trials, as described in Section 5.6.3, will recruit many people and will define a very simple primary outcome variable that can be assessed in nearly all participants. In that case, the study population will be very close to the target population. Trials with complex outcomes requiring frequent measurement will necessarily recruit participants who are likely to complete the trial; by their very nature, such people differ from the target population. Perhaps the starker example comes from trials of treatment for psychoses, for psychotic patients who are likely to complete a protocol without deviating from it do not represent the population of psychotic patients at large. A less extreme example comes from studies of pain: if one is comparing a new pain medication to placebo and the protocol precludes the use of rescue medication, the population chosen to study must be a group of people who are unlikely to experience unbearable pain over the course of the study (or a population of martyrs!).

5.2.3 Outcomes

A typical randomized clinical trial addresses a number of questions specified in a series of null and alternative hypotheses. In general, each hypothesis is associated with an *outcome* variable, sometimes called an *end point*. Examples of such variables are time to death in a trial studying a treatment for heart attack, level of pain for a trial studying an analgesic, number of lesions for a trial investigating a drug for leprosy (Hansen's disease), or cure for a trial of a new antibiotic. Most trials have a single *primary outcome*, the measure that drives the sample size, as well as a set of other outcomes, which are called *secondary, supportive, tertiary*, and, sometimes, simply *other*.

5.2.3.1 Primary Outcome

The primary outcome, which forms the basis for the calculation of the sample size of the trial, is the variable that the intervention is designed to affect. In a trial of pain, the primary outcome may be a score on a pain scale. In a trial of an intervention for type II diabetes, the primary outcome may be a laboratory measure like glycosylated hemoglobin (HgA1c) or a clinical measure like development of diabetic retinopathy. The definition of this primary outcome, the frequency of measuring it, and the precision of measurement all affect the design of the trial.

The primary outcome may be a continuous or ordinal variable; it may be a binary outcome; or it may be a time-to-event. Again, consider choice of primary outcome in a study of pain. A continuous outcome might be the level of pain on a visual analog scale, or VAS. The typical VAS is a 10-cm line anchored at “no pain” on the left and “worst imaginable pain” on the right. Participants in the trial mark their perceived level of pain. An ordinal outcome might be a point on a 7-point scale of pain where each number corresponds to a verbal description of intensity of pain. The VAS or the seven-point scale could be used to create a binary variable where “success” is a level of pain below, say, 3 cm on the VAS or below 2 on the ordinal scale. A time-to-event outcome might be the number of minutes until the pain falls below 3 cm on the VAS. Thus, in this case, the same instrument can give rise to outcome measures that have very different statistical properties (see also HK1, section 2.25).

5.2.3.2 Supportive, Secondary, and Exploratory Outcomes

Clinical trials rarely measure only a single outcome. The typical trial measures many variables both at baseline and at various follow-up times. Trials that are performed not for regulatory review often fail to distinguish among these outcomes; rather, the results are reported often with no correction for multiplicity. In many trials, all the outcomes except the primary outcome are called “secondary”.

Trials performed for the purpose of regulatory review, on the other hand, often classify each nonprimary outcome as “supportive,” “secondary,” or “exploratory.” A “supportive” outcome is a variation on the theme defined by the primary. It may be the same outcome analyzed in a somewhat different way. An example is a study of an eye disease where the outcome is best corrected visual acuity. Suppose the primary analysis compares the mean acuity in the treated and control groups with baseline acuity as the only covariate. A supportive analysis might use the same outcome but now with other covariates as well. In a trial comparing interventions for breast cancer, the primary outcome may be progression-free survival as assessed by a masked outcome committee; a supportive outcome might be progression-free survival as assessed by the investigator. Generally, the choice of supportive analyses does not influence the design of the study; their purpose is to assess the robustness of the conclusions regarding the primary outcome.

Secondary outcomes, on the other hand, are measures that add to the understanding of the breadth of the intervention's effect. In studies of pain where the level of pain is the primary outcome, a secondary outcome might be the ability of participants to carry out activities of daily living. In studies of eye disease with best corrected visual acuity as the primary outcome, a secondary outcome might be the ability of a participant to drive, or it may be a measure of peripheral vision. In a trial of breast cancer where overall survival is the primary outcome, secondary outcomes might be the complete response of the tumor, progression-free survival, occurrence of bone metastases, or a measure of quality of life. In a trial studying replacement enzymes for patients with lung disease caused by an enzyme deficiency, the primary outcome might be a measure of lung function. A secondary outcome might be the level of enzyme in the blood. In the regulatory setting, if the primary outcome shows statistically significant evidence of benefit, the label for the product may make claims based on a secondary outcome's showing benefit. Note that if the primary outcome is not statistically significant, nominal significance observed for a secondary outcome rarely leads to a claim because the failure of the primary outcome to show benefit means that all the type I error rate is "used up," and the secondary outcomes cannot be declared statistically significant. Therefore, statisticians designing a clinical trial are well advised to calculate the power for the secondary outcomes as well as for the primary outcome. The trial should include a formal strategy for protecting the type I error rate. Furthermore, if the secondary outcomes are to be used in a formal way for claims in a label, the methods of data collection must be as careful for them as for the primary outcome.

Many trials that are not conducted with the purpose of gaining regulatory approval for a product do not adopt a formal statistical framework for secondary outcomes. Nonetheless, as described in HK1, statisticians designing such trials are well advised to incorporate careful adjustments for multiplicity for the secondary outcomes. In so doing, the scientific community can know how to judge the reliability of the findings.

Exploratory outcomes are those outcomes that the investigators regard as hypothesis generating. The statistical power for these is often low; they are frequently, but not always, less important clinically than the primary and secondary outcomes, and the method of data collection and validation for them is often less rigorous than for the primary and secondary outcomes.

Snappin and Jiang (2011) suggest that instead of designating outcomes as primary, secondary, and other, a more useful categorization would classify the set of outcomes into two groups: those with and without type I error rate control.

5.2.4 Projected Timeline

In trials where all participants are followed for the same fixed period of time, the concept of time is simple. A person enters the trial, is followed for a fixed

time period, and the outcome is measured. Time plays a role in this type of study only in relation to recruitment. The length of the trial is defined by how quickly the study population is recruited. Analytically, each person is considered to start at time 0. Especially in a trial with a long recruitment time, prudent investigators check for important drifts in event rate over time, but theoretically, the study—and hence the analysis—is time invariant.

Much more complicated are studies that follow people for different lengths of time. Later in this chapter we discuss such trials in more detail along with the statistical methodology appropriate to them. Here we simply point out that if recruitment occurs over several years and each person is followed until a fixed time after the last person has been randomized, time enters the conception of the trial and its analysis in several ways. Calendar time starts on the day the first participant enters the trial and ends when the last person exits the trial. Study time is structured differently. Each person is assumed to enter at time 0; study time lasts from that artificial time 0 to the time of follow-up for the last person entered. See Section 5.5 for a fuller discussion of calendar and study time.

5.2.5 Choice of Control Group

The choice of control group is a crucial aspect of the design of a randomized trial. The typical trial compares an experimental intervention to some control. The control can be a placebo; it can be “best supportive care”; it can be “usual care”; or it can be another agent already shown to be effective. The fact that the participants in the trial have given informed consent to joining the trial means that they must understand the nature of the control group. The clearest inference about the effect of the experimental agent or intervention occurs when the control group is placebo (or, in a trial of surgery, when the control group is a sham operation). In conditions for which no effective therapy is available, such trials are feasible. Potential participants can be told that the condition has no therapy that works; therefore, entering a trial where the probability of receiving placebo is, say, one-half does not put them at a disadvantage relative to what would have happened in the usual setting. Similarly, consider a short-term trial studying a reversible symptom in a setting where an effective treatment is available. The potential participant might be willing to be randomized to placebo for a short time, especially if the trial incorporates a method for receiving rescue therapy if the symptom becomes too severe.

When the trial is studying a serious disease for which an effective therapy is available, randomizing to placebo can put the participant at undue risk. In that case, the control therapy might be an established treatment already shown to be effective. In a disease where the consequence of not being treated is dire, randomizing to an experimental therapy or a known effective therapy may put those in the experimental group at undue risk. In such cases, all participants may receive the known effective therapy, and randomization assigns placebo or the experimental therapy in addition to the established treatment.

At the end of a trial, the conclusions should be framed in terms of the effect of the experimental therapy relative to the control.

5.3 BIAS

Bias, an important concept in statistics in general, in clinical trials can be defined as any process, willful or inadvertent, that leads to a systematic error in the measurement of the effect of treatment. In statistics, the bias of an estimator $\hat{\theta}$ of θ is defined as the difference between its expected value and its true value, that is,

$$E(\hat{\theta}) - \theta. \quad (5.1)$$

A statistical test is biased if the true type I error rate is greater than the type I error rate stated for the test; that is, a test is biased if the probability that it will declare an effect of treatment statistically significant when none exists is greater than the probability stated for the test. While the multiple use of the word *bias* can be confusing, this chapter uses the term loosely to refer to any tendency that systematically leads to a result different from the true answer to the question being posed. We concentrate on clinical biases that are difficult to control and to quantify, as opposed to statistical bias (such as the bias that arises from confounding variables). The trial's design can adjust for the former while the analysis can adjust for the latter.

Both bias and variability contribute to the total error of measurement. Statisticians often talk about minimizing bias, minimizing variability, and striking a balance between the two. Recall that precision is the inverse of the variance. Consider a distribution with true mean θ . If the estimated treatment effect is biased, then ideally its bias is small relative to its variability; the bias is small in relation to the total error as illustrated in Figure 5.2. The first panel displays an unbiased distribution with low precision. The second distribution, which has small bias but high precision, would yield a satisfactory estimate of the mean θ , which would perhaps be preferable to the unbiased distribution. The second panel illustrates two biased curves—one with little bias and low precision, the other with large bias and high precision. In this case, an estimate from the right-hand curve with small bias and less precision is preferable over an estimate from the left-hand curve with high precision but large bias. This panel also illustrates that bias poses a particularly serious problem if the treatment effect is small. Because the presence of bias can greatly affect the interpretation and generalizability of a clinical trial, many aspects in trial design exist to ensure its minimization. The following sections outline these types of protection.

5.3.1 Unbiased Entry Criteria and Recruitment

While randomization provides balance, on average, in the distribution of measured and unmeasured patient characteristics among treatment groups in a

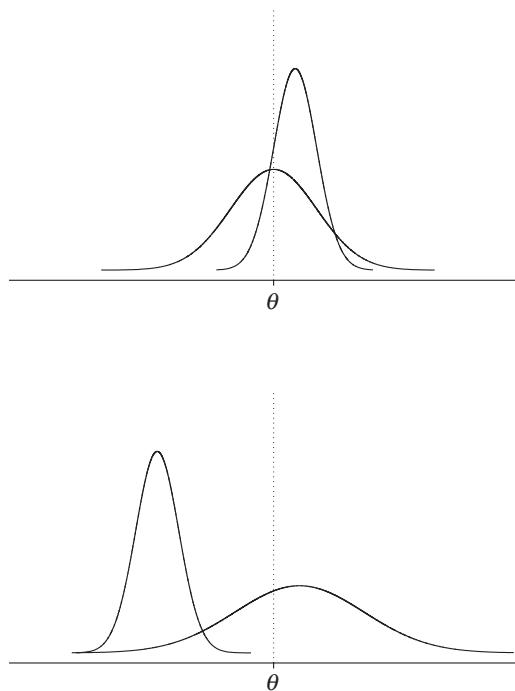


Figure 5.2 Bias and variability.

trial, preventing bias begins even before patients are randomized. The selection criteria used to screen potential participants into a trial should ensure that comparable criteria are used to enroll participants in different treatment groups. If the criteria differ, as could occur in a nonrandomized study, then this selection bias could lead to a biased estimate of the treatment effect. For example, in the nonrandomized setting, differential selection criteria that assigned healthier patients to the control arm would dilute the treatment effect. Alternatively, if all sicker patients were in the control arm, then the treatment effect would appear larger than it really is because all the healthy patients would experience even better results on the treatment arm.

Selection criteria should also ensure that participants will complete the study, and that their results are generalizable to the target population. Trials of a longer term therapy in breast cancer, for example, may be designed to include those with good performance status to better ensure that (1) those who are healthier can better tolerate a potentially harsh therapy and (2) outcomes such as survival reflect the effects of therapy rather than a patient's already poor prognosis. This screening occurs prior to treatment assignment to avoid selection bias in a nonrandomized setting, where physicians may select participants with certain characteristics, such as different risk profiles, for an experimental (rather than control) group.

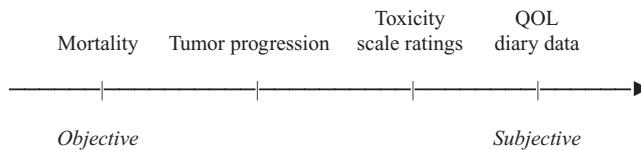


Figure 5.3 Potential trial outcomes and objectivity spectrum in a cancer trial.

Sometimes, investigators enroll into a trial patients who do not satisfy all the selection criteria. These so-called “exemptions” may occur for a variety of reasons, including the belief that the exception to the criteria is minor, or pressure to meet enrollment goals. The magnitude of effect on potential results of a trial depends on the extent of exemptions by investigators during study recruitment and on the type of exemption itself. Bias in the estimated treatment effect could occur if, under the exemptions, the study enrolled participants who are unlike the target population, such as the example mentioned early in Section 5.2.1 of a target population of NYHA class III and class IV participants. Allowing exemptions of healthier class I and class II patients into the study could dilute the treatment effect. In a smaller study, the magnitude of the bias could be large if most of these healthy patients were randomized to the same treatment group by chance.

5.3.2 Outcome Measures—Unbiased Assessment

Bias can also enter the assessment of outcomes. The objectivity of both investigators and participants may be compromised, particularly in open-label or unmasked studies, by personal beliefs about an exciting experimental treatment, financial interests, or time already invested in a product. Figure 5.3 shows a range of outcomes on a spectrum of objectivity.

Masking helps minimize bias in the measurement of trial outcomes. Even if the investigator is unmasked, endpoints may be adjudicated in a masked fashion, such as reviews of scans evaluating organ measurements or disease progression. The endpoint should be evaluated in such a way that the treatment assignment cannot affect its measurement. If the endpoint is fully objective (e.g., mortality), then an unmasked study may be unbiased. The more subjective the endpoint, the greater the rigor of masking necessary to prevent bias in the evaluation of outcome.

5.3.3 Once Randomized, Always Analyzed (Intent-to-Treat)

The analysis of data from randomized clinical trials must include all persons randomized, even those who do not receive their assigned treatment; the mantra of such clinical trials is *once randomized, always analyzed*. Analysis by randomized assignment, or *intent-to-treat* (ITT), is essential to ensure the

comparability of study groups. Randomization constructs study groups of equivalent risk; removal of people from the randomization set can bias the comparison of treatment to control. Many investigators find this rule arbitrary and counterintuitive, preferring an analysis by treatment received which, they say, produces more accurate estimates of the true biological effect of the treatment. While there is considerable debate in the medical community about use of ITT, most statistical arguments support the ITT approach. In studies of time to failure in chronic disease, an individual participant's treatment often ends earlier than planned in the protocol. One common approach to analyzing such data is to censor observations made after cessation of treatment. Lagakos, Lim, and Robins (1990), in a discussion of inferences from ITT when treatment terminates early in time-to-failure trials, point out that censoring observations in this way produces biased estimates of the distribution of time to failure had the therapy not been terminated. They show that an ITT approach yields valid inferences concerning the unconditional distribution of time to failure, regardless of whether and when the treatment was terminated. Lee, Ellenberg, Hirtz et al. (1991) show that using compliance to determine groups for comparison in analysis by treatment received may cause bias. Peduzzi, Witzes, Detre et al. (1993) point to the pitfalls of methods of analysis other than by random assignment even in trials comparing surgery to medical therapy.

Another interpretation of the ITT is that trials really test treatment policy; see, for example, Piantadosi (2005). A failure to complete assigned therapy may be viewed as an outcome of the trial, and may result from many factors other than treatment alone. In this light, a clinical trial tests the treatment program as opposed to simply the treatment alone, and testing by groups other than the assigned ones could produce biased results.

5.3.4 Masking Participants, Investigators, and Others

Gould (1981), in his book on measuring human intelligence, points to many ways in which one's own personal bias can lead to biased measurements, even when the measuring device appears to be objective. This tendency of the observer to measure somewhat subjectively has led the field of clinical trials to incorporate some masking of procedures. As discussed above, in the simple case where the experimental agent is a pill and the control agent a placebo, the treating physicians and participants—indeed all people involved in the operations of the trial except those who package the drug—can be masked to treatment allocation. Such masking is more difficult and sometimes impossible in trials where the two regimens differ qualitatively. Examples are trials comparing two therapeutic agents where the timing of cycles of therapy differ, comparing two different kinds of surgery or surgery to medical therapy, or testing a treatment associated with known, frequent adverse events.

Even when masking of treatment is not possible, two aspects of the design, treatment allocation and measurement of outcome, must incorporate masking. First and most important is the masking of treatment allocation. The investiga-

tor recruiting the potential participant must not be aware of the next entry in the randomization list because knowledge of the next assignment can influence whether the investigator will recruit the potential participant. For this reason, if randomization uses a blocked design, the investigators must not know the size of the blocks; if they do, they can guess what the next assignment will be with increasingly high probability as recruitment reaches the end of the block.

Depending on the subjectivity of the outcome measure (recall Figure 5.3), the person evaluating the outcome should be masked to the participant's treatment assignment. If the outcome is death, masking is not essential to unbiased measurement. If, on the other extreme, the outcome is an assessment of quality of life or severity of symptom, where either the investigator or the participant makes the assessment, then knowledge of the treatment assignment can influence the outcome. Therefore, to the extent feasible, when a study is not double masked, a person or committee that is masked to treatment assignment should assess the outcome.

5.3.5 Noncompliance and Study Dropout

Both noncompliance and dropout from the study have major roles in the design and interpretation of a clinical trial. People who participate in trials, like patients in the ordinary practice of medicine, do not always adhere to their assigned therapeutic regimen. Some simply forget to take their medication, some are too sick to accept therapy, and some stop study therapy because they are feeling better. Rigorous statistical analysis of the data should include those who do not comply with the assigned therapy. Even in the presence of noncompliance, analysis should follow an intent-to-treat philosophy and include all people in the group to which they were randomized.

5.3.5.1 Noncompliance

Noncompliance potentially dilutes the treatment difference between treated and control groups. The consequences are lower power for a study and introduction of bias.

Certain types of noncompliance will not affect the power of a study. For example, a placebo patient who does not take the assigned placebo is, like the placebo complier, still acting as a nontreatment control (unless, of course, the placebo truly has an effect). Similarly, a person on an assigned treatment regimen who stops the assigned treatment but adopts a regimen with similar effects does not adversely affect the power of the study. The problems arise from those situations if the nature and extent of the noncompliance compromises the expected difference between treated and control groups. When patients cross over to the other treatment, the trial's sensitivity to detect a treatment difference is reduced. This happens when those on treatment drop therapy or adopt the control therapy, or when a control patient adopts the treatment therapy or one similar to the one being tested. An example described

by Zelen (1993) considers a study of therapy for head and neck cancer comparing two regimens: radiation followed by surgery and surgery followed by radiation. In the first group, those who responded well to radiation may refuse surgery; noncompliance occurs in patients with better prognosis. In the second group, those who have difficulty recovering from surgery may refuse radiation; here, noncompliance occurs in those with worse prognosis.

In calculating sample size, many investigators ignore noncompliance and perform computations as if everyone will adhere to their assigned therapies. Ignoring the problem, however, invites more serious difficulties later, for if a trial suffers considerable noncompliance, the sample size will be insufficient to ensure the desired power.

To illustrate the effect of noncompliance on a two-arm trial, let μ_e and μ_c denote the means for the experimental and control compliers, respectively. Similarly, let p_e equal the proportion of participants in the experimental group who “drop out” of active therapy, and p_c equal the proportion of participants in the control group who “drop in” to active therapy. Then the actual means, μ_e^* and μ_c^* , are weighted averages of compliers and noncompliers:

$$\mu_e^* = \mu_e(1 - p_e) + \mu_c(p_e) \quad (5.2)$$

$$\mu_c^* = \mu_c(1 - p_c) + \mu_e(p_c) \quad (5.3)$$

The observed treatment difference will thus be

$$\mu_c^* - \mu_e^* = (1 - p_c - p_e)(\mu_e - \mu_c),$$

where $p_c + p_e < 1$; otherwise $\mu_c^* - \mu_e^*$ would be negative, with a result that the control fared better than the treatment. Noncompliance reduces the statistical power because it attenuates the treatment difference by a factor of $(1 - p_c - p_e)$.

A simpler case is a trial with 100% compliance in the control group, or negligible drop-in to treatment in the control group. This could occur in a trial where no alternative therapy is available. In this case, $p_c = 1$ so that $\mu_c^* - \mu_e^* = (1 - p_e)(\mu_e - \mu_c)$.

To deal with noncompliance, some researchers increase the sample size by a factor that represents the proportion of people who are not expected to comply. A typical approach is to multiply the sample size by an adjustment factor of $1/(1 - c)$, where c is the proportion not complying. This method is insufficient for two reasons. First, this adjustment only assumes one type of noncompliance, typically those who stop taking active therapy. A better adjustment should consider both drop-ins and dropouts—that is, by using $1/(1 - p_c - p_e)$ as discussed above. Second, the multiplying factors for adjusting sample size should be on the same scale as the statistical efficiency. Thus, a more appropriate sample size adjustment is to multiply the sample size by $1/(1 - p_c - p_e)^2$.

Table 5.1 shows the sample size adjustments needed for varying proportions of noncompliance. To read the table, specify the percentages of people expected

Table 5.1 Sample Size Required Relative to a Trial with Full Compliance as a Function of the Proportion *Dropping In* to Active Therapy in the Control Group and the Proportion *Dropping Out* of Active Therapy in the Treated Group

Percentage of Treated Participants <i>Dropping Out</i> of Treatment	Percentage of Controls <i>Dropping In</i> to Active Therapy			
	0	5	10	15
0	1	1.11	1.23	1.38
10	1.23	1.38	1.56	1.78
20	1.56	1.78	2.04	2.37
30	2.04	2.37	2.78	3.31
50	4.00	4.94	6.25	8.16

to drop in and drop out of active therapy. Thus, if you expect 10% of the treatment group to drop out of treatment and 5% of the control group to drop in to treatment, then the sample size necessary to achieve the prespecified type I error rate and power would be 1.38 times the size needed if all participants complied with their assigned treatment.

Designers of the trial should anticipate the extent of potential noncompliance and adjust the sample size accordingly. Operational aspects of the trial should build in measures to minimize noncompliance, for example, by increasing the palatability of an oral treatment.

5.3.5.2 Study Dropout, or Loss-to-Follow-Up

Loss to follow-up is statistically related to noncompliance. Noncompliers may not adhere to the assigned treatment and therefore “behave” differently than expected, thereby affecting their outcomes. Patients who drop out of treatment (noncompliers) but who remain in the study may experience outcomes that differ from expected. Patients who are lost to follow-up, or drop out of the study, may not have assessments of outcomes at all.

While trials should aim for complete follow-up and assessment of treatment effects, loss to follow-up is a reality that investigators should account for in the protocol, statistical analysis, and sample size calculations. Efforts written into the protocol and practiced during the trial often include phone calls or final interview visits to obtain final assessment for those who withdrew from the study. Ignoring dropouts because they do not have outcome data and excluding them from the analysis leads to potential bias in the inference about treatment effect. An example of nondifferential dropout is a therapy for a chemotherapy treatment compared with a nontoxic control. If most of the dropout occurs in the chemotherapy group because the treatment is harsh and the patients’ bodies cannot tolerate the treatment, then compared with control, the treatment may seem more beneficial than it actually is.

In addition to examining the distributional properties of those with missing outcomes, many researchers adopt a conservative or a worst-case scenario approach. The idea is that if imputing data with these assumptions produces a treatment effect, then the true treatment effect is at least as strong as the observed effect. For example, if the outcome is binary, then a missing outcome could be treated as a failure; if the outcome is continuous with multiple measurements over time for an individual, then the measurement at the last time point could be imputed from the overall placebo trend; if the outcome is survival, then the outcome could be treated as censored. Another approach is to use a *worst reasonable* scenario (see e.g., Proschan, Lan, and Wittes 2006). For example, rather than treating all missing values as failures, impute for missing values the most extreme percentage consistent with data from the literature.

While the method for accounting for missing endpoint data is highly dependent on the study itself and defined in a statistical analysis plan, the implications for design are reflected in the adjustment of sample size. Often a conservative approach is favored and ties back into the discussion of noncompliance. In this case, one may assume the scenario of patients in the treatment arm dropping out of treatment and assuming a control response. The sample size would be adjusted accordingly. In fact, when researchers discuss adjusting their sample size for drop-out, they are really referring to patients who drop out of assigned treatment. If referring to those who drop out of a study, whether they know it or not, investigators are really assuming that the endpoint for those drop outs will be imputed.

5.4 STATISTICAL ANALYSIS OF RANDOMIZED CLINICAL TRIALS

Statistical analysis of randomized clinical trials follows the same principles of statistical analysis as appropriate to any experiment. In a classical statistical framework, one specifies null and alternative hypotheses along with the desired type I and type II error rates, the effect size, the variability, and the planned statistical analysis. These attributes allow calculation of the required sample size, as well as decisions about such issues as stratification or the number of centers needed, and they may influence the choice of study population. For binary, categorical, ordinal, and continuous data, the methods of calculation described in HK1 and in standard statistical tests apply directly. Many clinical trials with binary outcomes, however, use methods based on time-to-event. Such trials ask not whether the participant experienced the event, but rather whether the intervention modified the timing of the event. At least two reasons lead to the use of this type of analysis. First, for an event that occurs eventually to everyone (e.g., death), no intervention will serve as DeSoto's fountain of youth; rather, all one can expect from the intervention is to lengthen the time to event, not to prevent it entirely. The second reason for use of survival methods stems from the practicalities of most clinical trials.

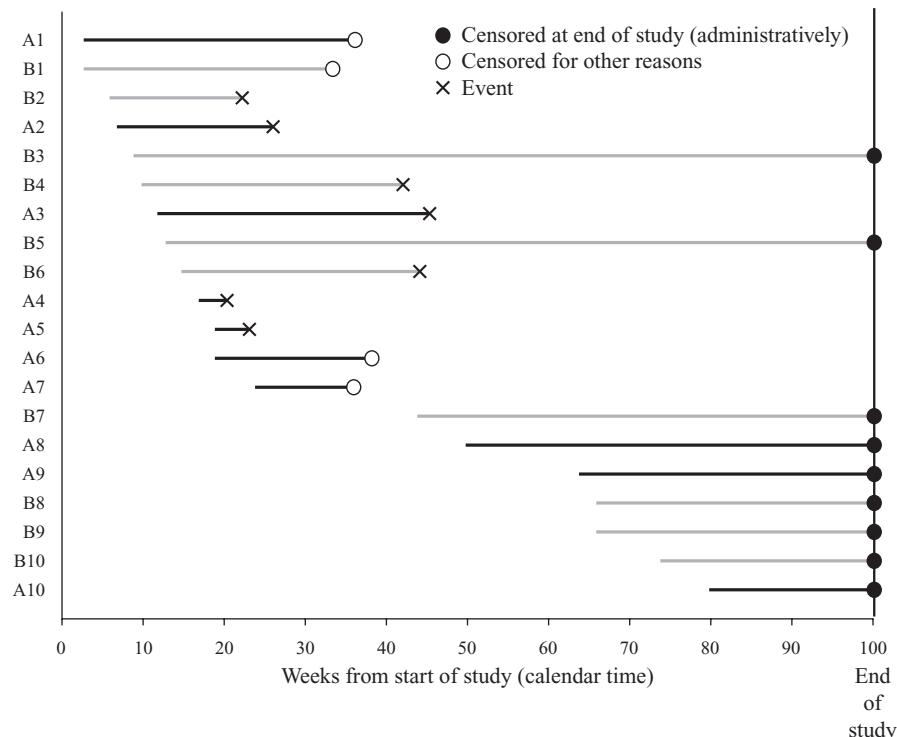


Figure 5.4 Experience of a clinical trial cohort in calendar time.

Even if one expects the treatment to cure the disease completely, or prevent the disease from occurring, the participants in the trial rarely have the same period of follow-up; therefore, binomial distributions are not strictly applicable.

Typically, recruitment to a trial occurs over a period of weeks, months, or sometimes years. The experience of a typical, but artificially small study population, is depicted in Figure 5.4. In this example, the total sample size is 20: 10 each in group A and group B. The study lasts 100 weeks from the time the first person is randomized until the time the trial ends. Because this trial, like many real trials, ends at a prespecified time after the last participant is randomized, follow-up time varies from person to person, violating the assumption of equal risk essential to binomial distributions. Each line in the figure represents a single participant and starts at the calendar time when the participant entered the trial. An \times at the end of a line means that the participant experienced the primary outcome of the trial.

When the trial is over, those still participating are *administratively censored*. A closed circle shows that at the end of the trial nine participants, three in Group A and six in Group B, were still being followed. This censoring is

independent of the process leading to the event of interest; it adds no bias to the system.

Some people will not experience a primary outcome event or reach the end of the study. These people, denoted by an open circle, are censored for a variety of other reasons. They may be lost to follow-up or they may withdraw their consent to participate. If the primary outcome is some event other than all-cause mortality, the primary outcome may be censored by a competing risk. For example, suppose one is studying the effect of a new lipid-altering drug on the occurrence of fatal or nonfatal heart attack. A person in the trial who dies of something other than a heart attack, say an automobile accident, or a stroke, or a cancer, is censored by a competing event. Similarly, in trials studying interventions to cure cancer, if the outcome of the trial is progression of the tumor and the participant dies before progression (or before progression is measured), a competing risk has prevented the assessment of the outcome. For both these reasons—administrative censoring caused by the fact that recruitment occurs over time, and more complicated types of censoring caused by loss to follow-up or competing risks—many long-term clinical trials studying clinical outcomes (as opposed to symptoms) use methods of survival techniques as the primary approach to analyzing data. Furthermore, many trials incorporate formal sequential analysis to allow early stopping if the data are showing compelling evidence of benefit, if they suggest unacceptable harm, or if the data thus far indicate that the chance of the trial's ever showing benefit is small. In the lingo of clinical trials, these three reasons for stopping are called stopping for efficacy, stopping for harm (often euphemistically called “stopping for safety”), or stopping for futility. The next sections introduce survival analysis and methods of sequential analysis commonly used in clinical trials.

5.5 FAILURE TIME STUDIES

5.5.1 Basic Theory

Fundamental to the design of studies that compare groups with respect to survival time, or, equivalently, time to failure, is the concept of a survival function and its related hazard function. (For simplicity of language, this section considers the outcome to be death unless otherwise specified. The theory holds for any event where the outcome is time to event.) We consider a variable $T \geq 0$, typically time, that has a density function $f(t)$ and a distribution function $F(t)$. The survival function is defined as $S(t)$:

$$S(t) = 1 - F(t) = \Pr\{T > t\}, \quad (5.4)$$

while the hazard function (or the hazard rate at time t , or, to use the language of demographers, the *force of mortality*) is:

$$\lambda(t) = \frac{f(t)}{1 - F(t)}. \quad (5.5)$$

The survival function is thus the probability that an individual survives (or does not experience the outcome of interest) longer than t . The hazard rate at time t ,

$$\lambda(t)dt \equiv \Pr\{t < T < t + dt | T > t\}, \quad (5.6)$$

is approximately the conditional probability of dying in the interval $(t, t + dt)$ given survival past time t .

The two functions are intimately related:

$$\int_0^t \lambda(u)du = \int_0^t \frac{f(u)}{1 - F(u)} du = -\log[S(t)]. \quad (5.7)$$

Consequently,

$$S(t) = \exp\left\{-\int_0^t \lambda(u)du\right\}. \quad (5.8)$$

5.5.2 Actuarial and Product-Limit Survival Curves

Demographers have long used survival curves to describe the pattern of survival of a population over time. The typical demographic, or actuarial, survival curve starts with a population of 100,000 people and applies a set of probabilities of death $q_k (k = 1, 2, \dots, K)$ to the number of people alive at the beginning of each interval. These tables allow calculation of median time to death, expectation of life, and other parameters describing the life experience of a population. To calculate the K -th-year survival rate, we partition $S(\tau_K)$, the probability of surviving to time τ_K , as:

$$\begin{aligned} S(\tau_K) &= \Pr\{T > \tau_K\} \\ &= \Pr\{T > \tau_1\} \Pr\{T > \tau_2 | T > \tau_1\} \Pr\{T > \tau_3 | T > \tau_2\} \dots \Pr\{T > \tau_K | T > \tau_{K-1}\} \\ &= p_1 p_2 \dots p_K, \end{aligned} \quad (5.9)$$

where T is the survival time, τ_k is the time point at the end of the k th interval, and $p_k = 1 - q_k$ is the conditional probability of survival from $k - 1$ to k . The typical classical survival curve partitions the time course into equal intervals.

Classical *single-decrement* actuarial tables account only for a single mechanism, death, for removal from the population. Life tables for clinical studies, however, must account not only for deaths but also for losses and withdrawals. A person who withdraws is one who leaves the study without experiencing

the event of interest. These losses and withdrawals are collectively called *censoring*. If the trial is studying a cause-specific reason for death, a person may *withdraw* because of death from a different cause. If the outcome is something other than death, the withdrawal may be due to death. These withdrawals are called *competing risks*.

A censored observation in survival analysis is an observation that gives partial information about the time of the event of interest. For example, suppose one has designed a trial to study the survival time of women after the diagnosis of breast cancer. Rather than wait until all women in the study population die, one might recruit participants over a period of, say, 3 years and follow them until 5 years after the last person enters the trial. All those who die within the period of follow-up have provided full information on their time to death; however, those who are still alive at the end of the study have provided only partial information. A woman who has entered the study 2 years after it starts and is alive at the end of the study will have provided 6 years of data: we know only that her survival time is at least 6 years. This woman is *censored* at year 6. She is considered to have been *administratively censored* because the failure to know when she died is due to the design of the study. This type of censoring is possibly noninformative and, if so, produces no bias.

Another type of censoring may come from loss to follow-up. A woman was known alive at the end of time interval $(k - 1, k]$, but no information is available about her at the end of time interval $(k, k + 1]$; the study investigators do not know whether she is still alive or whether she died in the interval. She is said to be *lost*. This censoring may be noninformative, but being lost may be related to the chance of dying in the interval. A person who is very sick may opt not to continue in the study because her imminent death overwhelms her desire to participate; conversely, a person who is totally cured may not want to participate because she is too busy with other aspects of her life.

To understand the relationships among calendar time, study time, times of events, and times of censoring, recall the study of 20 participants depicted in Figure 5.4, which depicts the time course of the trial in terms of calendar time. In analyzing the data, however, we imagine each participant to have been randomized to the study at time 0. Now the time course of the trial is displayed as in Figure 5.5 by study time. The administratively censored observations no longer occur at the end of the graph, as they did in Figure 5.4, but at times during the period covered by the graph. Time starts at 0 in both the calendar time and study time graphs, but in the study-time graph, time ends at the longest time of follow-up. If the first patient randomized had been present at week 100, the graph would have run from 0 to 100.

Consider now Figure 5.6, which represents the intervals within which events and censorings occur. Here time is partitioned into equal intervals and a participant in the trial can be removed from the study in one of three ways: (1) by experiencing the event of interest, denoted by an \times ; by being censored either (2) because the entire study ends, denoted by a closed circle, or (3) because of some other reason like loss to follow-up or competing risk, denoted

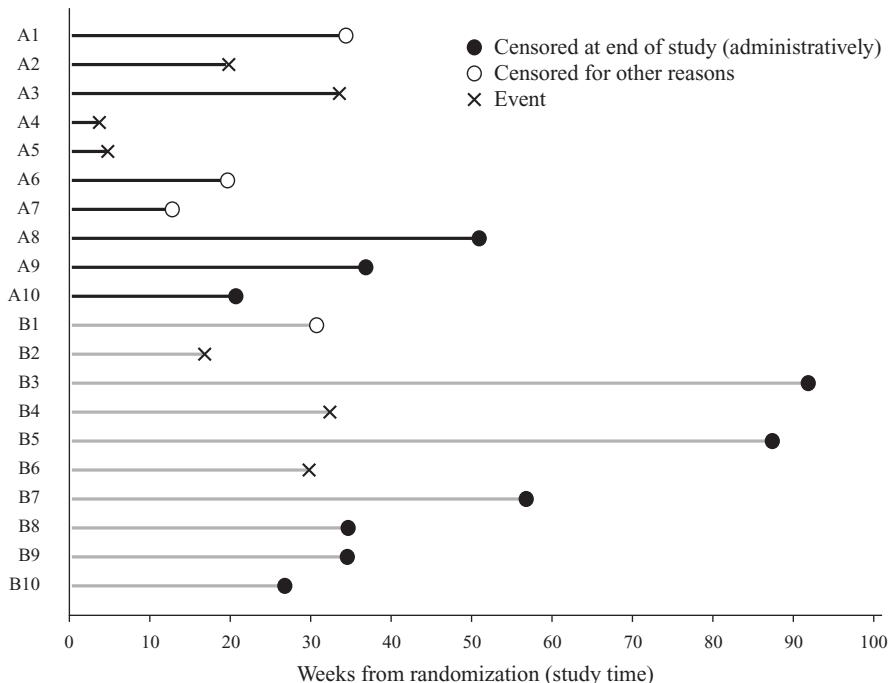


Figure 5.5 Experience of a clinical trial cohort in study time.

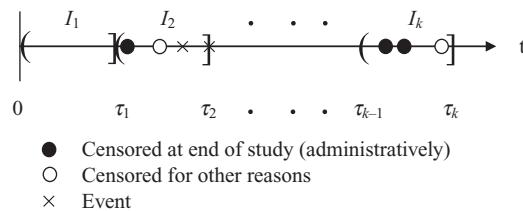


Figure 5.6 Equal intervals of time in a clinical trial. Note that some intervals do not contain any events; some contain more than one event; and some contain one or more censorings; some contain both events and censorings.

by an open circle. Note that censoring and events can occur either within the interval or at the end of the interval. (Some actuarial life tables used in demography partition the first year of life into months while subsequent partitions represent years. In general, an actuarial life table uses prespecified intervals, not necessarily equal sized ones.)

A life table represents the data in terms of study time. Cutler and Ederer (1958) show the calculation of 5-year survival rate from an actuarial life table relevant to a clinical study of 126 patients diagnosed at time 0. Table 5.2, also

Table 5.2 Actuarial Life Table

t_k	n_k	d_k	l_k	w_k	$m_k = n_k - (l_k + w_k)/2$	$q_k = d_k/m_k$	$p_k = 1 - q_k$	$S(t_k) = \prod_{i=1}^k p_i$
Years after Diagnosis	Alive at Start of Interval	Died during Interval	Lost to Follow-Up during Interval	Withdrawn Alive during Interval	Effective Number Exposed to the Risk of Dying	Proportion Dying	Proportion Surviving	Cumulative Proportion Surviving from Diagnosis Through end of Interval
0-1	126	47	4	15	116.5	0.40	0.60	0.60
1-2	60	5	6	11	51.5	0.10	0.90	0.54
2-3	38	2	0	15	30.5	0.07	0.93	0.50
3-4	21	2	2	7	16.5	0.12	0.88	0.44
4-5	10	0	0	6	7.0	0.00	1.00	0.44

given in Miller (1981), shows the life table when the data are categorized by year of death. Three columns distinguish life tables in the clinical setting from single-decrement actuarial life tables: l_k , the losses to follow-up; w_k , those withdrawn alive; and m_k , the effective number at risk for dying during the interval. The calculations assume that the losses and withdrawals occur half-way through the interval so that if n_k represents the number alive at the beginning of the interval, $m_k = n_k - (l_k + w_k)/2$. Therefore, the probability of dying is the number of deaths d_k divided not by the number at the beginning of the interval, but by the number assumed to be present at its mid-point. This type of life table is often called a *multiple decrement* table because a person has more than one way to exit from the table within a given time period. Interest lies in the exits due to death, but the calculations must account for other methods of exit. In the parlance of clinical trials, as described above, these other exits are called *censoring*.

In 1926, Major Greenwood (1926) derived a formula for the variance of the estimated survival probability $\hat{S}(\tau_K)$ through its logarithm and successive applications of the delta method. Consider:

$$\log[\hat{S}(\tau_K)] = \sum_{k=1}^K \log(\hat{p}_k).$$

Under the assumption that each \hat{p}_k is binomially distributed, application of the delta method (see HK1, section 6.10.3) yields the following derivation for the variance:

$$\text{Var}[\log(\hat{p}_k)] \sim \text{Var}(\hat{p}_k) \left[\frac{d}{dp_k} \log(p_k) \right]^2 \sim \frac{p_k q_k}{m_k} \frac{1}{p_k^2} = q_k / (m_k p_k). \quad (5.10)$$

Assuming now that each $\log(\hat{p}_k)$ is independent of every other $\log(\hat{p}_k)$,

$$\text{Var}(\log[\hat{S}(\tau_K)]) = \sum_{k=1}^K q_k / (m_k p_k). \quad (5.11)$$

Since $\hat{q}_k = d_k/m_k$, the estimated variance is therefore:

$$\widehat{\text{Var}}(\log[\hat{S}(\tau_K)]) = \sum_{k=1}^K \hat{q}_k / (m_k \hat{p}_k) = \sum_{k=1}^K d_k / [m_k (m_k - d_k)] \quad (5.12)$$

Another application of the delta method yields Greenwood's formula:

$$\widehat{\text{Var}}(\hat{S}(\tau_K)) = [\hat{S}^2(\tau_K)] \sum_{k=1}^K d_k / [m_k (m_k - d_k)]. \quad (5.13)$$

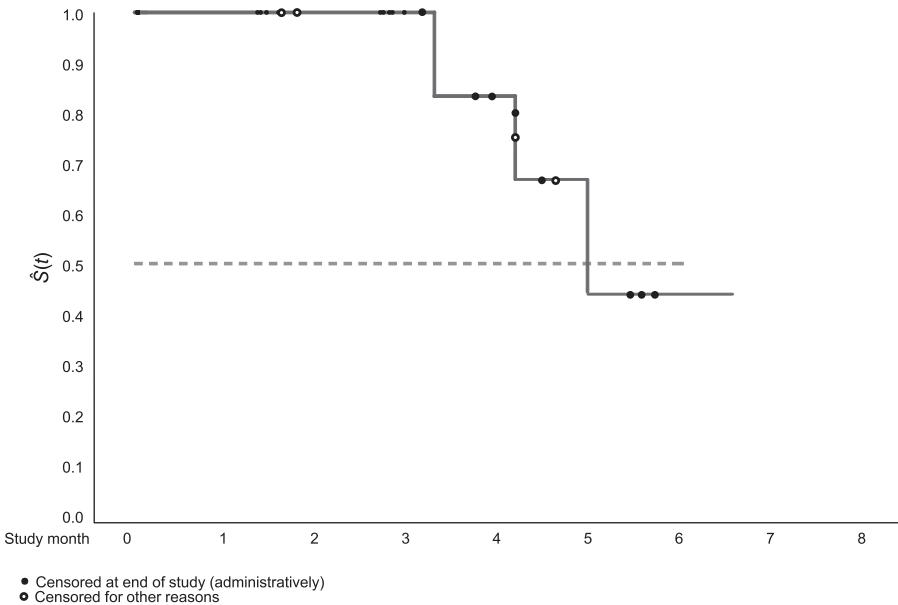


Figure 5.7 A Kaplan–Meier curve.

In a seminal paper, Kaplan and Meier (1958) proposed a modification of the standard life table to apply to clinical trials. Instead of visualizing the steps representing events occurring at each prespecified period of time, the curve would step down each time an event occurred. A patient would be removed from the risk set at the time of censoring. The resulting so-called *product-limit curve*, often called a *Kaplan–Meier curve*, would describe the experience of the cohort of patients randomized into a trial. The median survival time could be picked off the curve, and the Greenwood formula already developed for actuarial survival curves could be applied to calculate standard errors of survival at different times. A Kaplan–Meier curve for the data of the trial depicted in Figure 5.4 is shown in Figure 5.7. Note that for illustrative purposes, the curve includes all 20 participants. A real clinical trial would have a curve for each treatment group.

To produce the curve of Figure 5.7, we had to consider time in a slightly different way from the depiction in Figure 5.6. Again, as before, imagine that the study, as expressed in study time t rather than calendar time, runs from 0 to T , that is, $0 < t \leq T$. But now, rather than dividing the interval $(0, T]$ into equal intervals of time, we partitioned the timeline into intervals defined by the occurrence of an event. Thus, as shown in Figure 5.8, events always occur at the end of intervals and censoring occurs within them. Of course, in practice, some events occur simultaneously, but for the purpose of this chapter, we assume no tied times.

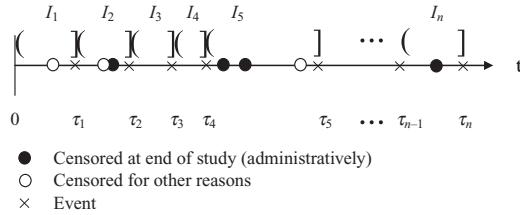


Figure 5.8 Intervals of time in a clinical trial defined by times of events.

Recall from Equation (5.9) that in the setting of the actuarial life table, the estimated probability $\hat{S}(\tau_K)$ of survival to time τ_K , where time is partitioned into prespecified, often equal, time periods, is:

$$\hat{S}(\tau_K) = \prod_{k=1}^K \hat{p}_k. \quad (5.14)$$

To generalize this equation to the product-limit setting, observe for the i th participant in the trial the pair $X_i = \{\min(T_i, C_i), \delta_i\}$ where δ_i is an indicator function identifying whether an observation is a time of event or a censoring. That is,

$$\delta_i = I(T_i \leq C_i) = \begin{cases} 1 & \text{if } T_i \leq C_i; \text{ i.e., the time of event } T_i \text{ is observed (not censored)} \\ 0 & \text{if } T_i > C_i; \text{ i.e., the time of event } T_i \text{ is censored} \end{cases}$$

Thus, the pair X_i tells us the last time of observation of the person and whether that observation represents an event or a censoring time. In the absence of ties, let $X_{[1]} < X_{[2]} \dots < X_{[n]}$ represent the order statistics of X_1, X_2, \dots, X_n and define $\delta_{[i]}$ as the value of δ_i that corresponds to the associated time $X_{[i]}$. At an arbitrary time t , the cohort at risk for an event is the set of subjects still alive immediately prior to t (i.e., at t^-).

Define the following quantities:

N = the total number of participants in the trial

N_i = the number of participants alive at time just prior to $X_{[i]}$

d_i = the number who died at time $X_{[i]}$; note that because we are not allowing ties, d_i is either 0 or 1.

The product-limit estimator of $S(t)$ is:

$$\hat{S}(t) = \prod_{X_{[i]} \leq t} \left(\frac{N-i}{N-i+1} \right)^{\delta_{[i]}}.$$

The Greenwood formula for the product limit estimator is:

$$\widehat{Var}[\hat{S}(t)] = [\hat{S}^2(t)] \sum_{\delta_i \leq t} \delta_i / [(N-i)(N-i+1)]. \quad (5.15)$$

For derivations of these results, and for analogous formulas when some of the events are tied, see Miller (1981).

5.5.3 Exponential Survival, Hazard Rates, and Ratios and Proportional Hazard Ratios

When survival time has an exponential distribution, the formulas simplify considerably. In that case, the hazard $\lambda(t)$, as given in Equation (5.5) is a constant λ , and the survival function Equation (5.8) is:

$$S(t) = e^{-\lambda t}. \quad (5.16)$$

If both the experimental and control groups have exponential distributions with hazards λ_E and λ_C , respectively, then the hazard ratio

$$\theta = \lambda_E / \lambda_C.$$

becomes the parameter of interest for testing whether the two distributions differ. Note that statistical tests of the hazard ratio are based not on the hazard ratio itself but on its logarithm because, under mild conditions, the latter is approximately normally distributed. For a justification of this assertion, see, for example, Peto and Peto (1972).

Thus, if the two time-to-event random variables being compared have exponential distributions, the resulting statistical theory is simple and well described in many elementary textbooks in statistics or probability. In actual clinical trials, the assumption of exponential distribution is usually considered too strong to form the basis of statistical testing, although, as we shall see below, exponential distributions are often assumed for the purpose of calculating sample size. A generalization of the exponential distribution that allows semi-parametric modeling is to assume that the distributions of interest have proportional hazards. In other words, while $\lambda_E(t)$ and $\lambda_C(t)$ are not constant, their ratio is, that is,

$$\theta(t) = \frac{\lambda_E(t)}{\lambda_C(t)} = \theta. \quad (5.17)$$

5.5.4 The Logrank Family of Tests

The Kaplan–Meier curve and the associated Greenwood formulas provide estimates of survival probabilities at various times. The goal of most randomized clinical trials, however, is to compare the two survival curves (or the set

Table 5.3 A 2×2 Table at the Time of the j th Event

Group	Dead	Alive	Total
Experimental	$O_j = 1$ if the j th death occurs in the experimental arm $= 0$ otherwise	$m_j - O_j$	m_j
Control	$1 - O_j$ 1	$n_j - (1 - O_j)$ $N_j - 1$	n_j $n_j + m_j = N_j$

of survival curves), testing whether they differ significantly from each other and estimating the magnitude of that difference. Mantel (1966) proposed a simple nonparametric test to compare survival curves. Peto and Peto (1972) called a version of this test the *logrank test*. They showed that it has locally optimal power if the hazard ratio is constant over time and its expected value is independent of the patterns of censoring. If the hazards are not proportional (i.e., the hazard ratio is not constant), the test is still valid in the sense that it is unbiased under the null hypothesis, but it is no longer optimal.

The logrank and related tests are variants of the Mantel–Haenszel test (Mantel and Haenszel 1959) for 2×2 tables. Recall that time in the survival setting is partitioned into semi-open intervals that end when a death (or, more generally, the outcome event of interest) occurs. Therefore, the survival experience can be characterized by a series of 2×2 tables, each one with a single event. Table 5.3 shows the j th such table. A trial with D deaths would have D such tables.

Let $T_{[j]}$ represent the time of the j th death. If at time t , just before the j th death, a total of m_j participants are in the experimental group and n_j are in the control group, then the numerator of the Mantel–Haenszel statistic is:

$$S_{MH} = \sum_{T_{[j]} \leq t} (O_j - E_j),$$

where E_j is the expected number of deaths in the experimental group which, under the null hypothesis, is m_j/N_j . A weighted version of the above statistic would be

$$S_{WMH} = \sum_{T_{[j]} \leq t} w_j(O_j - E_j),$$

where w_j is a weight and $\sum_j w_j = 1$.

To create a statistic with a standard normal distribution, divide S_{WMH} by its estimated standard error to produce:

$$Z = \frac{\sum_{T_{[j]} \leq t} w_j(O_j - E_j)}{\sqrt{\sum_{T_{[j]} \leq t} w_j^2 E_j(1 - E_j)}}. \quad (5.18)$$

Table 5.4 Some Examples of the Weighted Logrank Family of Tests

Weight (w_j)	Test	Comment
1	Logrank	The logrank is the most commonly used of this family of tests.
$N - j + 1$ where $N = \sum_j N_j$	Wilcoxon	Appropriate when there is no censoring.
N_j	Gehan	Applicable when there is censoring; however, in this test, censoring and survival are confounded. See Lan and Witter (1986).
$S(T_{[j]})$	Peto–Prentice	$S(T_{[j]})$ is the Kaplan–Meier estimate of the probability of surviving to time $T_{[j]}$. This test modifies the Gehan test in a way that does not confound censoring and survival.

If randomization is stratified, then one can perform a stratified logrank test that is a straightforward generalization of the unstratified version of Equation (5.14).

The statistic in Equation (5.18) has various names depending on the choice of w_j , as illustrated in Table 5.4.

5.5.5 The Cox Proportional Hazards Model

Investigators often want more than a test of significance. Typically, in the survival setting, they want to calculate the estimated hazard ratio along with its 95% confidence limits. They may further want to adjust the estimated hazard ratio θ for baseline covariates that may differ between the treatment groups. (If the trial is randomized, the groups should be well balanced.) The Cox proportional hazards model addresses these issues. The model is a generalization of the logrank test, but now the assumption of proportional hazards is more central to the validity of the model.

As described above, assume each person is followed until an event or censoring occurs. Assume also that the i th person is characterized by an s -dimensional vector of covariates $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{is})'$. Writing the hazard function (Eq. 5.5) to show its dependence on \mathbf{x} ,

$$\lambda(t; \mathbf{x}) = \frac{f(t; \mathbf{x})}{1 - F(t; \mathbf{x})}$$

and assuming that the hazards are proportional, we have

$$\lambda(t; \mathbf{x}) = \exp\{\boldsymbol{\beta}' \mathbf{x}\} \lambda_0(t), \quad (5.19)$$

where $\beta = (\beta_1, \beta_2, \dots, \beta_s)'$ is a vector of regression coefficients. Note that time enters the model only through the function $\lambda_0(t)$. The Cox model allows estimation of the parameter β where each component β_i represents the log hazard ratio associated with variable x_i . In this framework, the form of $\lambda_0(t)$ does not follow any particular distribution, so that estimation of β can occur without having to make any assumptions about $\lambda_0(t)$. The model is therefore considered *semi-parametric*. In the special case where the only covariate is the yes/no variable indicating whether an observation comes from an experimental or control participant, β has only one element and $\exp(\beta)$ is the estimated hazard ratio. Further, Cox derived a partial likelihood for the model. If no observation is censored, then the score test from this partial likelihood and the logrank test are asymptotically identical. See Miller (1981) and Kalbfleisch and Prentice (2002) for detailed descriptions and derivations of the Cox model.

5.5.6 Some Sample SAS Code

The following code shows a SAS program applicable to the data of Figure 5.5. The program shows the data, the code necessary to create a life table, and the code for the Cox model.

```
data TIME2_EG;
  input PID$ TRT DAYS2DTH F_DEATH ;
  cards;
*The first variable is the participant's identifier. The first
numeric column shows the treatment group, the second shows the
number of days from randomization until the participant died or
was censored; and the last column indicates whether the par-
ticipant died (1) or was censored (0). For ease of reading this
example, the participant's identifier includes the treatment
code A or B. In a real example, the treatment code would not be
part of the identifier.:
A1 1 33 0
A2 1 19 1
A3 1 33 1
A4 1 3 1
A5 1 4 1
A6 1 19 0
A7 1 12 0
A8 1 50 0
A9 1 36 0
A10 1 20 0
B1 2 30 0
B2 2 16 1
B3 2 91 0
B4 2 32 1
B5 2 87 0
B6 2 29 1
```

```
B7 2 56 0
B8 2 34 0
B9 2 34 0
B10 2 30 0
;
run;
```

SAS's PROC LIFETEST generates life tables and a logrank test. The *time* statement specifies the value for censoring. Consistent with the input data set above, we use 0 for censoring and 1 for failure.

```
* Sample PROC LIFETEST;

PROC LIFETEST data = TIME2_EG;
    time      DAYS2DTH*f_death(0);
    strata   TRT;
    id       PID;
run;
```

The resulting output contains two life tables, one for each treatment. The table orders the event times for each participant, marks the censored observations with an asterisk (*), and lists both the number and proportion surviving or failing, taking censoring into account.

Time-to-event analyses often report the median and compare groups with respect to this statistic. The median is the value for which $S(t) = 0.50$. For this example, in both groups, the proportion surviving is always above 50%. Thus, neither distribution has an estimated median, which is reflected in the SAS output.

The procedure also automatically yields a logrank test and a Wilcoxon test. Note that what SAS terms the Wilcoxon test is what this chapter terms the Gehan test. The logrank test indicates that the two treatment groups are not significantly different ($p = 0.43$).

The LIFETEST Procedure

Stratum 1: TRT = 1							
Product-Limit Survival Estimates							
DAYS2DTH	Survival		Standard		Number Failed	Number Left	PID
	Survival	Failure	Error				
0.0000	1.0000	0	0		0	10	
3.0000	0.9000	0.1000	0.0949		1	9	A4
4.0000	0.8000	0.2000	0.1265		2	8	A5
12.0000*	.	.	.		2	7	A7

19.0000	0.6857	0.3143	0.1515	3	6	A2
19.0000*	.	.	.	3	5	A6
20.0000*	.	.	.	3	4	A10
33.0000	0.5143	0.4857	0.1870	4	3	A3
33.0000*	.	.	.	4	2	A1
36.0000*	.	.	.	4	1	A9
50.0000*	.	.	.	4	0	A8

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable DAYS2DTH

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval	
		[Lower]	Upper)

75	.	33.0000	.
50	.	19.0000	.
25	19.0000	3.0000	.

Mean	Standard Error
------	----------------

25.5000	4.3664
---------	--------

NOTE: The mean survival time and its standard error were under-estimated because the largest observation was censored and the estimation was restricted to the largest event time.

Stratum 2: TRT = 2

Product-Limit Survival Estimates

DAYS2DTH	Survival	Failure	Survival				PID
			Standard Error	Number Failed	Number Left		
0.0000	1.0000	0	0	0	10		
16.0000	0.9000	0.1000	0.0949	1	9	B2	
29.0000	0.8000	0.2000	0.1265	2	8	B6	
30.0000*	.	.	.	2	7	B1	
30.0000*	.	.	.	2	6	B10	
32.0000	0.6667	0.3333	0.1610	3	5	B4	
34.0000*	.	.	.	3	4	B8	
34.0000*	.	.	.	3	3	B9	
56.0000*	.	.	.	3	2	B7	
87.0000*	.	.	.	3	1	B5	
91.0000*	.	.	.	3	0	B3	

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable DAYS2DTH

Quartile Estimates

Percent	Point Estimate	95% Confidence Interval [Lower Upper)
75	.	.
50	.	32.0000 .
25	32.0000	16.0000 .

Mean Standard Error

30.1000	1.8530
---------	--------

NOTE: The mean survival time and its standard error were under-estimated because the largest observation was censored and the estimation was restricted to the largest event time.

Summary of the Number of Censored and Uncensored Values

Stratum	TRT	Total	Failed	Censored	Percent Censored
1	1	10	4	6	60.00
2	2	10	3	7	70.00
Total		20	7	13	65.00

The LIFETEST Procedure

Testing Homogeneity of Survival Curves for DAYS2DTH over Strata

Rank Statistics

TRT	Log-Rank	Wilcoxon
1	1.0249	19.000
2	-1.0249	-19.000

Covariance Matrix for the Log-Rank Statistics

TRT	1	2
1	1.68755	-1.68755
2	-1.68755	1.68755

Covariance Matrix for the Wilcoxon Statistics

TRT	1	2
1	403.000	-403.000
2	-403.000	403.000

Test of Equality over Strata

Test	Chi-Square	DF	Pr >
			Chi-Square
Log-Rank	0.6225	1	0.4301
Wilcoxon	0.8958	1	0.3439
-2Log (LR)	1.5225	1	0.2172

The following SAS code using PROC PHREG (short for *proportional hazards regression*) also needs an indicator for censoring. As before, the value for censoring is 0. The output of this program is a Cox model with a single covariate, the treatment indicator. The procedure may accommodate further covariates by adding the desired variable names on the model statement. While PROC LIFETEST presents a test statistic and associated *p*-value, PROC PHREG can provide an estimate of the hazard ratio, as well as 95% confidence limits.

*Sample Cox model;

```
PROC PHREG data=TIME2_EG;
  model DAYS2DTH*F_DEATH(0) = TRT /rl;
  ods output parameterestimates=HAZARD;
  run;

  proc print data = HAZARD;
  run;
```

The output includes parameter estimates, as well as statistics describing the goodness of fit of the model and global tests of the hypothesis of vector $\beta = 0$. Note that the Cox model yields a significance value for the Wald test of 0.4367, which is close to the logrank *p*-value of 0.4301 from PROC LIFETEST. In the special case using only treatment in the model, the score test from the Cox model is equal to the logrank test from PROC LIFETEST (*p* = 0.4301). Recall that the *p*-values from the output come from statistics that are approximations to the true values; when one actually reports the results of such analysis, the cited *p*-value should have fewer significant digits.

The parameter estimate for treatment (variable TRT) is the logarithm of the hazard ratio. The hazard ratio is 0.55 with a 95% confidence interval of (0.12, 0.48); the risk of death in treatment group 1 is approximately half that of treatment group 2.

The PHREG Procedure

Model Information

Data Set	WORK.TIME2_EG
Dependent Variable	DAYS2DTH
Censoring Variable	F_DEATH
Censoring Value(s)	0
Ties Handling	BRESLOW

Number of Observations Read	20
Number of Observations Used	20

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent
			Censored
20	7	13	65.00

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without	With
	Covariates	Covariates
-2 LOG L	37.221	36.608
AIC	37.221	38.608
SBC	37.221	38.554

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	0.6134	1	0.4335
Score	0.6225	1	0.4301
Wald	0.6050	1	0.4367

Analysis of Maximum Likelihood Estimates

Variable	Parameter	Standard	Hazard			
	DF	Estimate	Error	Chi-Square	Pr > ChiSq	Ratio
TRT	1	-0.59837	0.76932	0.6050	0.4367	0.550
95% Hazard Ratio						
Confidence Limits						
0.122		2.483				

Obs	VARIABLE	DF	ESTIMATE	STDERR	CHISQ	PROBCHISQ	HAZARDRATIO
1	TRT	1	-0.59837	0.76932	0.6050	0.4367	0.550
	RLOWERCL						
0.122	HRUPPERCL		.483				

5.5.7 Some Sample Splus Code

Splus produces similar results using a different set of functions. The Splus and SAS data sets have nearly the same structure:

```
> # show dataset time2eg
> time2eg
  PID TRT DAYS2DTH F.DEATH
 1  A1   1      33     0
 2  A2   1      19     1
 3  A3   1      33     1
 4  A4   1       3     1
 5  A5   1       4     1
 6  A6   1      19     0
 7  A7   1      12     0
 8  A8   1      50     0
 9  A9   1      36     0
10 A10  1      20     0
11 B1   2      30     0
12 B2   2      16     1
13 B3   2      91     0
14 B4   2      32     1
15 B5   2      87     0
16 B6   2      29     1
17 B7   2      56     0
18 B8   2      34     0
19 B9   2      34     0
20 B10  2      30     0
>
```

The three Splus functions, `survfit`, `survdiff`, and `coxph`, have similar structures, requiring outcome (`DAYS2DTH`), censoring (`F.DEATH`), and predictor (`TRT`) variables. The code below defines three Splus objects, `tableeg`, `logrankeg`, and `coxeg`, corresponding to specific calls of these functions.

```
# survival table
tableeg_survfit(Surv(DAYS2DTH, F.DEATH) ~ TRT, data = time2eg)
summary(tableeg)

# log-rank test
logrankeg_survdiff(Surv(DAYS2DTH, F.DEATH) ~ TRT, data = time2eg, method="breslow")
logrankeg

# cox model
coxeg_coxph(Surv(DAYS2DTH, F.DEATH) ~ TRT, data = time2eg, method="breslow")
coxeg
summary(coxeg)
```

The object `tableeg` uses the `survfit` function and produces summary statistics for the survival data; `summary(tableeg)` produces the life tables for each treatment group.

```
> tableeg_survfit(Surv(DAYS2DTH, F.DEATH) ~ TRT, data = time2eg)

> summary(tableeg)
Call: survfit(formula = Surv(DAYS2DTH, F.DEATH) ~ TRT, data = time2eg)

          TRT=1
time n.risk n.event survival std.err lower 95% CI upper 95% CI
  3      10      1     0.900  0.0949   0.732        1
  4       9      1     0.800  0.1265   0.587        1
 19      7      1     0.686  0.1515   0.445        1
 33      4      1     0.514  0.1870   0.252        1

          TRT=2
time n.risk n.event survival std.err lower 95% CI upper 95% CI
 16      10      1     0.900  0.0949   0.732        1
 29       9      1     0.800  0.1265   0.587        1
 32       6      1     0.667  0.1610   0.415        1
```

The function `survdiff` produces a logrank test, with a *p*-value that is the same as the logrank *p*-value seen in proc lifetest using SAS.

```
> # log-rank test
> logrankeg_survdiff(Surv(DAYS2DTH, F.DEATH) ~ TRT, data = time2eg)
> logrankeg
Call:
survdiff(formula = Surv(DAYS2DTH, F.DEATH) ~ TRT, data = time2eg)
```

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
TRT=1	10	4	2.98	0.353	0.622
TRT=2	10	3	4.02	0.261	0.622

Chisq= 0.6 on 1 degrees of freedom, p= 0.43

The call of the function `coxph` also uses the method option to handle ties in the same way as PROC PHREG, which uses the Breslow method. Typing `summary(coxeg)` produces further helpful output, such as summary estimates for the hazard ratio, corresponding confidence limits, and additional test statistics for the global hypothesis of the vector $\beta = 0$.

```
> # cox model
> coxeg_coxph(Surv(DAYS2DTH, F.DEATH) ~ TRT, data= time2eg,
method="breslow")

> summary(coxeg)
Call:
coxph(formula = Surv(DAYS2DTH, F.DEATH) ~ TRT, data= time2eg,
method = "breslow")

n= 20

      coef  exp(coef)  se(coef)      z      p
TRT-0.598     0.55     0.769   -0.778 0.44

      exp(coef)  exp(-coef) lower .95 upper .95
TRT      0.55      1.82    0.122     2.48

Rsquare= 0.03 (max possible= 0.844 )
Likelihood ratio test= 0.61 on 1 df,  p=0.434
Wald test          = 0.6 on 1 df,  p=0.437
Score (logrank) test = 0.62 on 1 df,  p=0.43
```

5.5.8 Calculations of Number of Replications, or Sample Size

In the experimental design literature, the term *replication* refers to repeated applications of a treatment to different experimental units (e.g., HK1, chapter 6) in the context of an appropriate error-control design. In randomized clinical trials, however, the individual units are people. As explained above, clinical trials typically deliberately recruit a heterogeneous population that allows formal inference to a wide target population. Perhaps for this reason, the field of randomized clinical trials conventionally uses the term *sample size* to refer to the *replications* in other types of experiments.

Various methods are available to calculate sample size for a logrank test. They range from very simple models to models that account for many deviations from ideal. This section describes four approaches in increasing order of complexity.

Before beginning, we present a remarkable result: without making any assumption about the shape of the survival curve, it is easy to show that at the time of the j 'th death (see Table 5.3), the variance V_j of the number of events O_j under the null hypothesis is approximately $1/4$. This result stems directly from the construction of the logrank test. Under the null hypothesis, we expect an equal number of observations in each group. Because the logrank test partitions each interval when an event occurs, the probability that the event will occur in any given group is $1/2$. Therefore, $E(V_j) = E\{E_j(1 - E_j)\} \approx \frac{1}{4}$, and the variance of the total number of deaths is just $N/4$. This simple relationship allows a very simple formula for the sample size. As shown by Schoenfeld (1983), for a two-sided level α test with power $1 - \beta$, the logrank tests requires the total number of events in the two treatment groups to be approximately:

$$4 \frac{(\xi_{1-\alpha/2} + \xi_{1-\beta})^2}{[\ln(\theta)]^2}. \quad (5.20)$$

where ξ_p is the p th percentile of the standard normal distribution, and θ is the hazard ratio.

The total sample size required in each treatment group is:

$$\frac{4}{(\pi_e + \pi_c)} \frac{(\xi_{1-\alpha/2} + \xi_{1-\beta})^2}{[\ln(\theta)]^2}. \quad (5.21)$$

where π_e and π_c are the probabilities of experiencing the event during the trial in the experimental and control group, respectively.

If the ratio of allocation to experimental and control group is $r:1$ rather than $1:1$, the “4” in Equation (5.21) becomes $(r + 1)^2/r$.

Freedman (1982) derived the following formula for use under equal allocation to the treated and control groups. Substituting Equation (5.20) by:

$$\left(\frac{\theta+1}{\theta-1}\right)^2 (\xi_{1-\alpha/2} + \xi_{1-\beta})^2.$$

gives a sample size per group of

$$\frac{1}{(\pi_e + \pi_c)} \left(\frac{\theta+1}{\theta-1}\right)^2 (\xi_{1-\alpha/2} + \xi_{1-\beta})^2. \quad (5.22)$$

As Table 5.5 shows, these two formulas give very similar results. The required sample size per group is the factor displayed times the ratio:

$$\frac{(\xi_{1-\alpha/2} + \xi_{1-\beta})^2}{(\pi_e + \pi_c)}.$$

Note that neither Equation (5.21) nor Equation (5.22) explicitly incorporates time. In fact, time only occurs in calculating the probabilities of events π_e and π_c . But of course, in practice, time plays many roles in designing a study

Table 5.5 Comparison between the Schoenfeld and Freedman Formulas for Computing Sample Size of the Logrank Test

Hazard Ratio θ	Schoenfeld Factor	Freedman Factor
	$4/[\ln(\theta)]^2$	$\left(\frac{\theta+1}{\theta-1}\right)^2$
0.5	8.3	9.0
0.6	15	16
0.7	31	32
0.8	80	81
0.9	360	361
0.95	1520	1521

comparing survival times. The sample size formulas above make simplifying assumptions about the nature of the trial and the behavior of the participants. The approaches above assume that all participants in the trial are fully compliant with therapy, that all are followed until the end of the study, and that each participant's outcome is assessed. In actual trials, many of these assumptions are violated, and, in our experience, are violated in a direction that requires increased sample size to achieve the desired power. Two general approaches are available that allow relaxation of the assumptions. Some methods assume an underlying failure time model but allow deviations from that model. Perhaps the most commonly used of these methods is one due to Lachin and Foulkes (1986). Another set of approaches allows the designer of the trial considerable latitude in specifying the likely course of the trial under both the null and alternative hypotheses. These methods, developed by Lakatos (1986, 1988), Halpern and Brown (1993), and Shih (1995) all account for dropouts, drop-ins, losses, nonproportional hazards, and other deviations from assumptions.

Lachin and Foulkes assume that recruitment times, failure times, and loss-to-follow-up times all follow exponential distributions with different parameters. Lakatos (1986, 1988) introduced Markov models for calculating sample sizes. Rather than assuming exponentially distributed times to failure, loss, drop-in, and dropout, he breaks time into units (weeks or months) and allows the user to assign binomial probabilities to each event in each time period. This added flexibility allows one to model the expected time course of the trial. Shih (1995) extended his methods for even more complex settings.

Finally, the methods of Halpern and Brown (1993) allow the investigator to sketch an assumed survival curve and then simulate trials from that curve to calculate the sample size that gives appropriate power.

The various methods can yield very different sample sizes. As seen above, the Freedman and Schoenfeld formulas produce very similar answers. The other methods can produce much larger sample sizes because they account for deviations from assumptions that tend to attenuate the effect size. Lakatos and Lan (1992) summarize the various approaches to sample size calculation for the logrank test. A reasonable strategy for design is to use several different

approaches and to see how close the answers are to each other. Varying the parameters over reasonable ranges will allow the designer of the trial to assess the sensitivity of the samples size to the various assumptions.

5.5.9 Group Sequential Analysis

A common feature of many randomized clinical trials, and one that sets them apart from many other types of experiments, is the use of sequential analysis. Recruitment often occurs over an extended period of time, sometimes over years. Moreover, the treatments studied often have harmful side effects as well as potential benefits. As a result, the design of a trial that studies time-to-event typically includes methods for terminating the trial early in a way that preserves the type I error rate should the trial stop for benefit. The standard approach is to choose a so-called *group sequential* method (see Chapter 6). In classical sequential analysis, a decision about whether to continue to add observations is made after each randomized experimental unit is evaluated. In group sequential designs, these decisions occur after periodic “looks” at the interim accumulating data; thus, the decisions are made not after each participant’s outcome is known, but after a group of events have occurred. The recommendation comes from a committee independent of the investigators. These committees—often called *Data Monitoring Committees* or *Data and Safety Monitoring Committees*—look at the cumulative data several times during the trial and make recommendations either to continue the trial or to stop it. The latter is accompanied with a declaration of benefit, a concern for harm, or a conclusion that the trial is “futile”—that is, the trial will not yield meaningful answers if continued to its planned completion. The literature on group sequential analysis is large, and the methods are beyond the scope of this chapter. We refer the interested reader to four books, two that describe the operations of these committees and two that present more statistical discussions of the methods of group sequential analysis. Ellenberg, Fleming, and DeMets (2002) and Herson (2009) describe the operations of the committees that make these recommendations. The other two books, Jennison and Turnbull (2002) and Proschan, Lan, and Wittes (2006), provide detailed presentations of the relevant statistical methodology.

5.6 OTHER TOPICS

This very brief section mentions a few issues that arise frequently in clinical trials. Each of these topics has spawned a large literature.

5.6.1 Multiplicity

HK1 (chapter 7) presents various methods for dealing with multiplicity. In clinical trials, especially those that will be used for the basis of regulatory

approval, these methods are important not only because of their scientific rigor, but also because regulators are very cautious about ensuring the type I error rate is not inflated. A study that compares more than one treatment to control, that has more than one primary outcome, or that uses sequential analysis should incorporate methods for strict preservation of the type I error rate. Standard methods of group sequential analysis control the type I error rate, but when a study has a sequential design and other features that can lead to multiplicity, the statistician designing the trial must consider the experiment as a whole, ensuring preservation of the experiment-wise type I error rate.

Several methods are commonly used for controlling multiplicity in clinical trials. When the study has more than a single primary outcome, the type I error rate is split between them. For example, if a study has two primary outcomes, and an overall two-sided type I error rate of 0.05, then the error rate may be split into two equal (i.e., each outcome is assigned a two-sided error rate of 0.025) or unequal parts (e.g., one outcome is assigned an error rate of 0.01 and the other a rate of 0.04). This type of allocation ensures that under the null hypothesis, the probability is less than 0.05 that at least one outcome is statistically significant.

When the study has more than one experimental group and a control group, then the usual primary tests of significance compare each experimental arm to the control. Dunnett's test (e.g., HK1, chapter 7) is then used to adjust for multiplicity.

Perhaps the most contentious situation arises when a study has more than one secondary outcome. Some trials have two or three secondary outcomes, but as described earlier, many trials have a host of outcomes designated as "secondary." We recommend limiting the formal secondary outcomes to a few that have adequate power. (Adequate is, of course, in the eyes of the beholder, but we consider anything less than 50% power as not adequate.) Having defined a limited set of secondary outcomes, the investigators face a choice of strategies to control the type I error rate. One commonly used approach is to list the secondaries in a hierarchical order, and, if the primary outcome is statistically significant, then the investigators test the secondaries one at a time in the prespecified order. Suppose, for example, the study has listed s secondary outcomes, O_1, O_2, \dots, O_s . If the p -value for O_r is less than, say 0.05, then the investigator can test O_{r+1} . In other words, testing stops when the first outcome is not statistically significant. This procedure provides strict control of the type I error rate, but it places a high premium on correct *a priori* ordering of the hypotheses. A nominally significant hypothesis far down the list may not be declared statistically significant. Consider a study with four secondary outcomes where the nominal p -values are, in the prespecified order, 0.026, 0.052, 0.001, and 0.0002. Strict adherence to the hierarchical principle would declare O_1 statistically significant, but because the p -value for O_2 is above 0.05, neither O_3 and O_4 would be called "significant." When investigators are not allowed to make a claim when p -values are as small as 0.001 or 0.0002, they

often express the view that statistical principle has trumped science and common sense.

Another commonly used method is a Bonferroni correction, or, preferably, one of its many variants. These methods, while in some cases less statistically powerful than the hierarchical method, do not share the problem described above; however, there are situations in which hierarchical methods allow claims that the Bonferroni and related methods do not.

Statisticians often find it challenging to convince investigators to limit their secondary outcomes and to specify a strategy for testing them that preserves the type I error rate. We recommend in the design phase of a trial that the statistician provide the clinical investigators with the statistical power for each secondary outcome along with a range of methods for dealing with multiplicity and a set of hypothetical results for the various methods. This information as part of the planning of a trial can help investigators choose a reasonable number of secondary hypothesis and a sensible strategy for testing them.

5.6.2 Subgroups

Often, investigators are interested not only in the effect of a treatment on the study population as a whole but in its effect on various subgroups of the population. The goal is to ask whether the overall result of the trial applies to subsets defined by various baseline parameters. If the trial overall shows benefit, practicing physicians want to know whether some identifiable subgroups of patients do not benefit at all, or for whom the risks outweigh the benefits. Conversely, if the overall trial does not show statistically significant evidence of benefit, physicians may ask whether the benefit outweighs the risk in some identifiable subgroups.

Of particular interest are tests of subgroup by treatment interactions to investigate whether the study has shown evidence of differential subgroup effects. Readers familiar with general principals of experimental design are likely to view this type of question as naturally leading to a generalized randomized block design with the subgroups constituting blocks. In clinical trials, however, recruitment is often so difficult that designers of trials usually calculate the total sample size and then randomize within strata without selecting the size of the strata.

Many papers in the statistical and medical literature point to the statistical difficulty in interpreting analyses of subgroups. Small subgroups are associated with highly variable results, and many physicians overinterpret surprisingly large or surprisingly small observed effects. Statistical caveats abound. Looking at many subgroups without proper adjustment for multiplicity leads to a high probability of false positive results. On the other hand, because some subgroups are very small, failure to find an effect in a specific subgroup may simply reflect low statistical power.

In thinking about analysis of subgroups, three dualities are useful. First, and most important, is the distinction between baseline and improper subgroups.

Baseline subgroups are those defined by baseline parameters. If the analysis plan appropriately adjusts for multiplicity, then inference from such subgroups is unbiased. Improper subgroups are those defined by postbaseline variables. Typical of such analysis is a comparison of compliers to noncompliers, or responders to nonresponders. These comparisons, while appealing to many clinicians, are difficult to interpret because these types of subgroups may have been influenced by the treatment itself.

A second duality comes from the difference between prespecified subgroups and those defined post hoc. Betting on a horse after the race (post hoc subgroups) is easier than betting on one before the race begins (prespecified).

The third duality comes from subgroups defined by randomization strata and other subgroups. Readers of this book will understand that the fact of randomization by strata leads to the requirement to analyze in a way that reflects the randomization; it does not force the investigators to create stratum-specific hypotheses. Conversely, strata that are not the basis of randomization should not be precluded from analysis.

See Furberg and Byington (1983) and Yusuf et al. (1991) for papers that describe for clinicians the purposes and pitfalls of subgroup analysis.

5.6.3 Large, Simple Trials

Some randomized clinical trials have many thousands of participants but collect very little information on each. This type of trial is especially useful for studies of prevention of disease when the intervention is both simple and long term or else very short term. For example, studies of pediatric vaccines may include tens of thousands of infants. Data collection is limited to a few demographic variables, immediate adverse reactions, serious adverse reactions experienced in the 40- to 60-day period after the infant's immunization, and the occurrence of the disease the vaccine was designed to prevent. Several trials of simple interventions in cardiology, for example, use of aspirin to prevent heart attack, have used this large, simple model. In designing such a trial, the investigators must ensure very low loss to follow-up rates. See Yusuf, Collins, and Peto (1983) for a description of this type of trial.

5.6.4 Equivalence and Noninferiority Trials

The trials discussed thus far as so-called *superiority* trials where investigators hypothesize that the experimental intervention is superior to the control. Some trials are designed to show that a new product is equivalent to one already marketed. Other trials are designed to show that a new product is noninferior to one already available. *Equivalent* operationally means *not unacceptably different from*, and *noninferior* operationally means *not unacceptably worse than*. Such trials are very challenging because if two interventions show very similar effect sizes, it is impossible to know whether the control

intervention did anything at all. Perhaps the control was ineffective and the similarity between the new therapy and the control simply reflects that neither product was better than no treatment at all. Another serious problem with equivalence and noninferiority trials is that the sloppier the execution, the more likely the two products are to appear similar to each other. In a superiority trial, operational carelessness invites noise, which tends to attenuate the treatment effect rendering the test of superiority conservative. In an equivalence or noninferiority trial, on the other hand, lack of operational rigor, by adding noise, renders the treatments more similar than they otherwise might be. For these and other reasons, statisticians involved in the design of such trials must be cautious and must be willing to spend a lot of time with investigators discussing the pitfalls of these designs. Fleming (2008) discusses statistical and medical issues in noninferiority trials.

REFERENCES

- Cutler, S.J. and F. Ederer (1958). Maximum utilization of the life table method in analyzing survival. *J. Chronic Dis.*, **8**, 699–712.
- Ellenberg, S.S., T.R. Fleming, and D.L. DeMets (2002). *Data Monitoring Committees in Clinical Trials. A Practical Perspective*. New York: Wiley.
- Fleming, T.R. (2008). Current issues in non-inferiority trials. *Stat. Med.*, **27**, 317–332.
- Freedman, L.S. (1982). Tables of the number of patients required in clinical trials using the logrank test. *Stat. Med.*, **1**, 121–129.
- Friedman, L.M., C.D. Furberg, and D.L. DeMets (1998). *Fundamentals of Clinical Trials* (3rd ed.). New York: Springer.
- Furberg, C.D. and R.P. Byington (1983). What do subgroup analyses reveal about differential response to beta-blocker therapy? The Beta-Blocker Heart Attack Trial experience. *Circulation*, **67**, 98–101.
- Gould, S.J. (1981). *The Mismeasure of Man*. New York: WW Norton and Company.
- Greenwood, M. (1926). The natural duration of cancer. Reports on public health and medical subjects. Her Majesty's Stationery Office, London, **33**, 1–26.
- Halpern, J. and B.J. Brown (1993). A computer program for designing clinical trials with arbitrary survival curves and group sequential testing. *Control. Clin. Trials*, **14**, 109–122.
- Herson, J. (2009). *Data and Safety Monitoring Committees in Clinical Trials*. Boca Raton, FL: Chapman and Hall.
- Jennison, C. and B.W. Turnbull (2002). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman and Hall.
- Kalbfleisch, J.D. and R.L. Prentice (2002). *The Statistical Analysis of Failure Time Data* (2nd ed.). New York: Wiley.
- Kaplan, E.L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, **53**, 457–481.
- Lachin, J. and M. Foulkes (1986). Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics*, **42**, 507–519.

- Lagakos, S.W., L.Y. Lim, and J.M. Robins (1990). Adjusting for early treatment termination in comparative clinical trials. *Stat. Med.*, **9**, 1417–1424.
- Lakatos, E. (1986). Sample size determination in clinical trials with time-dependent rates of losses and noncompliance. *Control. Clin. Trials*, **7**, 189–199.
- Lakatos, E. (1988). Sample size based on the log-rank statistics in complex clinical trials. *Biometrics*, **44**, 229–241.
- Lakatos, E. and K.K.G. Lan (1992). A comparison of sample size methods for the log rank statistics. *Stat. Med.*, **11**, 179–191.
- Lan, K.K. and J. Witter (1986). Rank tests for survival analysis: A comparison by analogy with games. *Biometrics*, **41**, 1063–1069.
- Lee, Y.J., J.H. Ellenberg, D.G. Hirtz, and K.B. Nolan (1991). Analysis of clinical trials by treatment actually received: Is it really an option? *Stat. Med.*, **10**, 1595–1605.
- Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemoter. Report*, **50**, 163–170.
- Mantel, N. and W. Haenszel (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.*, **22**, 719–748.
- Miller, R.G., Jr (1981). *Survival Analysis*. New York: Wiley.
- Peduzzi, P., J. Wittes, K. Detre, and T. Holford (1993). Analysis as-randomized and the problem of non-adherence: An example from the Veterans Affairs Randomized Trial of Coronary Artery Bypass Surgery. *Stat. Med.*, **12**, 1185–1195.
- Peto, R. and J. Peto (1972). Asymptotically efficient rank invariant test procedures. *J. R. Stat. Soc. A.*, **135**, 185–198.
- Piantadosi, S. (2005). *Clinical Trials: A Methodological Perspective*. New York: Wiley.
- Pocock, S. (1983). *Clinical Trials: A Practical Approach*. New York: Wiley.
- Proschan, M.A., K.K.G. Lan, and J.T. Wittes (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach*. New York: Springer.
- Schoenfeld, D. (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics*, **39**, 499–503.
- Shih, J. (1995). Sample size based on the log-rank statistic in complex clinical trials. *Control. Clin. Trials*, **16**, 395–407.
- Snappin, S. and Q. Jiang (2011). Analysis of multiple endpoints in clinical trials: It's time for the designations of primary, secondary and tertiary to go. *Pharm. Stat.*, **10**, 1–2. Published online in Wiley Interscience <<http://www.interscience.wiley.com>; DOI: 10.1002/pst.402>.
- Yusuf, S., R. Collins, and R. Peto (1983). Why do we need some large, simple randomized trials? *Stat. Med.*, **3**, 409–420.
- Yusuf, S., J. Wittes, J. Probstfield, and H. Tyroler (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA*, **266**, 93–98.
- Zelen, M. (1993). Theory and practice of clinical trials. In: *Cancer Medicine* (3rd ed.). J.F. Holland, E. Frei III, R.C. Bast, D.W. Kufe, D. Morton, and R.R. Weichselbaum (eds.). Philadelphia: Lea and Febiger, pp. 299–313.

CHAPTER 6

Monitoring Randomized Clinical Trials

Eric S. Leifer and Nancy L. Geller

6.1 INTRODUCTION

A randomized clinical trial is a medical experiment conducted in human subjects who are randomized to receive one of two or more treatments, one of which may be a placebo (see Chapter 5). As such, it has to protect the safety of the trial's subjects while trying to obtain information that advances medical research for societal benefit. Meeting those two responsibilities requires monitoring trials in a careful manner. In this chapter, we discuss monitoring a randomized clinical trial in which subjects are randomized between two treatment arms, one of which may be a placebo. The Cardiac Arrhythmia Suppression Trial (CAST) provides an example of the importance of trial monitoring (CAST Investigators 1989). The trial's primary hypothesis was that pharmacologically suppressing arrhythmias (i.e., irregular heartbeats) in subjects who had a prior heart attack would decrease the number of sudden deaths and cardiac arrests. To test this hypothesis, it was planned to randomize 4400 subjects with a prior heart attack to receive either an antiarrhythmia drug or a placebo (Friedman et al. 1993). Subjects who were randomly assigned to receive an antiarrhythmic received one of encainide, flecainide, or moricizine. The primary end point for the trial was the comparison of all antiarrhythmia drugs versus placebo, and was tested using the logrank statistic. We will discuss the logrank statistic in greater detail in Section 6.6, including that the logrank statistic is approximately normally distributed for sufficiently large patient sample sizes.

Design and Analysis of Experiments: Special Designs and Applications, First Edition. Edited by Klaus Hinkelmann.

© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

As with most contemporary clinical trials, there was a data safety monitoring board (DSMB) appointed to monitor the trial's data. This DSMB was comprised of experts in cardiovascular medicine, clinical medicine, biostatistics, and ethics. The DSMB reported to the National Heart, Lung, and Blood Institute (NHLBI) director who had ultimate authority to conduct the trial (e.g., stop the trial early if circumstances warranted such action) since NHLBI funded the trial. Prior to the trial, there were clinical experts who were so convinced of the benefits of antiarrhythmics that they felt it was unethical to randomize subjects to a placebo (Moore 1995, p. 217). However, soon after the trial began, a difference between the two study arms emerged. The DSMB reviewed the data in September 1988 blinded to study arm (antiarrhythmic versus placebo), so the study arms were referred to as "drug X" and "drug Y." At the time of this review, about one-fourth of the planned 4400 subjects had been entered into CAST, and while only 3 out of 576 (0.5%) drug X subjects had a sudden death or cardiac arrest, 19 out of 571 (3.3%) drug Y subjects had such an event. This corresponded to a logrank statistic z -value of -3.43 with a one-sided p -value of 0.0003. While the event proportion for drug Y was over six times greater than that for drug X, the 22 events corresponded to a very small proportion of the 425 total expected events in the trial. Consequently, the DSMB decided that regardless of whether drug X was the antiarrhythmic or placebo arm, they would not recommend stopping the trial since the small proportion of events would not compellingly establish to the medical community whether or not antiarrhythmics were beneficial or harmful.

However, the absolute difference in proportions between drugs X and Y continued to grow wider. At the next DSMB meeting in April 1989, the DSMB reviewed the data in an unblinded fashion, that is, dispensing with the drug X and Y labels and using the actual drugs assigned. At that point, almost 40% of the planned 4400 subjects had been randomized in the trial. The DSMB found that the arrhythmic drugs arm was experiencing significantly more events corresponding to a logrank statistic z -value of -3.22 and one-sided p -value of 0.00064. In particular, two of the antiarrhythmics, encainide and flecainide, demonstrated substantial enough risk to be discontinued immediately. The third antiarrhythmic, moricizine, was subsequently determined to be harmful in the follow-up trial, CAST II (CAST II Investigators 1992).

The CAST experience demonstrates the importance of trial monitoring, in particular to guard against largely unanticipated clinical outcomes, that is, antiarrhythmic drugs were more harmful than a placebo. It also shows several fundamental aspects of trial monitoring. First, periodic or sequential review of batches or groups of patient outcomes. Indeed, the statistical aspects of clinical trial monitoring we discuss in this chapter are generally known as *group sequential methods*. Second, the use of a z -statistic at each interim analysis (e.g., the logrank test for the CAST primary outcome) to test for outcome differences by treatment arm. Finally, the synthesis of statistical analyses with clinical and ethical exigencies to determine, at an interim analysis, whether a clinical trial should be stopped or allowed to continue.

6.2 NORMALLY DISTRIBUTED OUTCOMES

To motivate the development in the remainder of this chapter, it will be helpful to consider the following simplified example. Suppose we are conducting a two-armed clinical trial in subjects with high systolic blood pressure in which the subjects are randomized to receive an experimental drug or a control drug. Suppose that N subjects are randomized to each treatment. The outcome is systolic blood pressure change from baseline to 3 months postrandomization, measured in millimeters of mercury (mmHg). Let $X_{T1}, X_{T2}, \dots, X_{TN}$ be the independent systolic blood pressure changes in the treatment arm which we assume to have a $N(\mu_T, \sigma^2)$ distribution where σ^2 is known, but μ_T is not. Let $X_{C1}, X_{C2}, \dots, X_{CN}$ be the independent changes in the control arm which we assume to have a $N(\mu_C, \sigma^2)$ distribution for unknown μ_C . Let $\delta = \mu_C - \mu_T$ be the treatment effect. We want to test the null hypothesis of no treatment effect, that is, $H_0: \delta = 0$. For $n = 1, 2, \dots, N$, let $S_n = \sum_{i=1}^n (X_{Ci} - X_{Ti})$, $v_n = \text{var}(S_n) = n\text{var}(X_{C1} - X_{T1}) = 2n\sigma^2$, and $Z_n = S_n / \sqrt{v_n}$. By definition, Z_n is normally distributed with mean $\sqrt{n}\delta / \sqrt{2}\sigma$ and variance 1. In this example, larger systolic blood pressure reductions are better. Thus, if we were to collect all $2N$ outcomes before observing the data, we would declare the experimental drug to be better if Z_N were large. Using a two-sided 0.05 significance level, we would reject the null hypothesis in favor of the superiority of the experimental drug if $Z_N > 1.96 = z_{0.025}$, where z_α is the upper α^{th} percentile of the standard normal distribution. Now, suppose instead of waiting until the end of the trial to look at the data, we planned in advance of the trial to test the null hypothesis after observing the $2n < 2N$ outcomes $X_{T1}, X_{T2}, \dots, X_{Tn}, X_{C1}, X_{C2}, \dots, X_{Cn}$. If Z_n were sufficiently large, we would reject H_0 in favor of the superiority of the experimental drug. However, if Z_n were not sufficiently large, we would continue the trial to observe the additional $2N - 2n$ observations $X_{T_{n+1}}, X_{T_{n+2}}, \dots, X_{TN}, X_{C_{n+1}}, X_{C_{n+2}}, \dots, X_{CN}$, and retest the null hypothesis at that time, rejecting H_0 if Z_N were sufficiently large. For the sake of simplicity, we assume that we would not stop early if $-Z_n$ were large, indicating inferiority of the experimental drug, although as we saw in the CAST example, this is something we would also need to consider in an actual trial.

By saying that we would reject H_0 if either Z_n or Z_N were sufficiently large, we mean that prior to the trial, we need to choose critical values c_1 and c_2 such that we reject H_0 if either: (1) $Z_n \geq c_1$, or (2) if $Z_n < c_1$ and $Z_N \geq c_2$. The values c_1 and c_2 need to be chosen to control the overall two-sided type I error at the 0.05 level. In mathematical terms, this means that under the null hypothesis,

$$\Pr\{Z_n \geq c_1\} + \Pr\{Z_n < c_1, Z_N \geq c_2\} = 0.025. \quad (6.1)$$

From Equation (6.1), we see that c_1 must be greater than $z_{0.025} = 1.96$, since $\Pr\{Z_n < c_1, Z_N \geq c_2\}$ is positive. However, to determine which values of c_1 and c_2 would preserve the equality in Equation (6.1), we use the following properties of the joint distribution of (Z_n, Z_N) :

- Z1': Z_n, Z_N have a bivariate normal distribution.
- Z2': $E(Z_n) = \sqrt{n}(\delta/\sqrt{2}\sigma)$.
- Z3': $\text{cov}(Z_n, Z_N) = \sqrt{n/N}$.
- Z4': $\sqrt{n/N} \cdot Z_n$ and $Z_N - \sqrt{n/N} \cdot Z_n$ are independent.

Properties Z1'–Z4' are straightforward to derive from the independence and normality assumptions about the X_{Ci}, X_{Ti} . Properties Z1'–Z4' also make straightforward the calculation of the monitoring boundary critical values c_1 and c_2 . Specifically, under H_0 and letting $\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard normal distribution and density functions, respectively,

$$\begin{aligned} 0.025 &= \Pr\{Z_n \geq c_1\} + \Pr\{Z_n < c_1, Z_N \geq c_2\} \\ &= \Phi(-c_1) + \Pr\left\{\sqrt{\frac{n}{N}} \cdot Z_n < \sqrt{\frac{n}{N}} \cdot c_1, Z_N - \sqrt{\frac{n}{N}} \cdot Z_n \geq c_2 - \sqrt{\frac{n}{N}} \cdot c_1\right\} \\ &= \Phi(-c_1) + \int_{-\infty}^{\sqrt{\frac{n}{N}} \cdot c_1} \int_{c_2-x}^{\infty} \left(1 - \frac{n}{N}\right)^{-1/2} \phi\left(\frac{y}{\sqrt{1 - \frac{n}{N}}}\right) dy \left(\frac{n}{N}\right)^{-1/2} \phi\left(\frac{x}{\sqrt{\frac{n}{N}}}\right) dx \\ &= \Phi(-c_1) + \int_{-\infty}^{\sqrt{\frac{n}{N}} \cdot c_1} \Phi\left(\frac{x - c_2}{\sqrt{1 - \frac{n}{N}}}\right) \left(\frac{n}{N}\right)^{-1/2} \phi\left(\frac{x}{\sqrt{\frac{n}{N}}}\right) dx. \end{aligned}$$

For fixed $c_1 > 1.96$, the above expression can be solved uniquely for c_2 by numerical integration since the integrand is a strictly decreasing function of c_2 . In other words, once we decide where to set the interim analysis critical value c_1 , the critical value c_2 at the end of the trial is determined.

Example 6.1. Suppose for the above described systolic blood pressure trial, we plan to randomize N subjects to each of the treatment and control arms. We will test the null hypothesis of no treatment difference at the two-sided 0.05 significance level (“alpha”). We also plan to conduct an interim analysis after half of the subjects in each arm have been followed for the end point, and we will spend 0.01 of the 0.05 alpha at the interim analysis. This means we calculate c_1 and c_2 so that $\Pr\{Z_n \geq c_1\} = 0.01/2 = 0.005$ and $\Pr\{Z_n < c_1, Z_N \geq c_2\} = 0.05/2 = 0.025$, hence $\Pr\{Z_n < c_1, Z_N \geq c_2\} = (0.05 - 0.01)/2 = 0.02$. Using the calculations shown above, it turns out that $c_1 = 2.58$ and $c_2 = 2.00$. There are several software packages that can be used to calculate c_1 and c_2 . We used version 2.1 of the freely available Lan–DeMets software, which can be downloaded from <http://www.biostat.wisc.edu/landemets>. Chapter 14 of Proschan, Lan, and Wittes (2006) gives a good discussion of this software.

6.3 BROWNIAN MOTION PROPERTIES

In this section, we state properties Z1'–Z4' more generally and make the connection to Brownian motion. These properties apply in a wide variety of clinical trial settings that we will indicate. Suppose we have random variables D_1, D_2, \dots, D_N and for $n = 1, 2, \dots, N$, let $S_n = \sum_{j=1}^n D_j$ and $v_n = \text{var}(S_n)$. Also, let $Z_n = S_n / \sqrt{v_n}$ denote the corresponding z -statistic although we have made no assumptions about D_1, D_2, \dots, D_N that imply Z_n is normally distributed. We call

$$t = \frac{v_n}{v_N} = \frac{\text{var}(S_n)}{\text{var}(S_N)} \quad (6.2)$$

the *information time* corresponding to S_n . For the systolic blood pressure example from Section 6.2, we form the random variables $D_1 = X_{C1} - X_{T1}$, $D_2 = X_{C2} - X_{T2}, \dots, D_N = X_{CN} - X_{TN}$. Recall that the $X_{C1}, X_{T1}, X_{C2}, X_{T2}, \dots, X_{CN}, X_{TN}$ are the observed systolic blood pressure changes of the control and treatment subjects. While pairing control subjects with treatment subjects to form the differences D_1, D_2, \dots, D_N is something that would not be typically done in an actual clinical trial, it will be useful for making the connection to Brownian motion below. Note that, D_1, D_2, \dots, D_N are independent with variance $2\sigma^2$, so $\text{var}(S_n)/\text{var}(S_N) = 2n\sigma^2/2N\sigma^2 = n/N$. Thus, the information time corresponding to S_n is the proportion of observed systolic blood pressures to the total number of systolic blood pressures to be observed.

For the information times $0 = t_0 < t_1 = n_1/N < t_2 = n_2/N < \dots < t_k = n_k/N < = 1$, let $Z(t_j)$ denote the z -statistic at information time $t_j = n_j/N$. Also, denote by θ the treatment effect parameter. In the two-armed setting (treatment versus control), we think of θ as representing a clinically meaningful difference between the treatment and control groups. Under the null hypothesis of no treatment difference, $\theta = 0$. Under the alternative hypothesis $\theta \neq 0$, θ is parametrized so that $E\{Z(1)\} = \theta$.

Suppose for the systolic blood pressure example from Section 6.2, a $\delta = 4$ mm Hg treatment effect is of clinical interest. Suppose the X_{Cj} and X_{Tj} have common standard deviation $\sigma = 10$ mm Hg, and the $D_j = X_{Cj} - X_{Tj}$ have common mean $\delta = 5$ mm Hg. Then, randomizing $N = 132$ subjects to each arm: gives

$$\theta = E\{Z(1)\} = E\left\{ \frac{\sum_{j=1}^N D_j}{\sqrt{2N\sigma^2}} \right\} = \frac{\delta}{\sqrt{2\sigma^2/N}} = \frac{4}{\sqrt{2 \cdot 10^2/132}} = 3.25.$$

The generalization of properties Z1'–Z4' that we present below assumes the treatment effect δ is a *local alternative*. A rigorous treatment of local alternatives is beyond the scope of this chapter, but for our discussion, it means that the treatment effect $\delta = \delta_N$ depends on N , such that $\theta = \delta_N / \sqrt{2\sigma^2/N}$

approaches a constant as N goes to infinity. The reason for this assumption is because the z -statistic $Z_N = \sum_{j=1}^N D_j / \sqrt{2N\sigma^2}$ has mean $\theta / \sqrt{2\sigma^2/N}$, which we need to approach a constant to obtain asymptotic (in N) results about Z_N .

Under local alternatives δ_N (which could be identically zero in the case of the null hypothesis), we have the following generalization of Z1'–Z4', which hold in a wide variety of contexts, including continuous outcomes (e.g., systolic blood pressure measurements), dichotomous outcomes (e.g., success or failure of a treatment), and time-to-event (e.g., survival) outcomes:

Asymptotically, in the number of observations N ,

- Z1: $Z(t_1), Z(t_2), \dots, Z(t_k)$ have a multivariate normal distribution.
- Z2: $E\{Z(t)\} = \sqrt{t}\theta$ for all t .
- Z3: $\text{cov}(Z(t_i), Z(t_j)) = (t_i/t_j)^{1/2}$ for $t_i \leq t_j$.
- Z4: (independent increments property): For $t_i < t_j \leq t_l < t_m$, $\sqrt{t_j}Z(t_j) - \sqrt{t_i}Z(t_i)$ is independent of $\sqrt{t_m}Z(t_m) - \sqrt{t_l}Z(t_l)$.

Theorems that establish Z1–Z4 in quite general settings can be found in Jennison and Turnbull (1997) and Scharfstein, Tsiatis, and Robins (1997). However, the basic idea of such theorems is to use the univariate central limit theorem applied to the sum S_{n_j} of the observations at each time $t_j = n_j/N$. We have already seen this explicitly in Section 6.2, where the standard normally distributed $Z_n = S_n / \sqrt{v_n}$ corresponded to the normalized sum of the independent and identically distributed normal random variables $D_1 = X_{C1} - X_{T1}$, $D_2 = X_{C2} - X_{T2}, \dots, D_n = X_{Cn} - X_{Tn}$. Note that by the Lindeberg central limit theorem (theorem 27.2, Billingsley 1986), it was not necessary for D_1, D_2, \dots, D_n to be normally, or even identically, distributed for Z_n to have an approximately standard normal distribution. The following result from Proschan, Lan, and Wittes (2006, result 2.1) highlights the role of univariate asymptotic normality in establishing Z1–Z4 when $\theta = 0$:

Result 6.1. Let S_N be a sum of independent (not necessarily identically distributed) random variables with mean 0, and $n_j \rightarrow \infty$ and $N \rightarrow \infty$ such that $v_{n_j}/v_N \rightarrow t_j, j = 1, \dots, k$. Then asymptotically, properties Z1–Z4 for $\theta = 0$ hold if and only if $Z(t_j)$ is asymptotically standard normal.

In several situations, such as in the evaluation of conditional power (Section 6.8), it is useful to recast Z1–Z4 in a Brownian motion formulation as in Lan and Wittes (1988). Lan and Wittes define

$$B(t) = \sqrt{t}Z(t) \tag{6.3}$$

to be the *B-value* at information t . This is because properties Z1–Z4, recast as B1–B4 below, characterize Brownian motion with drift parameter θ on the unit interval $0 \leq t \leq 1$:

- B1: $B(t_1), B(t_2), \dots, B(t_k)$ have a multivariate normal distribution.
- B2: $E\{B(t)\} = \theta t$ for all t .
- B3: $\text{cov}(B(t_i), B(t_j)) = t_i$ for $t_i \leq t_j$.
- B4: (independent increments property): For $t_i < t_j \leq t_l < t_m$, $B(t_j) - B(t_i)$ is independent of $B(t_m) - B(t_l)$.

6.4 BRIEF HISTORICAL OVERVIEW OF GROUP SEQUENTIAL METHODS

Jennison and Turnbull (2000, chapter 1) provide an excellent historical overview of sequential analysis, group sequential methods, and clinical trial monitoring. As they discuss, modern clinical trials monitoring has strong roots in the classified World War II work of Abraham Wald (Wald 1947) in the United States and George Barnard (Barnard 1946) in Great Britain. Both Wald and Barnard worked in advisory groups to their respective governments. Much of Wald's work centered on developing efficient statistical designs for testing munitions. Efficiency was defined in terms of minimizing the sample size of the munitions that were being tested. One of Wald's most famous conjectures, which he proved with Wolfowitz (Wald and Wolfowitz 1948), is the sequential probability ratio test (SPRT). In its simplest form, it considers testing the simple null hypotheses $H_0: \theta = \theta_0$ against the simple alternative hypothesis $H_1: \theta = \theta_1$, where $\theta_1 > \theta_0$. It sequentially considers independently and identically distributed observations x_1, x_2, \dots one-at-a-time and calculates after each observation the likelihood ratio

$$L_n = \frac{f_{1n}(x_1, x_2, \dots, x_n)}{f_{0n}(x_1, x_2, \dots, x_n)}. \quad (6.4)$$

In Equation (6.4), f_{1n} is the assumed joint density of (x_1, x_2, \dots, x_n) under H_1 , while f_{0n} is the assumed joint density under H_0 . For fixed constants $a < b$, the SPRT dictates that if $L_n \leq a$, no further observations are taken, and H_0 is accepted. Similarly, if $L_n \geq b$, no further observations are taken, and H_1 is accepted. Finally, if $a < L_n < b$, an $(n+1)^{\text{st}}$ observation x_{n+1} is taken, L_{n+1} is computed, and the same tests are repeated. The interval $a < L_n < b$ is called the *continuation region*, since the trial continues if L_n is between a and b .

Let α be the type I error of this SPRT (i.e., corresponding to boundary values a and b) and β be its type II error. Consider the class of all such statistical procedures that observe the data one observation at a time as the SPRT does, making a decision after each observation to accept H_0 or accept H_1 or continue sampling. If we restrict the class of such procedures to those that also have type I error α and type II error β , then Wald and Wolfowitz proved the SPRT's remarkable property that among all such statistical procedures, the SPRT uses on average the smallest number of observations both when H_0 is true and when H_1 is true. That is, among all such procedures, the SPRT has the

smallest expected sample size under both H_0 and H_1 . Moreover, if α and β are prespecified (e.g., $\alpha = 0.025$ and $\beta = 0.1$), then the boundary values $a^* = \beta/(1 - \alpha)$ and $b^* = (1 - \beta)/\alpha$ give an SPRT with approximate type I and II errors α and β , respectively.

The idea of minimizing the expected sample size under H_0 and H_1 is an important one whether the goal is to minimize the number of munitions tested or minimize the number of subjects enrolled to a clinical trial. However, Wald's original SPRT has limitations that Wald himself recognized. First, the SPRT is fully sequential in that after each observation x_n is taken, the SPRT decides to either accept H_0 , accept H_1 , or take another observation. Of course in several situations, it would be impractical to analyze the data after each observation. This is particularly true for large-scale clinical trials, but it is also true for industrial applications for which it would be infeasible to shut down a factory line for inspection after producing each unit. Indeed, Dodge and Romig (1929) developed a two-stage statistical monitoring plan for testing an industrial lot of objects that can be classified as defective (yes/no). The plan specifies a first-stage sample size n_1 and acceptance and rejection values a_1 and r_1 with $a_1 + 1 < r_1$. If the number d_1 of defective objects is less than or equal to a_1 , then the lot is accepted; if $d_1 \geq r_1$, then the lot is rejected. If $a_1 < d_1 < r_1$, then a second batch of n_2 objects is taken with cumulative acceptance and rejection bounds a_2 and r_2 such that $a_2 + 1 = r_2$. If the cumulative number of defectives $d_2 \leq a_2$, then the lot is accepted; otherwise, $d_2 \geq r_2$ and the lot is rejected. Such two-stage sampling was generalized to more than two stages by Bartky (1943) and the Columbia University Research Group (Freeman et al. 1948), which became the basis for the U.S. military standard of acceptance sampling, MIL-STD-105E (1989). Two-stage and three-stage procedures for normally distributed data were developed by Armitage and Schneiderman (1958), Schneiderman (1961), Dunnett (1961), and Roseberry and Gehan (1964).

In addition to the need for group sequential testing, Wald understood the need to truncate the SPRT, that is, limit the number of observations that could be taken. Wald showed (Wald 1947, section A.1) that with probability 1, the SPRT eventually terminates. Nevertheless, he studied the properties of truncating the SPRT at a fixed upper limit of observations (Wald 1947, section 3.8).

Armitage (1954, 1958, 1975) and Bross (1952, 1958) were largely responsible for introducing sequential methods in medical testing. Armitage, McPherson, and Rowe (1969) used recursive numerical integration to study the effect of repeated significance testing on type I error inflation when the fixed sample size critical value is not increased to account for multiple looks. In particular, they obtained the numerical results in Table 6.1, which gives the type I error inflation from equally spaced looks of normally distributed data. Noteworthy about this paper is that similar numerical integration techniques are used in most modern-day group sequential software (Proschan, Lan, and Wittes 2006, section 4.7).

Pocock (1977) provided substantial impetus to group sequential applications to clinical trials. In particular, he provided great detail showing how

Table 6.1 Two-Sided Type I Error from Equally Spaced Looks at Normally Distributed Data Using a 1.96 Critical Value at Each Look with Corresponding Pocock and O'Brien–Fleming Critical Values

Number of Looks	Type I Error	Pocock Critical Value c_P for z -Statistic	O'Brien–Fleming Critical Value c_{OF} for B -Value
1	0.050	1.960	1.960
2	0.083	2.178	1.977
3	0.107	2.289	2.004
4	0.126	2.361	2.024
5	0.142	2.413	2.040
10	0.193	2.555	2.087
20	0.248	2.672	2.126

Armitage, McPherson, and Rowe's results for normal response with known variance could be applied in the group sequential setting to normal responses with unknown variance, exponential responses (foreshadowing the applications to the logrank statistic), and binary responses. He also discussed variations in group sizes (unequally spaced interim looks), stratification of risk factors, and more than two treatment arms. Pocock also determined the constant monitoring boundary for the z -statistic. That is, for a two-armed trial with K equally spaced looks at the data at information times $j = 1/K, 2/K, \dots, K/K = 1$, Pocock proposed a constant critical value $c_P = c_P(\alpha, K)$ for the z -statistic which preserves overall two-sided type I error α according to

$$\Pr\left(\bigcup_{j=1}^K \{|Z(j/K)| > c_P\}\right) = \alpha. \quad (6.5)$$

For example, we see from Table 6.1 that for $\alpha = 0.05$ and $K = 5$, $c_P(0.05, 5) = 2.413$.

The constant monitoring boundary is known as the *Pocock boundary*, although Pocock and most other clinical trialists would not advocate using this boundary for most clinical trials (Pocock 1992). This is because the same level of evidence (i.e., critical value) is used for stopping the trial at an earlier look as at a later look. At the earlier looks when investigators are learning how to implement the protocol and the data are preliminary, it is often preferable to require a higher level of evidence for stopping the trial. Moreover, it is desirable to have the critical value at the final look be close to that of a trial with no interim looks (e.g., 1.96 corresponding to a two-sided 0.05 significance level) so that the trial will not be in the position of declaring nonsignificance despite the final z -statistic substantially exceeding the no-interim-look critical value. Although in such a situation it would be statistically proper to declare nonsignificance, it could confuse those unsophisticated with group sequential methodology.

O'Brien and Fleming (1979) proposed a boundary which handles this criticism of the Pocock boundary by using a constant monitoring value for the

B -value as opposed to the z -statistic used by Pocock. For a two-armed trial with K equally spaced looks at information times $j = 1/K, 2/K, \dots, K/K = 1$, the O'Brien–Fleming constant critical value, which we denote by $c_{OF} = c_{OF}(\alpha, K)$, preserves overall two-sided type I error α according to

$$\Pr\left(\bigcup_{j=1}^K \{|B(j/K)| > c_{OF}\}\right) = \alpha. \quad (6.6)$$

So $B(j/K)$ in Equation (6.6) replaces $Z(j, K)$ in Equation (6.5). For example, we see from Table 6.1 that for $\alpha = 0.05$ and $K = 5$, $c_{OF}(0.05, 5) = 2.040$.

Thus, unlike the constant Pocock boundary c_P for the z -statistic, the O'Brien–Fleming critical value for the z -statistic is $c_{OF}(j/K)^{1/2}$ so it decreases at each succeeding look. Consequently, it addresses the two criticisms we identified about the Pocock boundary by making it more difficult than Pocock's boundary to stop the trial at earlier looks and having a final look critical value that is closer to the no-interim-look critical value. Table 6.2 gives comparisons of Pocock and O'Brien–Fleming boundaries for the z -statistic for two-sided 0.05 significance level trials with 5, respectively, 10, equally spaced looks.

O'Brien and Fleming's boundary corresponds to Wald's truncated SPRT for normally distributed observations with variance 1. To see this, recall from Equation (6.4) that the SPRT is based on the likelihood ratio L_n , where the continuation region is an interval of the form $a < L_n < b$ where a and b are fixed constants. The truncated SPRT fixes an upper bound, say, N , on the number of independent observations X_1, X_2, \dots, X_N that will be taken. In O'Brien and Fleming's group sequential application of Wald's truncated SPRT, we understand each X_i to be the batch of outcomes obtained between the $(i - 1)^{\text{st}}$ and i^{th} looks. Also, X_i is assumed to be normally distributed with mean μ and variance 1. Then the likelihood ratio L_n corresponding to $H_0: \mu = 0$ versus $H_1: \mu = \mu_1$ has the form

Table 6.2 Pocock and O'Brien–Fleming z -Statistic Critical Values for Two-Sided 0.05 Significance Level Trials with 5, Respectively, 10, Equally Spaced Looks

Information Time	Pocock 5-Look	O'Brien–Fleming 5-Look	Pocock 10-Look	O'Brien–Fleming 10-Look
0.1			3.117	6.600
0.2	2.413	4.562	3.117	4.667
0.3			3.117	3.810
0.4	2.413	3.226	3.117	3.300
0.5			3.117	2.951
0.6	2.413	2.634	3.117	2.694
0.7			3.117	2.494
0.8	2.413	2.281	3.117	2.333
0.9			3.117	2.200
1.0	2.413	2.040	3.117	2.087

$$L_n = \frac{\prod_{i=1}^n \exp\{-(x_i - \mu_1)^2/2\}}{\prod_{i=1}^n \exp\{-x_i^2/2\}} = \exp\left\{\mu_1 \sum_{i=1}^n x_i - \frac{n\mu_1^2}{2}\right\}.$$

But then the continuation region $a < L_n < b$ can be equivalently stated in terms of the B -value $B(n/N) = N^{-1/2} \cdot \sum_{i=1}^n x_i$.

As we saw in Table 6.1, more frequent looks at the data result in higher type I error inflation when the critical value is not adjusted upward to account for the multiple looks. Thus, the need to craft monitoring boundaries such as Pocock's or O'Brien and Fleming's to preserve the type I error at a prespecified level. However, as quoted in Proschan, Lan, and Wittes (2006, p. 175), Meier (1975) has remarked, "... it seems hard indeed to accept the notion that I should be influenced in my judgment by how frequently *he* peeked at the data while he was collecting it." In other words, it is possible that the same data at the end of the trial could lead to an accept or reject H_0 decision depending on which monitoring boundary, hence critical value, is used. This view, which has been shared by many others in the literature, including Cornfield (1966) and Berger and Berry (1988), has led to nonfrequentist methods for monitoring that do not consider how often the data are looked at. These include likelihood methods (Anscombe 1963; Cornfield 1966) and Bayesian methods (Arrow, Blackwell, and Girshick 1949; Berger 1985; Berry 1987). Bayesian methods, in particular, have gained some popularity in phase I and II clinical trials, but are not widely used in phase III trials. For a nice description of Bayesian monitoring, see chapter 10 of Proschan, Lan, and Wittes (2006).

6.5 DICHOTOMOUS OUTCOMES

In this section, we consider trial monitoring for a dichotomous end point such as the occurrence (yes/no) of an outcome. An example of this is in the bone marrow transplant setting for hematologic malignancies where relatively short-term outcomes are sometimes of interest. Acute graft versus host disease (aGVHD) can be a very tragic consequence of a bone marrow transplant in which the patient's body rejects the bone marrow transplant that had been given to the patient in the hope of curing the malignancy. This rejection is due to the patient's immune response to the transplant and can result in severe damage to various organ systems of the patient, such as the lungs, gastrointestinal tract, skin, liver, and mucosa (Ferrara, Deeg, and Burakoff 1990).

Suppose for a trial in bone marrow transplant subjects, we randomize N subjects to receive an experimental aGVHD prophylaxis and N subjects to receive a control aGVHD prophylaxis. Suppose our end point is the onset of aGVHD or death within 6 weeks of randomization. Let p_T be the end point proportion in the (experimental) treatment arm and let p_C be the end point proportion in the control arm. Let $X_{T1}, X_{T2}, \dots, X_{TN}$ be the binary outcomes

in the treatment group and $X_{C1}, X_{C2}, \dots, X_{CN}$ be the binary outcomes in the control group.

We want to test the null hypothesis $H_0: p_C - p_T = 0$. To do this, we use the z -test for proportions after n treatment subjects and n control subjects have been followed for the 6-week end point given by

$$Z_n = \frac{\bar{X}_C - \bar{X}_T}{\sqrt{2p(1-p)/n}}. \quad (6.7)$$

In practice, p in Equation (6.7) would be replaced by the pooled estimate $(\bar{X}_C + \bar{X}_T)/2$. We assume the sample sizes at the interim analyses are sufficiently large so that the pooled estimate approximately equals $(p_C + p_T)/2$.

Let $S_n = \sum_{j=1}^n X_{Cj} - \sum_{j=1}^n X_{Tj}$. We assume there are interim analyses planned for information times $0 < t_1 < t_2 < \dots < t_k = 1$ where $t_{n_j} = \text{var}(S_{n_j})/\text{var}(S_{N_j}) = 2n_j(p_C(1-p_C) + p_T(1-p_T))/2N(p_C(1-p_C) + p_T(1-p_T)) = n_j/N$. Thus, the information time for a dichotomous outcome corresponds to the number of subjects observed for the outcome divided by the maximum number of subjects to be enrolled in the trial.

To apply Result 6.1, we establish the asymptotic standard normality of the z -statistic Z_n . To do so, we subtract the mean $E\{Z_n\}$ from Z_n and properly scale to obtain a zero-mean, asymptotically standard normal statistic Z_n^0 . We have

$$E\{Z_n\} = \frac{p_C - p_T}{\sqrt{2p(1-p)/n}}, \quad (6.8)$$

and we define

$$\begin{aligned} Z_n^0 &= \sqrt{\frac{2p(1-p)}{p_C(1-p_C) + p_T(1-p_T)}} \cdot (Z_n - E\{Z_n\}) \\ &= \frac{(\bar{X}_C - p_C) - (\bar{X}_T - p_T)}{\sqrt{(p_C(1-p_C) + p_T(1-p_T))/n}}. \end{aligned}$$

Z_n^0 is the normalized sum of the independent, zero mean random variables $X_{C1} - p_C, \dots, X_{Cn} - p_C, p_T - X_{T1}, \dots, p_T - X_{Tn}$, and has asymptotic standard normal distribution by the Lindeberg central limit theorem (Billingsley 1986, theorem 27.2). Consequently, by Result 6.1, the joint distribution of the zero mean B -values (see Eq. 6.3) corresponding to $(Z^0(t_1), \dots, Z^0(t_k))$ is asymptotically that of Brownian motion. By algebraically manipulating Equation (6.7), we have that the joint distribution of the B -values corresponding to $(Z(t_1), \dots, Z(t_k))$ is asymptotically that of Brownian motion with drift equal to $(p_C - p_T)/\sqrt{2p(1-p)/n}$ where $p_C - p_T$ is a local alternative.

Note that the above discussion assumes that there are equal numbers of observations in the two randomized groups at each interim analysis. Of course, exact equality does not always occur in practice. Nevertheless, the above discussion can be easily modified to the situation in which the number of treat-

ment and control observations, respectively, increase at approximately the same rate.

Example 6.2. Suppose we want to conduct the aGVHD trial described above, randomizing subjects to either the experimental or control aGVHD prophylaxes. We use the end point of aGVHD or death within 6 weeks of randomization and the z -test of proportions with a two-sided alpha level of 0.05. We want to have two interim analyses: the first analysis after one-third of the subjects have been followed for the end point and the second interim analysis after two-thirds of the subjects have been followed for the end point. By the Brownian motion properties for the z -test of proportions, we can use two-sided O'Brien–Fleming boundaries (Eq. 6.6). Since from Table 6.1 $c_{OF} = c_{OF}(0.05, 3) = 2.004$, the respective z -statistic critical values at the two interim analyses and the final look are $c_1 = 2.004/\sqrt{1/3} = 3.471$, $c_2 = 2.004/\sqrt{2/3} = 2.454$, $c_3 = 2.004/\sqrt{3/3} = 2.004$.

6.6 TIME-TO-EVENT OUTCOMES

The CAST trial (Section 6.1) used the logrank statistic to test its primary end point of time to sudden death or cardiac arrest. Since the logrank statistic is widely used in clinical trials with a time-to-event end point, we indicate how the logrank statistic satisfies properties Z1–Z4 from Section 6.3. In our discussion, we assume each subject can have at most one event, such as death, although properties Z1–Z4 can be extended to the case in which a subject can have more than one event. For simplicity, we assume the event of interest is death. Unless otherwise indicated, all of the results in this section assume the null hypothesis of no difference between the treatment and control groups.

For the logrank statistic, only subjects who are known to die during the trial contribute to the trial's information time. If the subject leaves the trial or the trial ends before the subject dies, the subject is a censored observation. Thus, our information about a subject who is censored is incomplete. If by the end of the trial we expect to have M deaths and thus far we have observed m deaths, then the current information time is m/M .

To establish this with more statistical rigor, we follow the development in section 2.1.3 of Proschan, Lan, and Wittes (2006). Suppose we plan to randomize a total of N subjects to each of the treatment and control groups, and we have thus far observed m deaths of an expected M deaths. We assume the m deaths occur at distinct times although the results can be generalized to allow for tied death times. For each death $i = 1, 2, \dots, m$, we can form a two-by-two table, as given in Table 6.3.

Just before the i^{th} death, there are r_{1i} treatment subjects and r_{0i} control subjects who are at risk of dying. We let O_i be the indicator that the i^{th} death occurred in the treatment group. Under the null hypothesis and conditionally

Table 6.3 2-by-2 Table Corresponding to the i^{th} Death

	Dead	Alive	No. at Risk
Treatment	O_i	$r_{1i} - O_i$	r_{1i}
Control	$1 - O_i$	$r_{0i} - (1 - O_i)$	r_{0i}
	1	$r_{1i} + r_{0i} - 1$	$r_{1i} + r_{0i}$

given the at risk numbers (r_{1i}, r_{0i}) , O_i is a Bernoulli random variable which takes the value 1 with probability $E_i \equiv r_{1i}/(r_{1i} + r_{0i})$.

After m deaths have been observed, the logrank statistic S_m is formed by summing up over the m deaths the *observed minus expected* $D_i = O_i - E_i$, so $S_m = \sum_{i=1}^m D_i$. Note that D_i has conditional mean $E\{D_i|r_{1i}, r_{0i}\} = 0$ so D_i has unconditional mean $E\{D_i\} = E[E\{D_i|r_{1i}, r_{0i}\}] = E(0) = 0$, hence S_m has unconditional mean 0. Next, note that D_i has conditional variance $V_i = E\{D_i^2 | r_{1i}, r_{0i}\} = E_i(1 - E_i)$. Using the general variance decomposition formula $\text{var}\{Y\} = E[\text{var}(Y|X)] + \text{var}\{E(Y|X)\}$, we find that D_i has unconditional variance $\text{var}\{D_i\} = E\{\text{var}(D_i|r_{1i}, r_{0i})\} + \text{var}\{E(D_i|r_{1i}, r_{0i})\} = E\{V_i\} + E\{0\} = E\{V_i\}$.

It can be shown (section 2.9.3 of Proschan, Lan, and Wittes 2006) that unconditionally, the D_i are uncorrelated. This may be surprising since the D_i are not independent because the at risk numbers $r_{1,i+1}$ and $r_{0,i+1}$ depend on whether the i^{th} event occurred in the treatment or control groups. Nevertheless, since the D_i are uncorrelated, S_m has unconditional variance $v_m = \text{var}\{\sum_{i=1}^m D_i\} = \sum_{i=1}^m \text{var}\{D_i\} = \sum_{i=1}^m E\{V_i\}$. It can further be shown that $E(V_i) = E\{E_i(1 - E_i)\}$ is approximately $(1/2) \cdot (1 - 1/2) = 1/4$. Hence, v_m is approximately $m/4$. Thus, from Equation (6.2), $Z_m = S_m/\sqrt{v_m}$ corresponds to information time $t_m = v_m/v_M$, which is approximately $(m/4)/(M/4) = m/M$. That is, the information time t_m is approximately the observed proportion m/M of expected deaths.

To obtain properties Z1–Z4 for the normalized logrank statistic $S_m = Z_m = S_m/\sqrt{v_m}$, we cannot use Result 6.1 as we did for the independent continuous and dichotomous outcomes, since the D_i are not independent. However, Z1–Z4 can be proven using martingale central limit theorems (Helland 1982; Tsiatis 1982). These central limit theorems also establish that

$$E\{Z(1)\} = \theta = \sqrt{M/4} \cdot \delta, \quad (6.9)$$

where δ is the log hazard ratio of the control to the experimental treatment. Peto gives a succinct discussion of this fact in the statistical appendix of Yusuf et al. (1985).

Example 6.3. Suppose we want to conduct a 2-year trial in subjects with pancreatic cancer for which we expect 300 deaths, which is the primary end point. Suppose we want to look at the data every 6 months, using a two-sided, 0.05 significance-level monitoring boundary, and we will use the logrank statistic

to test the null hypothesis of no treatment effect. Properties Z1–Z4 allow us to calculate critical values c_1 , c_2 , c_3 , and c_4 , corresponding to the 6-month, 1-year, 18-month, and 2-year looks, so that

$$\Pr\left(\bigcup_{j=1}^4 \{|Z(t_j)| > c_j\}\right) = 0.05. \quad (6.10)$$

In Equation (6.10), t_j is the ratio of observed deaths by the j^{th} look over the expected number of deaths by the end of the trial. In practice, the calculation of the c_j often uses spending function methodology which we discuss in Section 6.9. We note that, as Proschan, Lan, and Wittes (2006, section 2.1.3) point out, the type I error is preserved even if the final number of deaths in the trial turns out to be different from our pretrial guess. This is because Property Z3 says the covariance of the interim look z -statistics only depends on the ratio of the number of observed deaths, that is, $\text{cov}(Z(t_i), Z(t_j)) = (t_i/t_j)^{1/2}$ for $t_i \leq t_j$. For example, if there are 45 deaths by the first look and 120 deaths by the second look, the covariance of $Z(t_1)$ and $Z(t_2)$ is $(45/120)^{1/2}$ and does not depend on the total of number deaths which eventually occur in the trial.

6.7 UNCONDITIONAL POWER

Unconditional power is the pretrial probability of obtaining a statistically significant result corresponding to a prespecified hypothesized treatment effect. Having sufficient unconditional power (typically 80–90%) is vital to a trial's integrity. Indeed, a trial that has insufficient unconditional power and fails to obtain a statistically significant result leaves unanswered the question of whether the experimental treatment is superior to the control. In such an instance, the efforts of the trial's subjects were largely wasted, perhaps at possible risk to their well-being. Moreover, the trial's research and financial resources were also largely wasted.

Unconditional power is typically calculated with respect to a hypothesized treatment effect of clinical interest. As in Section 6.3, we parametrize the treatment effect by θ corresponding to a local alternative so that $E\{Z(1)\} = \theta$ and $\text{var}\{Z(1)\} = 1$. For a trial with no interim monitoring and with power $1 - \beta$ corresponding to the two-sided significance level α , we have the well-known equality

$$\theta = z_\beta + z_{\alpha/2}. \quad (6.11)$$

where z_p is the upper p^{th} percentile of the standard normal distribution. Equation (6.11) makes sample size calculations straightforward.

Example 6.4. Suppose for the systolic blood pressure example from Section 6.2, a $\delta = 4$ mm Hg difference in systolic blood pressure is of clinical interest. Assume the change-from-baseline systolic blood pressure measurements X_{Ti}

and X_{Ci} have standard deviation $\sigma = 10 \text{ mm Hg}$. If N subjects are randomized to each arm, then the difference-in-means z -statistic has drift parameter

$$\theta = E\{Z(1)\} = E\left\{\frac{\bar{X}_C - \bar{X}_T}{\sqrt{2\sigma^2/N}}\right\} = \sqrt{N} \frac{\delta}{\sqrt{2\sigma^2}} = \sqrt{N} \frac{4}{\sqrt{200}}. \quad (6.12)$$

From Equation (6.11), if the trial is being conducted using a two-sided 0.05 significance level with 90% power, θ satisfies

$$\theta = z_{1-0.90} + z_{0.05/2} = 1.28 + 1.96 = 3.24. \quad (6.13)$$

Thus, to obtain the desired per-arm sample size N , we put together Equations (6.12) and (6.13) and solve for N in

$$\sqrt{N} \cdot \frac{4}{\sqrt{200}} = \theta = 3.24. \quad (6.14)$$

Hence, $N = 132$ subjects per arm are needed.

Example 6.5. Suppose for the aGVHD example from Section 6.5, we wish to have 85% power for a two-sided 0.05 significance level corresponding to a hypothesized 33% reduction in the control group's event rate, which we assume to be $p_C = 0.30$. This corresponds to a $p_T = 0.20$ event rate in the treatment group. From Equation (6.11), the treatment effect parameter θ must satisfy

$$\theta = z_{1-0.85} + z_{0.05/2} = 1.04 + 1.96 = 3.00. \quad (6.15)$$

We recall from Section 6.5 that the z -test for proportions has the form

$$Z_N = \frac{\bar{X}_C - \bar{X}_T}{\sqrt{2(p(1-p))/N}}.$$

To obtain the desired per arm sample size N , we use

$$\theta = 3.00 = E\{Z(1)\} = E\{Z_N\} = \frac{p_C - p_T}{\sqrt{2(p(1-p))/N}} = \frac{0.3 - 0.2}{\sqrt{2(0.25 \cdot 0.75)/N}}, \quad (6.16)$$

so $N = 338$ per arm. Note that by taking into account p_T and p_C in the alternative hypothesis variance of $\bar{X}_C - \bar{X}_T$, equation (11.10) in Proschan, Lan, and Wittes (2006) gives a slightly more correct sample size formula for dichotomous outcomes.

Example 6.6. Suppose we are conducting a two-armed trial of an experimental treatment versus a control for heart failure, and we have as our end point event the time to heart failure hospitalization or death, whichever comes first. We will analyze the end point using the logrank statistic. Suppose the trial is being conducted using a two-sided 0.05 significance level with 90% power corresponding to a hypothesized control to treatment group hazard ratio of 1.20. When the control group's end point rate is low, this hazard ratio approximately corresponds to a 20% increase in the end point risk in the control group relative to the treatment group (Marubini and Valsecchi 1995, appendix A.9.2). From Equations (6.9) and (6.11), to obtain our desired power, we need to determine the total number M events so that

$$\sqrt{M/4} \cdot \log(1.20) = z_{1-0.90} + z_{0.05/2} = 1.28 + 1.96 = 3.24.$$

Thus $M = 1264$ events. If we think that approximately 40% of the subjects will have an event during the trial, then we need to recruit approximately $1264/0.40 = 3160$ subjects to the trial.

6.8 CONDITIONAL POWER

At an interim analysis of a clinical trial, the question may arise, “Given the data we have observed thus far, what is the probability of obtaining a statistically significant result for experimental treatment benefit at the end of the trial?” This probability is called *conditional power* since, unlike unconditional power, it is calculated using data from the trial. Often, this question is asked when the experimental treatment appears from the interim data to be little better than the control treatment and perhaps inferior to it. In such an instance, there is concern about randomizing additional subjects, as well as committing other resources to a trial, which may have little chance of showing experimental treatment benefit.

An example of this occurred in the Cardiac Arrhythmia Suppression Trial II (CAST II) (CAST II Investigators 1992), which was the trial started after the original CAST trial was stopped. CAST II tested the same primary hypothesis as did CAST, that is, pharmacologically suppressing arrhythmias in subjects who had a prior heart attack would decrease sudden deaths and cardiac arrests. As in CAST, CAST II randomized eligible subjects to receive either an antiarrhythmic drug or a placebo. However, while CAST subjects who were randomized to an antiarrhythmic received one of flecainide, encainide, or moricizine, CAST II subjects who were randomized to an antiarrhythmic only received moricizine. This was because CAST determined that flecainide and encainide posed greater harm than a placebo.

To test its hypothesis, CAST II originally proposed to enroll 2200 subjects between April 19, 1989, and January 1, 1992, with follow-up scheduled to continue until April 1, 1994. However, at a July 31, 1991, DSMB meeting, it was

found that 42 of 581 (7.2%) moricizine subjects had a primary event, while 32 of 542 (5.9%) placebo subjects had a primary event. This corresponded to a z -value of -1.05 for the logrank statistic (Friedman et al. 1993). Since CAST II chose its sample size based on 200 expected events, the $32 + 42 = 74$ events corresponded to information time $74/200 = 0.37$ (Section 6.6). Given those data, it was determined that there was a 0.08 probability of demonstrating a significant benefit for moricizine, that is, conditional power was 0.08. The 0.08 was calculated assuming the true moricizine benefit was the 30% reduction in the primary end point assumed for the pretrial, unconditional power calculation, and that the observed -1.05 z -value in favor of the placebo occurred for random reasons.

As we will see below, along with the observed interim data and the corresponding information time, the conditional power calculation depends on what the true treatment effect is assumed to be. Typically, conditional power is calculated under a range of hypothesized treatment effects, with the most optimistic treatment effect usually the one that was used for the pretrial unconditional power calculation. Partly due to the 0.08 conditional power, the DSMB decided to stop CAST II at their July 1991 meeting.

The calculation of conditional power is greatly facilitated by Properties B1–B4 from Section 6.3. Suppose a trial has observed the test statistic $B(t) = \sqrt{t}Z(t) = b$ at information time $t < 1$. Also, suppose the final (i.e., information time = 1) z -statistic critical value for declaring statistical superiority of the experimental treatment is z_f . Recall that by definition, $Z(1) = B(1)$. The conditional power assuming the treatment effect θ is the conditional probability (under θ) that $B(1) > z_f$ given $B(t) = b$. That is, conditional power is

$$\text{CP}(z_f, t, b, \theta) = \Pr_{\theta}\{B(1) > z_f \mid B(t) = b\}. \quad (6.17)$$

Often, the fixed sample critical value, for example, $z_{0.025} = 1.96$ is used for z_f in Equation (6.17). This provides for a slightly more optimistic conditional power calculation, since interim monitoring will make $z_f > 1.96$.

From the independent increments property B4, we know $B(1) - B(t)$ is independent of $B(t)$, so we rewrite Equation (6.17) as

$$\begin{aligned} \text{CP}(z_f, t, b, \theta) &= \Pr_{\theta}\{B(1) - B(t) > z_f - b \mid B(t) = b\} = \Pr_{\theta}\{B(1) - B(t) > z_f - b\}. \end{aligned} \quad (6.18)$$

To evaluate the last expression, we know from properties B1 and B2 that $B(1) - B(t)$ is normally distributed with mean $E_{\theta}\{B(1) - B(t)\} = \theta(1 - t)$. Moreover, using property B3, $\text{var}_{\theta}\{B(1) - B(t)\} = \text{var}\{B(1)\} + \text{var}\{B(t)\} - 2\text{cov}\{B(1), B(t)\} = 1 + t - 2t = 1 - t$. Consequently, $B(1) - B(t)$ is normally distributed with mean $\theta(1 - t)$ and variance $1 - t$ so we may rewrite Equation (6.18) as

$$\begin{aligned} \text{CP}(z_f, t, b, \theta) &= \Pr_{\theta} \left\{ \frac{B(1) - B(t) - \theta(1-t)}{\sqrt{1-t}} > \frac{z_f - b - \theta(1-t)}{\sqrt{1-t}} \right\} \\ &= 1 - \Phi \left(\frac{z_f - b - \theta(1-t)}{\sqrt{1-t}} \right). \end{aligned} \quad (6.19)$$

Equation (6.19) allows us to evaluate conditional power by using the standard normal distribution. Note that $E_{\theta}\{B(t)/t\} = \theta t/t = \theta$, so $\hat{\theta} = B(t)/t$ is an often-used estimate of the current trend (i.e., at information time t) of the data.

Example 6.7. Continuing with Example 6.5, suppose we have an interim analysis after observing outcomes for 210 treatment subjects and 216 control subjects for a total of 426 subjects. Since our maximum planned sample size is 338 subjects per arm or 676 subjects, the interim analysis corresponds to information time $t = 426/676 = 0.63$. Also suppose that the z -test of proportions at this interim analysis is $Z(0.63) = 0.32$. This corresponds to a B -value of $b = B(0.63) = \sqrt{0.63} \cdot Z(0.63) = \sqrt{0.63} \cdot 0.32 = 0.25$. To compute conditional power under the originally hypothesized treatment effect $\theta = 3.00$, we use Equation (6.19) to obtain

$$\text{CP}(1.96, 0.63, 0.25, 3.00) = 1 - \Phi \left(\frac{1.96 - 0.25 - 3.00 \cdot (1 - 0.63)}{\sqrt{1 - 0.63}} \right) = 0.16.$$

Thus, given the data that have been observed thus far, there is only a 0.16 probability of obtaining a statistically significant result for the superiority of the experimental treatment even if the remaining data followed the originally hypothesized treatment effect $\theta = 3.00$.

Similarly, computing the conditional power under the current trend of the data $\hat{\theta} = B(0.63)/0.63 = 0.25/0.63 = 0.40$, we have

$$\text{CP}(1.96, 0.63, 0.25, 0.40) = 1 - \Phi \left(\frac{1.96 - 0.25 - 0.40 \cdot (1 - 0.63)}{\sqrt{1 - 0.63}} \right) = 0.005.$$

It is intuitively clear that the conditional power under the current data trend $\hat{\theta} = 0.40$ must be lower than under the originally hypothesized treatment effect $\theta = 3.00$. It is noteworthy that while $\hat{\theta} = 0.40$ is between 1/8 and 1/7 of $\theta = 3.00$, the fact that $t = 0.63$ of the information time has elapsed further lowers the probability of obtaining a significant result under the current trend to 0.005, which is less than 1/25 of the 0.16 conditional power under the originally hypothesized effect.

Example 6.8. In the above discussion about the CAST II trial, we indicated that the trial was stopped early due, in part, to a conditional power of 0.08 assuming the remaining data followed the hypothesized treatment effect of a 30% end point rate reduction. We now show, using Equation (6.19), how the

0.08 was obtained. CAST II was designed to have 80% unconditional power corresponding to a one-sided 0.025 significance level to establish the superiority of the antiarrhythmic drug moricizine. From Equation (6.11), this corresponded to the treatment effect parameter

$$\theta = z_{1-0.80} + z_{0.025} = 0.84 + 1.96 = 2.80.$$

At information time $t = 0.37$, the logrank statistic had z -value = -1.05. From Equation (6.3) this corresponded to

$$b = B(0.37) = \sqrt{0.37}Z(0.37) = \sqrt{0.37} \cdot (-1.05) = -0.64.$$

To account for interim monitoring, the end of study critical value for treatment superiority was $z_f = z_{0.0125} = 2.24$. Substituting all of these values in Equation (6.19), $CP(2.24, 0.37, -0.64, 2.80) = 1 - \Phi((2.24 - (-0.64) - 2.80(1 - 0.37)) / \sqrt{1 - 0.37}) = 0.08$ corresponding to the 0.08 conditional power calculated for the CAST II DSMB.

A generalization of conditional power is predictive power. Predictive power is a Bayesian concept since it assumes a prior distribution for the treatment effect parameter θ and then computes the average conditional power with respect to the posterior distribution of θ given the current data. Spiegelhalter, Freedman, and Blackburn (1986) give a good discussion of predictive power.

6.9 SPENDING FUNCTIONS

In Section 6.8, we described the CAST II trial. Prior to the first interim look at the trial's data, the DSMB set up its statistical monitoring guidelines in the form of spending functions. In mathematical terms, a *spending function* (Lan and DeMets 1983) is an increasing function $\alpha^*(t)$ defined for all t in the interval $[0, 1]$ such that $\alpha^*(0) = 0$ and $\alpha^*(1) = \alpha$ for some fixed $0 < \alpha < 1$. The value $\alpha^*(t)$ is the cumulative one-sided type I or type II error, depending on whether $\alpha^*(t)$ corresponds to treatment superiority or inferiority, that is used (or *spent*) for interim treatment effect testing by information time t . Thus, $\alpha^*(1) = \alpha$ is the overall type I or type II error of the trial.

To monitor for the superiority of the antiarrhythmic drug, CAST II used a linear spending function with a jump at $t = 1$:

$$\alpha_U(t) = 0.0125 \cdot I(t < 1) \cdot t + 0.025 \cdot I(t = 1).$$

Thus, 0.0125 of the type I error was to be spent before $t = 1$ with an additional 0.0125 spent at $t = 1$ so the overall (one-sided) type I error was 0.025. To monitor for the inferiority of the antiarrhythmic drug, CAST II also used

a linear spending function with a jump at $t = 1$, but with coefficients that were twice as large as those for the superiority boundary:

$$\alpha_L(t) = 0.025 \cdot I(t < 1) \cdot t + 0.05 \cdot I(t = 1).$$

Setting the overall type II error (0.05) at twice the level of the overall type I error (0.025) meant that CAST II was monitored for antiarrhythmic inferiority, or harm, twice as strictly as for superiority. This was done in part because CAST, the predecessor of CAST II, showed that the antiarrhythmics encainide and flecainide were harmful. Thus, going into CAST II, there were concerns about its antiarrhythmic drug, moricizine.

At CAST II's first interim analysis, which occurred at information time $t = 0.135$, $\alpha_U(0.135) = 0.0125 \cdot 0.135 = 0.00169$. Thus, CAST II would have had statistical evidence for the superiority of the antiarrhythmic drug if $Z(0.135)$, the logrank z -statistic at information time 0.135, exceeded $z_{0.00169} = 2.93$. However, $Z(0.135) = 0.18 < 2.93$, and CAST II continued.

At the second interim analysis, which occurred at information time $t = 0.22$, the cumulative alpha was $\alpha_U(0.22) = 0.0125 \cdot 0.22 = 0.00275$. Since $\alpha_U(0.135) = 0.00169$ was spent at the first interim analysis, there was $\alpha_U(0.22) - \alpha_U(0.135) = 0.00275 - 0.00169 = 0.00106$ alpha to be spent at the second interim analysis. Since the logrank z -statistic at the second interim analysis $Z(0.22) = 0.36 < 3.07 = z_{0.00106}$, CAST II continued.

The above calculations show how to compute critical values for a general spending function $\alpha^*(t)$ that is being used to monitor an asymptotically Brownian motion process. If t_k is the information time at the k^{th} interim analysis, then the critical value corresponding to t_k is the upper $\alpha^*(t_k) - \alpha^*(t_{k-1})$ percentile of the standard normal distribution.

We noted above that the first two CAST II interim analyses slightly favored the antiarrhythmic drug with $Z(0.135) = 0.18$, $Z(0.22) = 0.36 > 0$. However, the third and fourth interim analyses reversed, with a trend towards antiarrhythmic harm with $Z(0.29) = -1.08$ and $Z(0.37) = -1.05$. The critical value for harm at the fourth interim analysis was $-z_{0.002} = -2.88$ since $\alpha_L(0.37) - \alpha_L(0.29) = 0.025 \cdot (0.37 - 0.29) = 0.002$. While $Z(0.37) = -1.05$ did not cross the -2.88 critical value for harm, CAST II was stopped at the fourth interim analysis due in part to low conditional power as discussed in Section 6.8.

6.10 FLEXIBILITY AND PROPERTIES OF SPENDING FUNCTIONS

A key reason why spending functions have gained wide use over the past 30 years is the flexibility they provide to DSMBs. In particular, spending functions are straightforward to implement when the information times at interim analyses are unpredictable. One example of this unpredictability is for clinical trials in which information time is measured by the observed proportion of expected events. This is because DSMB meetings are typically scheduled at regular

calendar intervals (e.g., every 6 months) and it is hard to predict the number of events that will occur by a prespecified calendar date. This was the case for CAST II, where the primary end point was time to arrhythmic death or cardiac arrest death and the logrank statistic was used. Nevertheless, as we saw in Section 6.9, the spending function $\alpha^*(t)$ only uses the information time at the previous interim analysis, say, t_{k-1} , and the information time at the current interim analysis, t_k , to calculate the critical value at the current interim analysis, that is, the upper $\alpha(t_k) - \alpha(t_{k-1})$ percentile of the standard normal distribution. Thus, the correct critical value can be calculated regardless of the number of events that occurred since the last interim analysis.

The early theoretical development for monitoring time-to-event outcomes at unpredictable information times is due to Slud and Wei (1982) and Lan and DeMets (1983). The latter authors introduced the spending function, which does not require the number of interim analyses to be specified before the trial. The theoretical development that established the validity of spending functions for the logrank and related linear rank statistics is largely due to Tsiatis (1981, 1982), Sellke and Siegmund (1983), and Slud (1984). Li and Geller (1991) discuss some practical issues related to spending functions. In particular, they study properties of spending functions that require more extreme z -statistic values to stop a trial at earlier look times, the desirability of which we discussed in Section 6.4 as an advantage of the O'Brien-Fleming boundary.

An important assumption in these papers is that the value of the z -statistic at one interim analysis does not affect the timing of the next interim analysis. This could be problematic in practice. For example, in the CAST II trial, the April 1991 interim analysis, at information time 0.29, revealed a trend toward harm for the antiarrhythmic arm with $Z(0.29) = -1.08$. Although the DSMB had been originally scheduled to next meet 6 months later in October 1991, the DSMB decided to move up that meeting to July 1991, partly due to this negative trend. However, moving up the DSMB meeting from October to July meant that a more extreme critical value would be used in July, since less information time would have elapsed. There were $0.37 - 0.29 = 0.08$ units of elapsed information time between April and July, so the lower boundary critical value was the $\alpha_L(0.37) - \alpha_L(0.29) = 0.025 \cdot (0.37 - 0.29) = 0.002$ percentile of the standard normal distribution, that is, -2.88 . Suppose instead that the DSMB had delayed their meeting until October as planned, and information time continued to elapse at the same rate as between April and July. Then the information time in October would have been 0.45 and the lower boundary critical value in October would have been the $\alpha_L(0.45) - \alpha_L(0.29) = 0.025 \cdot (0.45 - 0.29) = 0.004$ percentile of the standard normal distribution, i.e., -2.65 , which would have been less extreme than the -2.88 July threshold.

Of course, CAST II was stopped in July 1991, but these calculations suggest that a spending function may provide some protection against type I or II error inflation when the timing of the interim analyses are data driven. To investigate this issue, Lan and DeMets (1989) considered both the Pocock-like spending

function, $\alpha_p(t) = \sigma \log \{1 + (e - 1)t\}$, and O'Brien–Fleming-like spending function, $\alpha_{OF}(t) = 2\{1 - \Phi(z_{\alpha/2} t^{1/2})\}$. These spending functions are continuous approximations to the Pocock and O'Brien–Fleming boundaries, respectively, for equally spaced looks discussed in Section 6.4. For these spending functions, Lan and DeMets began with prespecified, equally spaced interim analysis information times and increased the frequency of interim analyses in the following manner: Suppose at a prespecified interim analysis time t_k for the spending function $\alpha^*(t)$, the z -statistic $Z(t_k)$ was greater than 80% of the corresponding critical value $c(t_k)$. That is, $0.80c(t_k) < Z(t_k) < c(t_k)$, where $c(t_k)$ is the upper $\alpha^*(t_k) - \alpha^*(t_{k-1})$ percentile of the standard normal distribution. Then the frequency of the remaining interim analyses was doubled. For example, consider the O'Brien–Fleming-like alpha spending function with four pre-trial specified analysis times at $t_1 = 0.25$, $t_2 = 0.50$, $t_3 = 0.75$ and $t_4 = 1$. In particular, the second-look critical value is $c(t_2) = 2.963$. If the second look z -statistic was $2.40 > 0.8 \cdot 2.963 = 2.37$, then remaining looks would be taken at times 0.625, 0.75, 0.875, and 1.0. Using this data-driven design, Lan and DeMets found that power and type I error were essentially unchanged.

Proschan, Follmann, and Waclawiw (1992) made a more determined effort to inflate the type I error through data-driven looks. For a three-look trial, they fixed the first and final looks at information times $t_1 = \delta$ and $t_3 = 1$, respectively. Then, using a grid search, they varied the second look time t_2 as a function of the first-look z -statistic so as to maximize the conditional power of the trial. Even when using this *intention to cheat* approach (Proschan, Lan, and Wittes 2006, p. 90), they found that for the O'Brien–Fleming, Pocock, and linear spending functions ($\alpha^*(t) = \alpha t$), the type I error was inflated by no more than about 10% for one-sided 0.025 and 0.05 significance levels, respectively, that is, type I errors of no more than about 0.0275 and 0.055, respectively. For more than three looks, they used a less computationally intensive method and obtained similar results.

Thus, spending functions provide great flexibility for monitoring trials, and as well as substantial type I and II error protection when look times are data driven. Equally important in their widespread use is the availability of user-friendly software for their implementation. Chapter 14 of Proschan, Lan, and Wittes (2006) provides a good introduction to the freely available software (<http://www.biostat.wisc.edu/landemets/>) developed by Reboussin et al. (2000).

6.11 MODIFYING THE TRIAL'S SAMPLE SIZE BASED ON A NUISANCE PARAMETER

As we saw in Example 6.4, specifically Equation (6.14), the sample size calculation for a continuous outcome depends on both the hypothesized mean difference between the treatment and control groups as well as the standard deviation of the outcome. In the sample size calculation (Eq. 6.14), a 4 mm Hg

mean difference in 3-month systolic blood pressures between the treatment and control groups was the treatment effect of interest. However, the 10 mm Hg standard deviation also appears in the denominator of Equation (6.14) as $\sqrt{200} = \sqrt{2 \cdot 10^2}$ and the resulting sample size is 132 subjects per arm. Since the standard deviation is not the main quantity of interest in the treatment versus control comparison, it is referred to as a *nuisance parameter*. If the standard deviation was 12 mm Hg, then we would need 189 subjects per arm to maintain 90% power. Moreover, if we ran the trial with 132 subjects per arm and the standard deviation was actually 12 mm Hg, the trial's power would drop to 77%.

Similarly, in Example 6.5, specifically Equation (6.16), we need 338 subjects per arm to have 85% power to obtain a significant result under the hypothesized 33% reduction in the control group event rate which we assume to be $p_c = 0.30$. However, if p_c were actually 0.20, we would need 557 subjects per arm to maintain 85% power to obtain a significant result under the hypothesized 33% event rate reduction, that is, $p_T = 0.133$. In such a situation, 338 subjects per arm would only provide 65% power. We see that the control group event rate is the nuisance parameter for a dichotomous outcome. It is not uncommon in a clinical trial to have the control group event rate be lower than what was initially hypothesized. This is because subjects who enroll in clinical trials are often healthier than the typical patient in the population of interest, whether by self-selection or because of the eligibility criteria. Also, the subjects' other medications or lifestyle modifications may make them healthier than previous patients upon whom sample size calculations are often based.

The above discussion shows that it would be useful to monitor nuisance parameters and consider making sample size adjustments to maintain unconditional power. One would want to be sure this is done in such a manner that overall type I error is not jeopardized. The next two subsections present such methods for modifying sample size.

6.11.1 Sample Size Modification for a Continuous Outcome Based on an Interim Variance Estimate

Up until now for the continuous outcome case, we have assumed that our sample sizes were large enough so that the standard deviation of the test statistic could be treated as the true value, and the test statistic was approximately normally distributed. However, one of the most important early papers on sample size modification due to Stein (1945) considered the one sample normally distributed outcome case in which the standard deviation was unknown, so a *t*-statistic was needed for testing.

We demonstrate Stein's procedure for the two-armed treatment versus control arm setting from Section 6.2 with $X_{T1}, X_{T2}, \dots, X_{TN}$ independent $N(\mu_T, \sigma^2)$ treatment arm observations and $X_{C1}, X_{C2}, \dots, X_{CN}$ independent $N(\mu_C, \sigma^2)$ control arm observations. With N observations per arm, the test statistic is

$$\frac{\bar{X}_C - \bar{X}_T}{\sqrt{2s^2/N}}. \quad (6.20)$$

where s^2 is the pooled sample variance. Of course, if s^2 were known to be σ^2 , the per arm sample size N required to obtain a statistically significant difference in means with power $1 - \beta$ under the hypothesized difference δ at the two-sided α significance level would be

$$N = \frac{2\sigma^2(z_{\alpha/2} + z_{\beta})^2}{\delta^2}. \quad (6.21)$$

However, a key point of Stein's procedure is that we do not know what σ^2 is.

In the first stage of Stein's procedure, n_1 independent observations per arm are taken to obtain the pooled sample variance estimate s_1^2 with $2(n_1 - 1)$ degrees of freedom. Stein then uses s_1^2 as the sample variance estimate in the t -statistic at the end of the trial (Eq. 6.21) based on $N \geq n_1$ observations per arm. That is, Stein's t -statistic is

$$t_S = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{2s_1^2/N}}.$$

The final per arm sample size N is chosen so as to guarantee at least $1 - \beta$ power. This is done by setting $\sigma^2 = s_1^2$ in Equation (6.21) and *assuming* t_S has a null hypothesis central t distribution with $2(n_1 - 1)$ degrees of freedom which we denote by $t_{2(n_1 - 1)}$. That is,

$$N = \max\{n_1, 2s_1^2(t_{\alpha/2,2(n_1-1)} + t_{\beta,2(n_1-1)})^2/\delta^2\} \quad (6.22)$$

where $t_{r,2(n_1-1)}$ is the upper r^{th} percentile of the $t_{2(n_1-1)}$ distribution. Stein showed that t_S does in fact have a null hypothesis $t_{2(n_1-1)}$ distribution.

Stein's procedure is an early example of using the first stage data as an *internal pilot study* (Wittes and Brittain 1990) to estimate the nuisance parameter σ^2 and possibly modify the total sample size for the trial after the trial has started. However, Stein's procedure is rarely used today for several reasons. First, it is somewhat unintuitive to disregard the second-stage data in the variance estimate s_1^2 for the end-of-trial test statistic t_S . Second, there is the possibility that the second-stage subjects are different from the first-stage subjects with a different σ^2 value. It could be that the $2n_1$ first-stage subjects were more selectively chosen, as the trial was in its early stage and the trial protocol was still being refined, than the $2(N - n_1)$ second-stage subjects. This could result in greater variability among the second-stage subjects. s_1^2 would then underestimate σ^2 , hence inflating the type I error of t_S . Finally, as Zucker et al. (1999) discuss, although Stein's procedure controls the *unconditional* type I error at the α level, it does not control the conditional, on the final sample size $2N$,

type I error. As Zucker et al. point out, there is an ongoing debate over whether conditional error needs to be controlled.

If we wish to use all of the data to estimate the variance of the end-of-trial t -statistic, a natural procedure might be as follows: Provisionally calculate the total sample size $2\tilde{n}$ using a guess for σ^2 in Equation (6.21). Then conduct an internal pilot study using the first $2n_1$ observations where $n_1 < \tilde{n}$ (e.g., $n_1 = \tilde{n}/4$) to estimate σ^2 by s_1^2 . Suppose, based on s_1^2 , we calculate in Equation (6.21) the required per arm sample size to be N . Then, for the end-of-trial t -statistic, we would use the usual t -statistic

$$\frac{\bar{X}_T - \bar{X}_C}{\sqrt{2s^2 / N}}, \quad (6.23)$$

where s^2 is the pooled variance based on all $2N$ trial observations. Wittes and Brittain (1990) call Equation (6.23) the *naïve t-test* since we pretend N was chosen before the trial began and refer to a t distribution with $2(N - 1)$ degrees of freedom for hypothesis testing. In fact, as shown rigorously in Wittes et al. (1999) and Proschan (2005), and heuristically in Proschan, Lan, and Wittes (2006, section 11.2.2), s^2 on average underestimates σ^2 . Fortunately, as shown in Wittes et al. (1999), the effect of this bias on type I error and power is quite small.

Returning to Example 6.4, suppose we wanted to conduct a smaller trial with 90% power to obtain a statistically significant difference in means under the larger hypothesized difference of $\delta = 6$ mmHg, but still assuming the common standard deviation of $\sigma = 10$ mmHg. Also, suppose we use a symmetric O'Brien–Fleming-like spending function with data looks at information times 0.50 and 1.0. Using the Lan–DeMets software, we find that the drift parameter is $\theta = 3.2473$ which, for N subjects per arm, satisfies

$$\theta = E\{Z(1)\} = E\left\{\frac{\bar{X}_C - \bar{X}_T}{\sqrt{2\sigma^2/N}}\right\} = \frac{\delta}{\sqrt{2\sigma^2/N}} = \sqrt{N} \cdot \frac{6}{\sqrt{2 \cdot 100}}. \quad (6.24)$$

Thus, we initially estimate $N = (3.2473 \cdot \sqrt{200}/6)^2 = 59$ subjects per arm.

However, there are two additional concerns that we need to address in the design phase of this trial. First, although the blood pressure change outcomes will be essentially normally distributed, our modest sample size may make z -statistic-based monitoring boundaries for type I and II error anticonservative since the z -statistic approximation to the t -statistic may be questionable. Second, our sample size calculation (Eq. (6.24)) may use an incorrect guess for σ^2 .

To address the first concern, Proschan, Lan, and Wittes (2006, section 8.1) suggest using the *nominal p-value approach* to adjust the monitoring boundaries to accomodate the t -statistic. To address the second concern, we can reestimate σ^2 before the first interim analysis was planned to occur.

Based on the 118 total subjects determined from Equation (6.24), the first interim analysis for treatment efficacy was originally planned to occur at infor-

mation time 0.50, that is, after a total of 60 subjects were observed for the end point. Thus, without looking at the observed treatment difference, we may reestimate σ^2 after the first 50 subjects, which would be before the first interim analysis for treatment efficacy was originally planned. Suppose we found the pooled variance estimate was 144 mm Hg, hence estimating σ to be 12 mm Hg. This gives a revised per arm sample size of $N = (3.2473 \cdot \sqrt{2 \cdot 144}/6)^2 = 85$. We reemphasize that the sample size was increased before taking any interim looks at the treatment effect. Thus, as we have discussed, modifying the trial's sample size based on an interim variance estimate has a very small impact on the type I error so the treatment efficacy monitoring may proceed as if the original total sample size had been 170 subjects.

Corresponding to the efficacy look times 0.50 and 1.0, the Lan–DeMets software gives corresponding one-sided nominal p -values of 0.00153 and 0.0245, respectively. For the first-look critical value corresponding to 85 total subjects, the nominal p -value approach uses $t_{0.00153, 85-2} = 3.051$ for the t -test statistic instead of $z_{0.00153} = 2.962$ for stopping the trial for treatment superiority. If the trial continued to the second and final look with 170 total subjects, the nominal p -value approach uses the critical value $t_{0.02450, 170-2} = 1.983$ instead of $z_{0.02450} = 1.969$. We see in this example that the moderate-to-large sample sizes result in modest differences between the t -statistic and z -statistic critical values. However, table 8.2 of Proschan, Lan, and Wittes (2006) shows that the nominal p -value approach performs quite well when the sample sizes are substantially smaller.

6.11.2 Sample Size Modification for a Dichotomous Outcome Based on an Interim Estimate of the Pooled Event Rate

As discussed in the introductory part of this section, a dichotomous outcome trial can be underpowered due to a pretrial underestimation of the control group's event rate. We use Example 6.5 to show how this can be resolved. We want 85% power to obtain a statistically significant result under a hypothesized 33% reduction in the control event rate p_C for the treatment event rate p_T . That is, $p_T = 0.67p_C$ with a pretrial guess of $p_C = 0.30$. Suppose we use a symmetric O'Brien–Fleming-like spending function with data looks at information times 0.50, 0.75, and 1.0. Using the Lan–DeMets software, we find that the drift parameter is $\theta = 3.025$. Thus, using the z -test for proportions and N subjects per arm, we have

$$\begin{aligned} 3.025 &= \theta = E(Z(1)) = E\left\{\frac{\bar{X}_C - \bar{X}_T}{\sqrt{2p(1-p)/N}}\right\} = \frac{p_C - p_T}{\sqrt{2p(1-p)/N}} \\ &= \frac{0.10}{\sqrt{2 \cdot (0.25 \cdot 0.75)/N}}. \end{aligned} \quad (6.25)$$

Hence our initial per-arm sample size estimate is $N = 344$. To estimate the trial's sample size using an internal pilot study, we follow an initial n_{C1} control

subjects and n_{T1} treatment subjects for the end point. Let \bar{X}_{C1} and \bar{X}_{T1} be the corresponding observed event rates, and let $\bar{X}_1 = (n_{C1}\bar{X}_{C1} + n_{T1}\bar{X}_{T1})/n_1$ be the observed pooled event rate where $n_1 = n_{C1} + n_{T1}$. Then we set

$$\bar{X}_1 = \frac{p_C + p_T}{2} \quad (6.26)$$

$$p_T = 0.67 \cdot p_C \quad (6.27)$$

and solve for p_C and p_T and use Equation (6.25) to determine the revised trial sample size. For example, suppose $n_{C1} = 157$, $n_{T1} = 143$, $\bar{X}_{C1} = 0.21$, and $\bar{X}_{T1} = 0.19$. Then $\bar{X}_1 = (157 \cdot 0.21 + 143 \cdot 0.19)/300 = 0.20$, and we solve Equations (6.26) and (6.27) to get $p_C = 0.24$ and $p_T = 0.16$. Substituting $p_C = 0.24$ and $p_T = 0.16$ with $\theta = 3.025$ into Equation (6.25) gives a revised trial sample size of $N = 458$ subjects per arm, i.e., 916 total subjects. The data look times 0.50, 0.75, and 1.0, correspond to 458, 687, and 916 total subjects. The Lan–DeMets software gives corresponding one-sided nominal p -values of 0.00153, 0.00916, and 0.02200, respectively, corresponding to critical values of 2.96, 2.36, and 2.01, respectively, at the three data looks. As in the previous subsection for a continuous outcome, the recalculation of the trial's sample size occurred before taking any interim looks for treatment efficacy.

A key feature of the above procedure is that we use the first stage's pooled event rate \bar{X}_1 , and not the control group event rate \bar{X}_{C1} , to recalculate the sample size. This is important since under $H_0: p_C = p_T$, \bar{X}_1 is asymptotically independent of the observed treatment effect $\bar{X}_{C1} - \bar{X}_{T1}$, but \bar{X}_{C1} is not. To see this, first note that

$$\begin{aligned} \text{cov}(\bar{X}_1, \bar{X}_{C1} - \bar{X}_{T1}) &= \frac{1}{2} \{ \text{cov}(\bar{X}_{C1}, \bar{X}_{C1}) + \text{cov}(\bar{X}_{T1}, \bar{X}_{C1}) - \text{cov}(\bar{X}_{C1}, \bar{X}_{T1}) \\ &\quad - \text{cov}(\bar{X}_{T1}, \bar{X}_{T1}) \} = \frac{1}{2} \{ p_C(1-p_C) + 0 - 0 - p_T(1-p_T) \} = 0. \end{aligned}$$

Since \bar{X}_1 and $\bar{X}_{C1} - \bar{X}_{T1}$ are asymptotically bivariate normal with covariance 0, they are asymptotically independent. On the other hand, $\text{cov}(\bar{X}_{C1}, \bar{X}_{C1} - \bar{X}_{T1}) = \text{cov}(\bar{X}_{C1}, \bar{X}_{C1}) - \text{cov}(\bar{X}_{C1}, \bar{X}_{T1}) = p_C(1-p_C) - 0 > 0$. Thus, when considering a sample size modification, the pooled event rate estimate can be provided to investigators while keeping them blinded to arm-specific event rates and not (asymptotically) inflating the type I error.

6.12 SAMPLE SIZE MODIFICATION BASED ON THE INTERIM TREATMENT EFFECT

Suppose we are interested in sample size modification after an interim look at the treatment effect. This may be because the interim treatment effect esti-

mate, albeit imprecisely measured due to the interim sample size, may be smaller than the one for which we initially powered the trial. Modification based on an interim treatment effect estimate is more controversial since the treatment effect is the primary end point so using an interim effect estimate to modify the trial design may appear to be “hedging the bet” that the treatment is efficacious.

There are two general methods for such sample size modifications. The first method requires that sample size modifications, as functions of interim treatment effect estimates, be specified in advance of the interim look. Such functions are based on unconditional criteria, such as minimizing a weighted average of expected sample sizes over a range of hypothesized treatment effects. The second method allows modifications to be made after observing an interim treatment effect. Such a modification can be based on conditional (on the interim treatment effect estimate) criteria, such as maximizing conditional power at a particular alternative. The latter methods are widely known as *adaptive designs* (see Chapter 7), although as Shih (2006) points out, any sequential design that uses interim data (such as a nuisance parameter estimate) to determine the course of the trial is adaptive. We will use Shih’s terminology and refer to the former, prespecified adaptive methods as *group sequential* (GS) methods. Our discussion of the latter adaptive designs will focus on a large class of such designs known as *variance spending* (VS) designs (Bauer and Köhne 1994; Proschan and Hunsberger 1995; Fisher 1998).

In this section alone, it will be useful to modify the notation from previous sections and allow Z_j to be the normalized z -statistic corresponding to the data obtained between the $(j - 1)^{\text{st}}$ and j^{th} looks at the data. Similarly, we let n_j denote the number of observations per arm corresponding to Z_j . Finally, let (c_{Lj}, c_{Uj}) denote look j s lower and upper critical values.

We largely restrict our discussion to trial designs with a maximum of two looks since the GS and VS methods are easiest to describe with only two looks, and the two-look designs capture the basic ideas of designs with more than two looks. The first and second look z -statistics used by the GS method are

$$\text{GS first-look } z\text{-statistic } Z_{GS,1} = Z_1 \quad (6.28)$$

$$\text{GS second-look } z\text{-statistic } Z_{GS,2} = \frac{\sqrt{n_1}Z_1 + \sqrt{n_2}Z_2}{\sqrt{n_1 + n_2}}. \quad (6.29)$$

$Z_{GS,1}$ and $Z_{GS,2}$ are simply the first- and second-look z -statistics from Section 6.3’s Brownian motion setup. In particular, conditional on n_1 and n_2 , $Z_{GS,2}$ is the unique uniformly most powerful test statistic for the drift parameter θ ; note that if $n_2 = 0$, then $Z_{GS,2} = Z_1$. The distinguishing feature of the GS method is that $n_1, n_2(z_1), c_{L1}, c_{U1}$, and $c_2(z_1) \equiv c_{L2}(z_1) = c_{U2}(z_1)$ can be chosen to optimize prespecified unconditional criteria, such as the average expected sample size (or average sample number, ASN) over a range of alternatives, subject to type I and II error constraints. Note that the second look values $n_2(z_1)$ and $c_2(z_1)$

can depend on the observed value z_1 of the first look z -statistic Z_1 , but this dependence is through the functions $n_2(\cdot)$ and $c_2(\cdot)$, which are specified before the interim look. Of course, with a maximum of two looks, $c_{L2}(z_1) = c_{U2}(z_1)$.

Jennison and Turnbull (2006) studied the operating characteristics of GS designs that were optimized with respect to a linear combination of expected sample sizes under the null and target alternatives and a more optimistic alternative than the target. More specifically, subject to fixed type I error α at $\theta = 0$ and type II error β at $\theta = \Delta > 0$, they calculated optimal GS designs for the criterion

$$\frac{1}{3}\{ASN(\theta = 0) + ASN(\theta = \Delta) + ASN(\theta = L\Delta)\}, \quad (6.30)$$

where they took $L = 2$ or 4 .

Jennison and Turnbull considered four classes of GS designs, each with increasing amounts of flexibility; below, let n_f denote the per arm sample size of the fixed sample test corresponding to α , β , and Δ :

- *Class A: Equal Batch Size GS Designs.* For the error spending functions $f(t) = \alpha t^\rho$ and $g(t) = \beta t^\rho$, optimally choose the inflation factor R where $n_1 + n_2 = R n_f$ and $n_1 = n_2$. Jennison and Turnbull showed that there is a one-to-one relationship between R and ρ (Jennison and Turnbull 2000, table 7.6). Also, c_{L1} , c_{U1} , c_2 are determined by the choice of ρ and R . Since neither n_2 nor c_2 depend on z_1 , Jennison and Turnbull called such a design a *non-adaptive group sequential design*.
- *Class B: Unequal Batch Size GS Designs.* As for Class A, but we also optimally choose the initial batch size n_1 . This is also a nonadaptive design.
- *Class C: Optimal Nonadaptive GS Designs.* Optimization with respect to n_1 , n_2 , c_{L1} , c_{U1} , and c_2 , where n_2 and c_2 do not depend on z_1 .
- *Class D: Optimal Adaptive GS Designs.* As for Class C, but n_2 and c_2 may depend on z_1 . These designs are also discussed in Schmitz (1993).

While the class D designs give the greatest flexibility, they require unblinding of the interim treatment effect estimate through z_1 since the second batch size $n_2(z_1)$ and critical value $c_2(z_1)$ depend on z_1 . In fact, Jennison and Turnbull's numerical results showed that little efficiency is lost with respect to criterion Equation (6.30) by using a class B design instead of class C or D. While we presented these designs as two-look designs, Jennison and Turnbull studied designs with up to six looks. They concluded that class B would be practical for many group sequential settings.

For the VS method, the first and second look z -statistics are

$$\text{VS first-look } z\text{-statistic } Z_{VS,1} = Z_1 \quad (6.31)$$

$$\text{VS second-look } z\text{-statistic } Z_{VS,2} = \frac{\sqrt{n_1}Z_1 + \sqrt{m}Z_2}{\sqrt{n_1+m}}. \quad (6.32)$$

where m is set before the first-stage result z_1 is observed. A key difference between the GS and VS methods are the weights in the second look z -statistic. For the VS method, n_1 and m are typically the anticipated first and second look batch sizes based on pretrial unconditional power calculations. However, in this method, n_2 , the number of observations actually taken for Z_2 , may not be equal to m . Nevertheless, the weight attached to Z_2 in $Z_{VS,2}$ is $\sqrt{m}/\sqrt{n_1+m}$, while the weight attached to Z_2 in $Z_{GS,2}$ is $\sqrt{n_2}/\sqrt{n_1+n_2}$. Thus, more (respectively, less) weight is attached to Z_2 in $Z_{VS,2}$ than in $Z_{GS,2}$, if $n_2 < m$ (respectively $n_2 > m$). The square of the $\sqrt{m}/\sqrt{n_1+m}$ weight attached to Z_2 corresponds to the proportion of the total variance of $Z_{VS,2}$ that is *spent* on the second look, hence the name of the method.

One motivation for this method is discussed by Proschan and Hunsberger (1995) who showed how to achieve a desired conditional power given $Z_1 = z_1$. To describe Proschan and Hunsberger's method, we assume $n_1 = m$ so the first look takes place halfway through the pretrial sample size choice. (This is only a matter of convenience, and their method can be easily implemented if $n_1 \neq m$.) Then, equal weights are attached to Z_1 and Z_2 in the second look z -statistic

$$\frac{\sqrt{n_1}Z_1 + \sqrt{n_1}Z_2}{\sqrt{2n_1}} = \frac{Z_1 + Z_2}{\sqrt{2}}. \quad (6.33)$$

To have a one-sided α -level test, we use the rejection region

$$\frac{Z_1 + Z_2}{\sqrt{2}} > z_\alpha. \quad (6.34)$$

In Equation (6.34), Z_2 corresponds to a batch size of $n_2(z_1)$, where n_2 is defined in Equation (6.39).

From Equation (6.34), the conditional type I error given $Z_1 = z_1$ is

$$A(z_1) = 1 - \Phi(\sqrt{2}z_\alpha - z_1). \quad (6.35)$$

Proschan and Hunsberger called $A(z_1)$ the *conditional error function*. Next, suppose the observations in the treatment and control groups have common variance σ^2 with hypothesized treatment effect by δ . Given $Z_1 = z_1$, Z_2 has normal distribution with mean $\sqrt{n_2(z_1)}\delta/\sqrt{2}\sigma$ and variance 1. It follows from Equation (6.34) that the conditional power under δ is

$$\Pr_\delta \left\{ Z_2 - \frac{\sqrt{n_2}\delta}{\sqrt{2}\sigma} > \sqrt{2}z_\alpha - z_1 - \frac{\sqrt{n_2}\delta}{\sqrt{2}\sigma} \right\} \quad (6.36)$$

$$= \Phi\left(\frac{\sqrt{n_2}\delta}{\sqrt{2}\sigma} - \sqrt{2}z_\alpha + z_1\right). \quad (6.37)$$

Thus, we may obtain conditional power $1 - \beta$ if we choose n_2 sufficiently large so that

$$\Phi\left(\frac{\sqrt{n_2}\delta}{\sqrt{2}\sigma} - \sqrt{2}z_\alpha + z_1\right) = \Phi(z_\beta) \quad (6.38)$$

or

$$n_2 = \frac{2\sigma^2(\sqrt{2}z_\alpha - z_1 + z_\beta)^2}{\delta^2}. \quad (6.39)$$

Equation (6.39) has the same form as the fixed-sample sample size formula (Eq. 6.21) except that $\sqrt{2}z_\alpha - z_1$, which is the one-sided critical value for Z_2 , plays the role of $z_{\alpha/2}$ in Equation (6.21) in the two-sided setting. Note that since n_2 must be positive, we can find an $n_2 > 0$ that satisfies Equation (6.38) if and only if $z_1 - \sqrt{2}z_\alpha < z_\beta$. If $z_1 - \sqrt{2}z_\alpha \geq z_\beta$, we might have wanted to have stopped the trial after the first stage based on the large z_1 value suggesting the treatment's superiority. Thus, in practice, there may be trials for which Proschan and Hunsberger's procedure would be modified to allow for stopping for treatment superiority after the first stage.

The flexibility of this procedure lies in the fact that in Equation (6.39), the hypothesized treatment effect δ and the desired conditional power β may be chosen *after* the first stage $Z_1 = z_1$ has been observed (but of course before Z_2 is observed). Proschan, Lan, and Wittes (2006, section 11.4) show why it is permissible to do this without inflating the type I error. The key point is that under the null hypothesis, Z_2 has standard normal distribution regardless of the choice of n_2 . Thus, Z_1 and Z_2 are independent, each with a standard normal distribution, so $(Z_1 + Z_2)/\sqrt{2}$ has a standard normal distribution under the null hypothesis.

While the equal weighting of Z_1 and Z_2 in the test statistic $(Z_1 + Z_2)/\sqrt{2}$ is what provides for the flexibility of this procedure, it is the equal weighting that has led to criticisms of this and other similar adaptive methods. Unusual circumstances can arise, such as the following: suppose $Z_1 = 0.1$, $Z_2 = 2.5$, and $n_2 = 3n_1$. In this case, the second-stage sample is three times larger than the first stage and also has a highly significant z -statistic. Consequently, the group sequential z -statistic $Z_{GS} = (\sqrt{1 \cdot 0.1 + \sqrt{3}} \cdot 2.5)/\sqrt{1+3} = 2.22 > 1.96$ is significant, but the equally weighted z -statistic $Z_{VS} = (0.1 + 2.5)/\sqrt{2} = 1.84 < 1.96$ is not significant. Of course this is because the equally weighted z -statistic attaches equal weights to $Z_1 = 0.1$ and $Z_2 = 2.5$ despite Z_2 corresponding to substantially more data. However, such a circumstance would rarely occur in practice.

When considering the relative advantages of GS methods versus VS methods, Jennison and Turnbull (2006) draw the distinction between adapting to an internal estimate of the treatment effect versus adapting to circumstances from outside of the trial. For the latter case, they give a hypothetical pharmaceutical industry example in which news of a competing drug's side effects might lead trial investigators to decrease the clinically relevant difference, for example, δ in Equation (6.21), they wish to detect. In such an instance, it may be appropriate to use a VS method such as Proschan and Hunsberger's to increase the sample size to obtain sufficient conditional power for the smaller effect size.

However, VS methods that modify future batch sizes based on interim *within-trial* data are contained within Jennison and Turnbull's class D adaptations for which they showed their class B designs are nearly as efficient. Jennison and Turnbull (2003) also showed that the VS method of Cui, Hung, and Wang (1999) performed noticeably worse with respect to unconditional power and average sample size as compared with a class B design using the linear type I and II error spending functions $f(t) = 0.025 t$ and $g(t) = 0.1 t$, respectively. Of course, comparisons based on unconditional power and average sample size are not entirely relevant if the trial's sample size needs to be modified due to external circumstances such as the one described above, so the comparison to Cui et al. is not entirely fair.

Bauer and Köhne (1994) discussed adaptive designs that can incorporate even more drastic mid-trial modifications than simply changing the total sample size. Such adaptations could involve dropping treatment arms in a multi-armed trial, changing doses, shifting interest to subgroups, or even changing the primary end point, all while preserving the overall type I error. For example, changing the primary end point may be considered if the original primary end point, such as mortality, had a lower than anticipated event rate. In such an instance, other end points such as disease-specific hospitalization could be included in the primary end point to obtain reasonable power. Of course, before such adaptations are enacted, the credibility of the resulting trial should be considered.

Bauer and Köhne accomplished these more general modifications by working directly with the p -values p_1 and p_2 corresponding to the test statistics of the first stage and second stage subjects, respectively. Under the null hypothesis of no treatment effect, it can be shown that p_1 and p_2 are independent and identically distributed uniform random variables on the unit interval. Bauer and Köhne applied Fisher's combination of p -values test (Fisher 1932), which uses the fact that $-2\ln p_1 p_2$ has a chi-square distribution with 4 degrees of freedom, χ_4^2 . Consequently, we may reject the null hypothesis at the α significance level if $-2\ln p_1 p_2$ is greater than the upper α^{th} percentile of the χ_4^2 distribution. For example, if $\alpha = 0.05$, then we reject the null hypothesis if $-2\ln p_1 p_2 > 9.49$. Of course, if p_1 and p_2 correspond to different end points, doses, treatments, or subgroups, care must be exercised in interpreting the results of the global $-2\ln p_1 p_2$ test.

The discussion in this section may suggest a strategy of using an optimal group sequential design for as long as it is feasible, but switching to an adaptive design if unforeseen circumstances occur. As Posch, Bauer, and Brannath (2003) pointed out, the stagewise test statistics Z_1, Z_2, \dots can always be combined according to the preplanned group sequential design regardless of whether a mid-trial adaptation occurs. If no such adaptation is undertaken, the original group sequential design is left intact. While such adaptive methods mathematically preserve the type I error, the medical and regulatory communities currently remain cautious about using them as demonstrated in the February 2010 draft guidance for industry from the Food and Drug Administration (<http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf>).

6.13 CONCLUDING REMARKS

We have reviewed various statistical aspects of clinical trial monitoring. In so doing, we have enormously benefitted from Proschan, Lan, and Wittes (2006), which we would recommend to anybody wishing to learn more about statistical monitoring of clinical trials. There are other important topics that we did not discuss, including monitoring multiple end points, estimation following a group sequential trial, Bayesian monitoring, and repeated confidence intervals. Both Proschan, Lan, and Wittes (2006) and Jennison and Turnbull (2000) are excellent resources that provide further discussion on the topics we covered, as well as areas that we did not discuss. Also, our discussion has focused on randomized confirmatory (phase III) trials; there are other issues involved in monitoring phase I and II trials. It is clear that clinical trial monitoring continues to provide many challenges and open areas for research.

REFERENCES

- Anscombe, F.J. (1963). Sequential medical trials. *Journal of the American Statistical Association*, **58**, 365–383.
- Armitage, P. (1954). Sequential tests in prophylactic and therapeutic trials. *Quarterly Journal of Medicine*, **23**, 255–274.
- Armitage, P. (1958). Sequential methods in clinical trials. *American Journal of Public Health*, **48**, 1395–1402.
- Armitage, P. (1975). *Sequential Medical Trials*. Oxford: Blackwell.
- Armitage, P., C.K. McPherson, and B.C. Rowe (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A*, **132**, 235–244.
- Armitage, P. and M. Schneiderman (1958). Statistical problems in a mass screening program. *Annals of the New York Academy of Science*, **76**, 896–908.
- Arrow, K.J., D. Blackwell, and M.A. Girshick (1949). Bayes and minimax solutions of sequential decision problems. *Econometrica*, **17**, 213–244.

- Barnard, G.A. (1946). Sequential tests in industrial statistics. *Journal of the Royal Statistical Society, Supplement*, **8**, 1–26.
- Bartky, W. (1943). Multiple sampling with constant probability. *Annals of Mathematical Statistics*, **14**, 363–377.
- Bauer, P. and K. Köhne (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, **50**, 1029–1041.
- Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.). New York: Springer.
- Berger, J.O. and D.A. Berry (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, **76**, 159–165.
- Berry, D.A. (1987). Interim analysis in clinical trials: The role of the likelihood principle. *The American Statistician*, **41**, 117–122.
- Billingsley, P. (1986). *Probability and Measure* (2nd ed.). New York: Wiley.
- Bross, I. (1952). Sequential medical plans. *Biometrics*, **8**, 188–205.
- Bross, I. (1958). Sequential clinical trials. *Journal of Chronic Diseases*, **8**, 349–365.
- Cardiac Arrhythmia Suppression Trial II Investigators (1992). Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. *The New England Journal of Medicine*, **327**, 227–233.
- Cardiac Arrhythmia Suppression Trial Investigators (1989). Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *The New England Journal of Medicine*, **321**, 406–412.
- Cornfield, J. (1966). Sequential trials, sequential analysis and the likelihood principle. *The American Statistician*, **20**, 18–23.
- Cui, L., H.M.J. Hung, and S.-J. Wang (1999). Modification of sample size in group sequential clinical trials. *Biometrics*, **55**, 853–857.
- Dodge, H.F. and H.G. Romig (1929). A method for sampling inspection. *Bell System Technical Journal*, **8**, 613–631.
- Dunnett, C.W. (1961). The statistical theory of drug screening. In: *Quantitative Methods in Pharmacology*, H. de Jonge (ed.). Amsterdam: North-Holland, pp. 212–231.
- Ferrara, J.L.M., H.J. Deeg, and S.J. Burakoff (1990). *Graft-Vs.-Host Disease: Immunology, Pathophysiology, and Treatment*. New York: Marcel Dekker.
- Fisher, L.D. (1998). Self-designing clinical trials. *Statistics in Medicine*, **17**, 1551–1562.
- Fisher, R.A. (1932). *Statistical Methods for Research Workers* (4th ed.). London: Oliver and Boyd.
- Freeman, H.A., M. Friedman, F. Mosteller, and W.A. Wallis (1948). *Sampling Inspection*. New York: McGraw-Hill.
- Friedman, L.M., J.D. Bristow, A. Hallstrom, E. Schron, M. Proschan, J. Verter, D. DeMets, C. Fisch, A.S. Nies, J. Ruskin, H. Strauss, and L. Walters (1993). Data monitoring in the Cardiac Arrhythmia Suppression Trial. *The Online Journal of Current Clinical Trials*. Document number: 79 (1993 July 31).
- Holland, I.S. (1982). Central limit theorems for martingales with discrete or continuous time. *Scandinavian Journal of Statistics*, **9**, 79–94.
- Jennison, C. and B.W. Turnbull (1997). Group-sequential analysis incorporating covariate information. *Journal of the American Statistical Association*, **92**, 1330–1341.

- Jennison, C. and B.W. Turnbull (2000). *Group Sequential Methods with Applications to Clinical Trials*. New York: Chapman and Hall/CRC.
- Jennison, C. and B.W. Turnbull (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine*, **22**, 971–993.
- Jennison, C. and B.W. Turnbull (2006). Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine*, **25**, 917–932.
- Lan, K.K.G. and D.L. DeMets (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, **70**, 659–663.
- Lan, K.K.G. and D.L. DeMets (1989). Changing frequency of interim analysis in sequential monitoring. *Biometrics*, **45**, 229–241.
- Lan, K.K.G. and J. Wittes (1988). The *B*-value: A tool for monitoring data. *Biometrics*, **44**, 579–585.
- Li, Z. and N.L. Geller (1991). On the choice of times for data analysis in group sequential clinical trials. *Biometrics*, **47**, 745–750.
- Marubini, E. and M.G. Valsecchi (1995). *Analysing Survival Data from Clinical Trials and Observational Studies*. New York: Wiley.
- Meier, P. (1975). Statistics and medical experimentation. *Biometrics*, **31**, 511–529.
- MIL-STD-105E (1989). *Military Standard Sampling Procedures and Tables for Inspection by Attributes*. Washington, DC: U.S. Government Printing Office.
- Moore, T.J. (1995). *Deadly Medicine: Why Tens of Thousands of Heart Patients Died in America's Worst Drug Disaster*. New York: Simon and Shuster.
- O'Brien, P.C. and T.R. Fleming (1979). A multiple testing procedure for clinical trials. *Biometrics*, **35**, 549–556.
- Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, **64**, 191–199.
- Pocock, S.J. (1992). When to stop a clinical trial. *British Medical Journal*, **305**, 235–240.
- Posch, M., P. Bauer, and W. Brannath (2003). Issues in designing flexible trials. *Statistics in Medicine*, **22**, 953–969.
- Proschan, M.A. (2005). Two-stage adaptive methods based on a nuisance parameter: A review. *Journal of Biopharmaceutical Statistics*, **4**, 559–574.
- Proschan, M.A. and S.A. Hunsberger (1995). Designed extension of studies based on conditional power. *Biometrics*, **51**, 1315–1324.
- Proschan, M.A., D.A. Follmann, and M.A. Waclawiw (1992). Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics*, **48**, 1131–1143.
- Proschan, M.A., K.K.G. Lan, and J.T. Wittes (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach*. New York: Springer.
- Reboussin, D.M., D.L. DeMets, K. Kim, and K.K.G. Lan (2000). Computations for group sequential boundaries using the Lan-DeMets spending function method. *Controlled Clinical Trials*, **21**, 190–207.
- Roseberry, T.D. and E.A. Gehan (1964). Operating characteristic curves and accept-reject rules for two and three stage screening procedures. *Biometrics*, **20**, 73–84.
- Scharfstein, D.O., A.A. Tsiatis, and J.M. Robins (1997). Semiparametric efficiency and its implication on the design and analysis of group sequential studies. *Journal of the American Statistical Association*, **92**, 1342–1350.

- Schmitz, N. (1993). *Optimal Sequentially Planned Decision Procedures. Lecture Notes in Statistics*, vol. 79. New York: Springer.
- Schneiderman, M. (1961). Statistical problems in the screening search for anticancer drugs by the National Cancer Institute of the United States. In: *Quantitative Methods in Pharmacology*, H. de Jonge (ed.). Amsterdam: North-Holland, pp. 232–246.
- Sellke, T. and D. Siegmund (1983). Sequential analysis of the proportional hazards model. *Biometrika*, **70**, 315–326.
- Shih, W.J. (2006). Group sequential, sample size re-estimation and two-stage adaptive designs in clinical trials: A comparison. *Statistics in Medicine*, **25**, 933–941.
- Slud, E.V. (1984). Sequential linear rank tests for two-sample censored survival data. *Annals of Statistics*, **12**, 551–571.
- Slud, E.V. and L.-J. Wei (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of the American Statistical Association*, **77**, 862–868.
- Spiegelhalter, D.J., L.S. Freedman, and P.R. Blackburn (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials*, **7**, 8–17.
- Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics*, **16**, 243–258.
- Tsiatis, A.A. (1981). The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika*, **68**, 311–315.
- Tsiatis, A.A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association*, **77**, 855–861.
- Wald, A. (1947). *Sequential Analysis*. New York: Wiley.
- Wald, A. and J. Wolfowitz (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics*, **19**, 326–339.
- Wittes, J. and E. Brittain (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, **9**, 65–72.
- Wittes, J., O. Schabenberger, D. Zucker, E. Brittain, and M. Proschan (1999). Internal pilot studies I: Type I error rate of the naïve t-test. *Statistics in Medicine*, **18**, 3481–3491.
- Yusuf, S., R. Peto, J. Lewis, R. Collins, and P. Sleight (1985). Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Progress in Cardiovascular Diseases*, **XXVII**, 335–371.
- Zucker, D.M., J.T. Wittes, O. Schabenberger, and E. Brittain (1999). Internal pilot studies II: Comparisons of various procedures. *Statistics in Medicine*, **18**, 3493–3509.

CHAPTER 7

Adaptive Randomization in Clinical Trials

Lanju Zhang and William F. Rosenberger

7.1 INTRODUCTION

Randomization is the hallmark of the well-conducted clinical trial. When properly implemented, randomization mitigates bias and provides a basis for inference. However, in many clinical trials, the importance of these benefits is minimized, and randomization is confused with *allocation*, the assignment of treatments to patients. While such allocation may be randomized, randomization is often a side note, double-masking is considered the key to eliminate bias, and inference is conducted without regard to the randomization procedure employed. Consequently, any random person could develop a reasonably unpredictable sequence of assignments in a double-masked trial, and there would be no need for a formal randomization procedure.

Randomization can promote certain objectives when properly applied; for example, it can promote balance among treatment assignments, improve efficiency, assign more patients to the better treatment, or promote balance across known and unknown covariates. These objectives can be achieved simultaneously while eliminating selection bias and providing a basis for inference. Here, we describe *adaptive randomization procedures* and discuss their use in achieving multiple objectives. The area of adaptive randomization is a subset of the broader topic *adaptive designs*.

In the context of clinical trials, adaptive designs refer to changing aspects of the trial design in mid-course, often using accruing data to make interim decisions. It is in the spirit of increasing flexibility of design modification

Design and Analysis of Experiments: Special Designs and Applications, First Edition. Edited by Klaus Hinkelmann.

© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

without sacrificing trial integrity that adaptive designs are proposed and promoted.

Dragalin (2006) distinguishes among four aspects of clinical trials that characterize the broad concept of adaptive designs:

- *Adaptive Randomization.* The randomization probabilities are changed during the course of the trial.
- *Adaptive Sampling Rule.* The sample size is reestimated based on accruing data during the trial.
- *Adaptive Stopping Rule.* Sequential monitoring is employed to potentially stop the trial early due to either efficacy, futility, or safety.
- *Adaptive Decision Rule.* Flexible interim decisions that may alter the protocol of the clinical trial, for example, changing the trial endpoint, the test statistics, or dropping treatment arms during the course of the trial.

While each of these four aspects has been heavily researched independently for years (see e.g., Chapter 6), the broader theoretical construct of adaptive designs as an umbrella for any interim changes during clinical trials provides a new subclass of experimental design that has recently become a hot topic in statistics. The primary focus of theoretical research has been the development of adaptive designs that do not undermine the validity or integrity of the clinical trial. Preservation of error rates, historically the purview of sequential analysis, is the prime focus of much of the research. Virtually the entire area of sequential analysis (as it is applied to clinical trials) is captured under the third bullet point: *adaptive stopping rule*.

Here we focus on item 1, *adaptive randomization*, another broad area about which books have been written (e.g., Rosenberger and Lachin 2002; Hu and Rosenberger 2006). We focus on how adaptive randomization can be used to achieve certain objectives while preserving inferential validity and error rates. In Section 7.2, we describe different classes of adaptive randomization procedures. In Section 7.3, we discuss likelihood-based properties of these procedures. In Section 7.4, we describe randomization as a basis for inference. In Section 7.5, we consider practical implications of these procedures and regulatory perspectives and draw conclusions.

7.2 ADAPTIVE RANDOMIZATION PROCEDURES

Blocking, randomization, and replication constitute the three pillar principles of design of experiments, as described by Fisher (1935). While Fisher may not have envisioned clinical trials, today we might say that the principles of randomization and stratification are the pillars of design of clinical trials. Through stratification on known covariates, balance and comparability can be obtained within important subgroups. Through randomization, accidental bias and selection bias can be mitigated such that subjects in treatment groups are as

comparable as possible, and thus the treatment effect can be more accurately estimated or tested. Accidental bias is caused by omitting an unknown covariate of importance in the analysis. Selection bias occurs in trials where an investigator, consciously or unconsciously, tries to “beat” randomization by guessing the next treatment assignment and allocate it to a patient he/she deems to have a better (or worse) response. In combination, stratification and randomization can be used to balance known or unknown covariates among treatment groups such that their confounding effect on treatment is controlled to a minimum. Another benefit of randomization is that it provides a naturally attractive way to conduct inferential tests, known as randomization tests (see e.g., HK1, chapters 5 and 6).

Traditionally randomization procedures in clinical trials attempt to balance treatment assignments throughout the trial, or within a stratum in stratified randomization. However, the number of strata can easily multiply to be unrealistically large relative to the total sample size. However, many patients are allocated in blocks with probability 1 leading to the potential of selection bias, and perfect balance may be elusive if there are unfilled blocks at the end of the trial.

Adaptive randomization is proposed to address these shortcomings. An adaptive randomization procedure assigns patients to treatment arms with a probability that can change dependent on the history of patient assignments, patient responses, and/or covariates. Adaptive randomization can be classified into different categories dependent on the breadth of available trial information the adaptation is based on; for details, see Zhang and Rosenberger (2011). In the following, we focus on three classes of adaptive randomization procedures.

7.2.1 Restricted Randomization Procedures

First, we consider randomization procedures that promote balanced treatment assignments while being *fully randomized*, in that all patients are assigned to a treatment with probability less than 1. It is a long-held belief that balanced treatment assignments lead to the most efficient test of the treatment effect, although there are a number of scenarios where unbalanced treatment assignments are more desirable that we will discuss in a later section. Complete randomization balances treatment assignments asymptotically. Since assignments of patients to one of the two treatments are independent Bernoulli trials with probability 1/2, a nonnegligible probability of imbalance can occur in case of finite sample size. *Restricted randomization* determines the randomization probabilities based on the current treatment imbalance, skewing it to favor the treatment that has been assigned least often.

Assume there are two treatments, A and B . Let $N_i(j), i = A, B$ be the number of patients randomized to treatment $i = A, B$ when $j = 1, \dots, n$ patients have been assigned a treatment. Define an allocation function, $\phi(j)$, to be the probability that patient $j + 1$ is randomized to treatment A . Different definitions

of this function characterize different restricted randomization procedures proposed in the literature. Complete randomization is defined by

$$\phi(j) = 1/2, \quad \text{if } j = 0.$$

We now describe *forced balance* designs that force exactly $n/2$ subjects on each treatment, provided n is known and even. In the following we will present the definitions of $\phi(j)$ for $j \geq 1$.

7.2.1.1 Random Allocation Rule (RAR)

According to this rule, the probability for a patient to be randomized to a treatment is proportional to the difference between $n/2$ and number of patients already randomized to this treatment. The RAR is defined by:

$$\phi(j) = \frac{\frac{n}{2} - N_A(j)}{n - j}.$$

7.2.1.2 Truncated Binomial Design (TBD)

According to this design, treatment assignments follow independent Bernoulli trials with probability 1/2 until one treatment receives half of the patients, then all patients to be randomized are assigned to the other treatment. The TBD is defined by:

$$\begin{aligned}\phi(j) &= 1/2, && \text{if } \max\{N_A(j), N_B(j)\} < n/2, \\ &= 0, && \text{if } N_A(j) = n/2, \\ &= 1, && \text{if } N_B(j) = n/2.\end{aligned}$$

Both the RAR and TBD force exact balance at the termination of a randomization process and require knowledge of the total number of patients to be randomized. There can also be a positive probability of imbalance during the course of the trial, although it is corrected at the end. Because of this, a *permuted block design (PBD)* is often used by randomizing to fixed length blocks in which equal number of patients will be randomly assigned, either using the RAR or TBD within each block. If the total number of patients is unknown, blocks can be added until patients are exhausted, but this may lead to imbalance if there is an unfilled final block. Another drawback of the PBD is that many patients will be randomized with probability 1 (at the end of each block), which can induce selection bias.

To ensure that each patient is randomized with probability less than 1, various authors have proposed adaptive randomization procedures that are fully randomized, and also tend to balance throughout the course of the trial, although they are not guaranteed to provide perfect balance at the end.

7.2.1.3 Efron's Biased Coin Design (BCD)

Let $D_j = N_A(j) - N_B(j)$. Efron (1971) introduced an adaptive randomization procedure with an attempt to balance treatment assignments in the course of randomization. Specifically, with $p \in (0.5, 1]$,

$$\begin{aligned}\phi(j) &= 1/2, && \text{if } D_j = 0, \\ &= p, && \text{if } D_j < 0, \\ &= 1-p, && \text{if } D_j > 0.\end{aligned}$$

7.2.1.4 Wei's Urn Design (UD)

Wei (1978) proposed an urn model for adaptive randomization. An urn initially contains α balls of each of two types A and B . When a patient is to be randomized, a ball is drawn and returned to the urn. If the ball is of type A (B), the patient receives treatment A (B) and β type B (A) balls are added to the urn. This urn design is denoted $UD(\alpha, \beta)$, and,

$$\phi(j) = \frac{\alpha + \beta N_B(j)}{2\alpha + \beta j}.$$

7.2.1.5 Smith's Generalized Biased Coin Design

Smith (1984) generalized Wei's UD procedure and proposed the following allocation function:

$$\phi(j) = \frac{N_B(j)^\gamma}{N_A(j)^\gamma + N_B(j)^\gamma},$$

where γ is a constant.

For a detailed description of these procedures and their mathematical properties, refer to Rosenberger and Lachin (2002). Now we consider the effect of these procedures in balancing treatment assignments. Complete randomization can result in severe terminal imbalance and mid-trial imbalance. The RAR and TBD can lead to severe mid-trial imbalance. The BCD, UD, Smith's design, and the PBD can mitigate mid-course imbalance. Rosenberger and Lachin (2002) use a simulation study to compare mean and standard deviation of proportion of patients assigned to treatment A by some of these procedures. They confirm that complete randomization is not desirable but that other procedures are similar with regard to terminal balance only. Efron's BCD is the least variable procedure. Here we conduct a simulation study to compare proportion of patients assigned to treatment A throughout the randomization process. Figure 7.1 shows results of such a simulation study. For the PBD, we assume all blocks are filled and block size is 4 except for the last block with size 6. In Figure 7.1, the top two panels depict the trajectory of proportion of patients assigned to treatment A throughout the randomization process for different procedures. It is clearly seen from the left panel that severe imbalances can occur for complete randomization, the RAR and TBD. In the case of complete randomization, terminal imbalance obviously exists. The right panel depicts procedures that mitigate mid-trial imbalances. They work well after 10 patients have been randomized. The lower two panels are a replication of the same simulation and the same conclusions can be drawn.

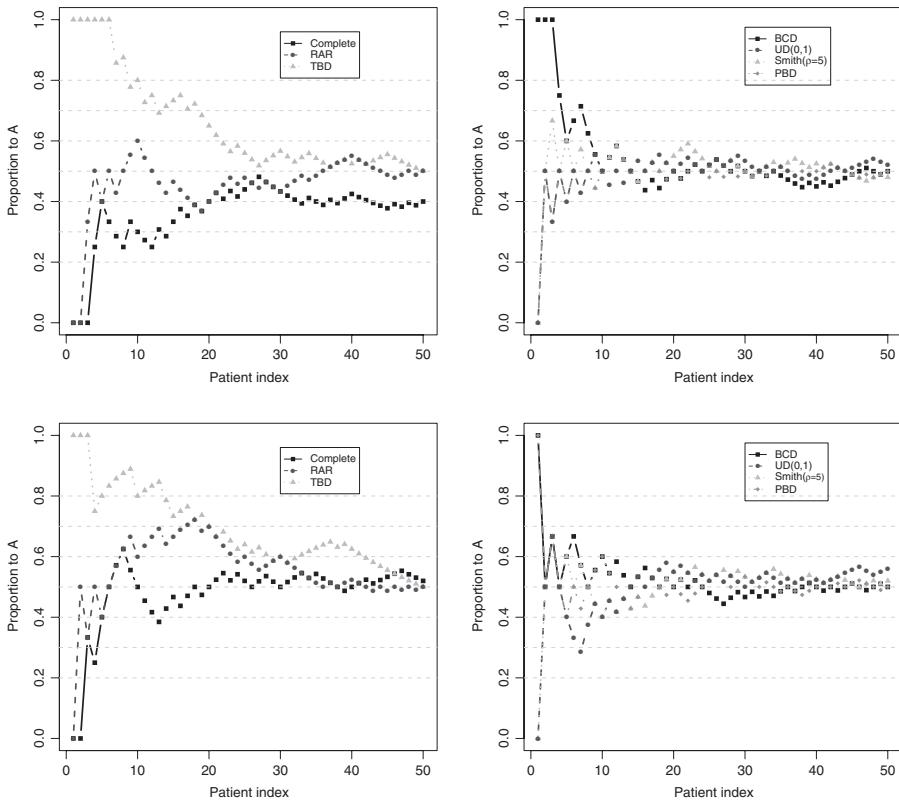


Figure 7.1 Simulation of proportion of patients assigned to treatment A with $n = 50$. The two rows correspond to two runs of the randomization process.

A fundamental question is why it is deemed important to balance treatment assignments. As we mentioned earlier, one argument is that if treatment assignment is balanced, then the test of treatment effect will be most *efficient*. This is a myth; in fact, this is true only when the variability of patient responses is the same between two treatment groups. Rosenberger and Sverdlov (2008) describe the trade-off among balance, efficiency, and ethics, and describe scenarios in which balanced treatment assignments in an unstratified trial or within strata are inefficient.

7.2.2 Covariate-Adaptive Randomization

In every clinical trial, an investigator will determine certain covariates of interest that may affect a subject's response to a treatment, such as demographic factors, severity of illness, or clinical center. If a covariate is known to have a significant effect on treatment responses, stratification is often used to ensure

that treatment assignments are equally distributed across the levels of the covariate. In fact, some baseline covariates are often considered “default” stratification factors in clinical trials regardless of their effect on treatment responses (e.g., clinical center, age, gender).

Each combination of a level of different factors is called a stratum. For example, if there are three age groups, six combinations of age and gender lead to six strata. When the number of strata is small, a simple way to achieve stratified randomization is to use restricted randomization within each stratum. However, in practice, factor level can easily multiply to an intractable number relative to total sample size of the trial. Many strata may end up with a few or no patients. In this case, applying restricted randomization within each stratum is not practical.

To address this issue, *covariate-adaptive randomization* has been proposed. Covariate-adaptive randomization can be considered as a generalization of restricted randomization. The idea is to balance treatment assignments at each stratum as far as possible. In case of many strata, a metric is invoked to assess the overall treatment imbalance. When the next patient is to be randomized, the metric is computed within covariate levels. Then a larger probability is applied to the treatment that is underrepresented. Zelen (1974) proposed the first method for covariate-adaptive randomization. Here we will discuss a procedure proposed by Pocock and Simon (1975).

The Pocock–Simon procedure is a generalization of Efron’s BCD. Let $N_{ijk}(n)$ be the number of patients at level j of covariate i assigned to treatment k after n patients have been randomized. If the $(n+1)$ th patient comes with level l_i for covariate i , let $D_i(n) = N_{il_i1}(n) - N_{il_i2}(n)$. Recall that in Efron’s BCD, we simply compute the treatment assignment difference over all randomized patients. Here we calculate the treatment assignment difference at the level associated with the $(n+1)$ th patient of each covariate. The next step is to define a metric that combines the $D_i(n)$ ’s. A natural metric is a weighted sum: $D(n) = \sum_i w_i D_{ilk}(n)$, where weight w_i for covariate i is chosen such that covariates deemed of greater importance should carry a larger weight. The allocation probability to treatment A is defined similarly to Efron’s design. Let $p \in [1/2, 1]$. The allocation function is given by

$$\begin{aligned}\varphi(n+1) &= 1/2, && \text{if } D(n) = 0, \\ &= p, && \text{if } D(n) < 0, \\ &= 1-p, && \text{if } D(n) > 0.\end{aligned}$$

Note that this procedure balances *marginally* across covariates instead of within each stratum. The implications of this, and a description of other covariate-adaptive procedures, can be found in Rosenberger and Sverdlov (2008). Pocock and Simon (1975) also generalize the procedure to more than two treatments.

Covariate-adaptive randomization procedure can be applied to balance treatment assignments over covariates when number of strata is so large that

stratified randomization is not practical. This procedure can be easily implemented using SAS. An SAS macro can be found in Kuznetsova and Ivanova (2007).

7.2.3 Response-Adaptive Randomization

As we mentioned earlier, restricted randomization is proposed to achieve treatment assignment balance, with the belief that the treatment effect can be tested most efficiently. However, this is true only when the variability of patient responses is the same between the two treatment groups. When the variability is not the same, unbalanced treatment assignments can often provide a more efficient test. On the other hand, we need to recall that “the fundamental ethical concern raised by clinical research is the fact that research subjects are exposed to risks for the benefit of others, thus raising concerns of exploitation” (Wendler 2009). If, in the course of randomization, a trend has shown that one treatment is performing better than the other, albeit not conclusively, then changing the allocation probability to favor the better treatment may mitigate the ethical concern somewhat. *Response-adaptive randomization* is proposed to address such ethical concerns, where the allocation probability for the next participant is updated using accumulated responses so that by the end of recruitment more participants will be allocated to the better-performing treatment arm.

There are several approaches to response-adaptive randomization. One intuitive and heuristic approach is through urn models. Different colors of balls in the urn correspond to different treatment arms. One updates the urn composition based on available participant outcomes and randomizes participants to a treatment according to the color of a ball drawn randomly from the urn. The adaptation of the randomization probability is reflected by updated urn composition. Typical urn models include the *randomized play the winner* (Wei and Durham 1978) and the *drop the loser* (Ivanova 2003).

But to what extent should the probability be skewed? If there is no treatment effect, then response-adaptive randomization should lead to equal allocation. If there is a treatment effect, as many patients should be randomized to the better performing arm as possible, from an ethical point-of-view. However, it has been shown that if skewing is too large, loss of power of test can be significant (e.g., Rosenberger and Hu 2004). Therefore, there is a trade-off between efficiency and ethics. Urn models, as discussed below, do not take into account power loss. Another approach, the optimal allocation approach, derives an optimal allocation proportion based on an optimization problem which minimizes some ethical cost with a constraint on preservation of power. A randomization procedure is then utilized to implement the optimal allocation at current values of the parameter estimates.

Response-adaptive randomization is covered in Rosenberger and Lachin (2002) and Hu and Rosenberger (2006). In this section, we will discuss response-adaptive randomization methods based on these two approaches. We

will focus on response-adaptive randomization without covariate adjustments. Note that in most trials, there is a delay in the response of a patient upon receiving a treatment. This delayed effect has been discussed in many papers, for example, Bai, Hu, and Rosenberger (2002), Zhang et al. (2007), and Zhang and Rosenberger (2006). In essence, as long as responses are not “too far out” relative to the patient recruitment stream, asymptotic properties of the procedure will hold. For simplicity, we will assume responses are immediately available after treatment assignments. We consider two treatments A and B with a fixed total sample size n .

7.2.3.1 Response-Adaptive Randomization Based on Urn Models

In response-adaptive randomization, the urn composition changes based on the response of the patient most recently randomized. We will give two examples of urn models that apply to clinical trials with binary outcomes. In the following, we assume patients receiving treatment A (B) have responses following an i.i.d. Bernoulli distribution with success probability p_A (p_B).

The randomized play-the-winner (RPW) rule was proposed by Wei and Durham (1978). The urn initially contains α balls of each of two colors, say black and white, corresponding to treatments A and B , respectively. Suppose the j th patient is to be randomized. A ball is drawn from and placed back into the urn. If its color is black (white), the patient is randomized to treatment A (B). The patient’s response is recorded. If it is a success (failure), then β balls of the same (opposite) color are added to the urn. Let $N_A(n)$ be the number of patients finally randomized to treatment A . It has been shown (Wei and Durham 1978) that

$$\lim_{n \rightarrow \infty} \frac{N_A(n)}{n} = \frac{q_B}{q_A + q_B},$$

where $q_i = 1 - p_i$. This implies that if the sample size is large enough, the number of patients randomized to a treatment is inversely proportional to the failure probability of patients randomized to the treatment. Thus, more patients will be randomized to the treatment with a higher success probability.

The drop-the-loser (DL) rule was proposed by Ivanova (2003). For this procedure, the urn contains balls of three colors: black, white, and red. The black and white balls represent treatments A and B , respectively; red balls are called immigration balls. Suppose the j th patient is to be randomized. A ball is drawn from the urn. If its color is red, then a black ball and white ball are added to the urn. If its color is black (white), the patient is randomized to treatment A (B). The patient’s response is recorded. If it is a success, the ball is returned to the urn. If it is a failure, the ball is not returned to the urn. There is the possibility that the urn will eventually contain no white or black balls. Then a red ball is drawn and black and white balls are added to the urn. It has been shown that the DL rule results in the same limiting allocation proportion as the RPW with smaller variance.

Although these two urn models are easy to implement, there are a few limitations. On the one hand, they can only be used for clinical trials with binary endpoints. On the other hand, if q_B is much larger than q_A , then the allocation probability to treatment A could be very large. Such extreme skewing to a treatment can induce significant loss of power, as we discussed before.

7.2.3.2 Optimal Allocation Approach

In this approach, a target allocation proportion is first determined based on an optimization problem. Then a randomization procedure is used that targets the optimal allocation. Let $Y_i, i = A, B$ denote response of a patient randomized to treatment i and $\eta_i = E(Y_i)$. Let $f_i(\eta_A, \eta_B)$ be ethical cost functions of the parameters. In the following, we consider testing the hypotheses

$$H_0 : \eta_A = \eta_B \text{ versus } H_1 : \eta_A \neq \eta_B.$$

We will use Wald test

$$\frac{\hat{\eta}_A - \hat{\eta}_B}{\sqrt{\text{var}(\hat{\eta}_A - \hat{\eta}_B)}},$$

where $\hat{\eta}_i$ is the maximum likelihood estimator of η_i .

The optimization problem takes into account the ethical cost and power of the test. Let n_A be the number of patients desired to be allocated to treatment A . Similarly to Jennison and Turnbull (2000), the optimization problem is given by

$$\begin{cases} \min & n_A f_A(\eta_A, \eta_B) + (n - n_A) f_B(\eta_A, \eta_B) \\ \text{subject to} & \text{var}(\hat{\eta}_A - \hat{\eta}_B) = C_0, \end{cases}$$

where C_0 is a constant. Solving this problem, we can obtain an optimal allocation proportion $\rho = n_A/n$, which does not depend on C_0 in the scenarios considered in the following.

Table 7.1 lists the optimal allocation proportions based on different ethical metrics for clinical trials with binary, continuous, and survival outcomes. For each of the three types of outcomes, the first ethical metric function is $f = 1$, therefore total sample size is minimized. The resultant allocation proportion is also called Neyman allocation. In the case of binary outcomes, the second metric function is the failure rate, thus total expected number of failures is minimized. In case of continuous outcomes, the second ethical metric function is the mean response, thus the total expected responses of all patient are minimized under an assumption that a smaller response is more desirable. The resulting allocation proportion from this optimization problem is modified such that at least 50% patients be randomized to the treatment with a smaller mean response. The third ethical metric function is the probability of a patient's response exceeding C , thus the total expected number of patients with response larger than C is minimized. This dichotomizes a continuous response into a binary one. In the case of survival outcomes, the second ethical metric function is the hazard rate. The third ethical metric function is the probability that a

Table 7.1 Optimal Allocation Proportions for Different Trial Outcomes

Outcome	$f_i(\eta_A, \eta_B)$	Allocation Proportion	Reference
Binary ($Y_i \sim Bernoulli(\eta_i)$)	1	$\frac{\sqrt{\eta_A(1-\eta_A)}}{\sqrt{\eta_A(1-\eta_A)} + \sqrt{\eta_B(1-\eta_B)}}$	Rosenberger and Lachin (2002)
	$1 - \eta_i$	$\frac{\sqrt{1-\eta_B}}{\sqrt{1-\eta_A} + \sqrt{1-\eta_B}}$	Rosenberger et al. (2001)
Continuous ($Y_i \sim N(\eta_i, \sigma_i^2)$)	1	$\frac{\sigma_A}{\sigma_A + \sigma_B}$	Rosenberger and Lachin (2002)
Survival ($Y_i \sim \exp(\eta_i)$)	η_i	$\max\left\{\frac{1}{2}, \frac{\sigma_A \sqrt{\eta_B}}{\sigma_A \sqrt{\eta_B} + \sigma_B \sqrt{\eta_A}}\right\}$ if $\eta_A \leq \eta_B$ $\min\left\{\frac{1}{2}, \frac{\sigma_A \sqrt{\eta_B}}{\sigma_A \sqrt{\eta_B} + \sigma_B \sqrt{\eta_A}}\right\}$ if $\eta_A > \eta_B$	Zhang and Rosenberger (2006)
	$\Phi\left(\frac{\eta_i - C}{\sigma_i}\right)$	$\frac{\sqrt{\Phi\left(\frac{\eta_B - C}{\sigma_B}\right)\sigma_A}}{\sqrt{\Phi\left(\frac{\eta_B - C}{\sigma_B}\right)\sigma_A} + \sqrt{\Phi\left(\frac{\eta_A - C}{\sigma_A}\right)\sigma_B}}$	Biswas and Mandal (2004)
	1	$\frac{\eta_A \sqrt{\varepsilon_B}}{\eta_A \sqrt{\varepsilon_B} + \eta_B \sqrt{\varepsilon_A}}$	Zhang and Rosenberger (2007)
	$\frac{1}{\eta_i}$	$\frac{\sqrt{\eta_A^3 \varepsilon_B}}{\sqrt{\eta_A^3 \varepsilon_B} + \sqrt{\eta_B^3 \varepsilon_A}}$	Zhang and Rosenberger (2007)
	$1 - e^{-C/\eta_i}$	$\frac{\eta_A \sqrt{\varepsilon_B (1 - e^{-C/\eta_B})}}{\eta_A \sqrt{\varepsilon_B (1 - e^{-C/\eta_B})} + \eta_B \sqrt{\varepsilon_A (1 - e^{-C/\eta_B})}}$	Zhang and Rosenberger (2007)

It is assumed here that a smaller response is desirable for continuous outcomes. For survival outcomes, ε_i is the probability of the i th patient's death.

patient's survival time is less than C . In the survival case, the exponential survival function is assumed. The event rate ε_i depends on the hazard rate, as well as censoring time. In Zhang and Rosenberger (2007), different censoring schemes are considered and the formulas for ε_i are derived.

An optimal allocation proportion gives the desirable proportion of participants to be randomized to a treatment. From Table 7.1, one notes that all the allocation proportions are dependent on unknown parameters. A natural way to randomize patients using these proportions is to replace the unknown parameters with their estimates based on available responses. We then require a rule that will skew the allocation probabilities accordingly. Procedures for response-adaptive randomization based on optimal allocation proportions are

a generalization of restricted randomization procedures. In fact, the optimal allocation proportion for restricted randomization can be considered as 1/2; the update of the allocation function for these procedures does not depend on responses of randomized patients. In response-adaptive randomization procedures, the allocation probability for the next patient is a function of the optimal allocation proportion. In the following, we consider several such procedures. All these procedures can be used for all scenarios in Table 7.1. Note that the drop the loser and randomized play-the-winner rules cannot target an optimal allocation proportion.

The real-valued urn model (RVUM) was proposed by Yuan and Chai (2008). This urn model is designed to target any desired allocation proportion and can be described as follows. Suppose the current urn composition is (α_A, α_B) . The probability for the next patient to be randomized to treatment i is $\alpha_i/(\alpha_A + \alpha_B)$. Then the patient's response is observed. The urn composition is updated by adding some real numbers to (α_A, α_B) , and the increments depend on all available responses. By carefully selecting the increments, a specific optimal allocation proportion can be targeted. Since (α_A, α_B) and their increments can take real values, these procedures are called *real-valued urn models*.

Here we give an example of RVUM targeting the optimal allocation proportion of Rosenberger et al. (2001), denoted by $R\rho$ from this point on. Suppose the initial urn composition is $(\alpha_{0A}, \alpha_{0B})$. After j patients, the urn composition is $(\alpha_{jA}, \alpha_{jB})$. Randomize the $(j+1)$ th patient ($j \geq 1$) to treatment i with probability $\alpha_{jA}/(\alpha_{jA} + \alpha_{jB})$. The $(j+1)$ th patient's response is observed. Let $\hat{R}\rho_{j+1}$ be the estimate of $R\rho$ based on responses of the first $j+1$ patients. Update the urn composition to $(\alpha_{jA} + \hat{R}\rho_{j+1}, \alpha_{jB} + 1 - \hat{R}\rho_{j+1})$. The process is repeated until all patients have been randomized. Note that the probability for the $(j+1)$ th patient to be randomized to treatment A is given by:

$$\frac{\alpha_{0A} + \sum_{k=0}^j \hat{R}\rho_k}{\alpha_{0A} + \sum_{k=0}^j \hat{R}\rho_k + \alpha_{0B} + \sum_{k=0}^j (1 - \hat{R}\rho_k)}.$$

If $\alpha_{0i} = 0$, then the above probability is given by

$$\frac{\sum_{k=1}^j \hat{R}\rho_k}{j},$$

which is the average of the first j sequential estimates of $R\rho$.

The doubly-adaptive biased coin design (DBCD) was proposed by Hu and Zhang (2004), based on Eisele (1994). Suppose after j patients, the estimate of optimal allocation proportion ρ is $\hat{\rho}_j$, and $N_A(j)$ patients are randomized to treatment A . In the DBCD procedure, an allocation function defines the probability of next patient to be randomized to patient A , which is larger than the target ρ if $N_A(j)/j < \hat{\rho}_j$ and smaller than ρ if $N_A(j)/j > \hat{\rho}_j$. Ultimately, the estimates of ρ and $N_A(j)/j$ become very close, and both should converge to ρ .

as $j \rightarrow \infty$. Let $x = N_A(j)/j$ and $y = \hat{\rho}_j$. The allocation function $g(x, y)$ given by Hu and Zhang (2004) is defined as

$$g(x, y) = \begin{cases} 0 & \text{if } x > y, \\ \frac{y\left(\frac{y}{x}\right)^\gamma}{y\left(\frac{y}{x}\right)^\gamma + (1-y)\left(\frac{1-y}{1-x}\right)^\gamma} & \text{if } 0 < x < 1, \\ 1 & \text{if } x = 0, \end{cases}$$

where $\gamma \geq 0$ is a tuning parameter controlling the degree of randomness of the procedure. The $(j+1)$ th patient is randomized to treatment A with probability $g(N_A(j)/j, \hat{\rho}_j)$. When $\gamma = 0$, it becomes the *sequential maximum likelihood procedure* of Melfi, Page, and Gerald (2001). This procedure can be applied to target all allocation proportions in Table 7.1. Note that when $y = 1/2$, $g(x, y)$ reduces to Smith's allocation function in Section 7.2.1. Simulation evidence shows that $\gamma = 2$ is a reasonable value (Hu and Zhang 2004).

The efficient randomized adaptive design (ERADE) was proposed by Hu, Zhang, and He (2009). The ERADE uses the following allocation function:

$$g(x, y) = \begin{cases} \gamma & \text{if } x > y, \\ y & \text{if } x = y, \\ 1 - \gamma(1-y) & \text{if } x < y, \end{cases}$$

where $0 \leq \gamma < 1$ is a tuning parameter for degree of randomness. When $\gamma = 0$, this procedure is deterministic. A larger γ leads to a more random procedure. Also note that when $\rho = 1/2$, this procedure becomes Efron's biased coin design (Efron 1971).

7.2.3.3 Properties of Randomization Procedures Targeting Optimal Allocation Proportions

A basic requirement for a response-adaptive randomization procedure is consistency, that is, $N_A(n)/n \rightarrow \rho$ as $n \rightarrow \infty$. The three response-adaptive randomization procedures presented above meet this requirement. Very importantly, they can be applied to all types of outcomes and target different allocation proportions.

When choosing an optimal allocation or a procedure, what should one consider? Hu and Rosenberger (2003) present a template that relates the power of the test, the allocation proportion, and the variability of N_k/n for binary outcomes. This template allows one to compare asymptotic properties of different randomization procedures targeting the same allocation proportion, or compare asymptotic properties of different allocation proportions using the same randomization procedure. Zhang and Rosenberger (2006) discuss such a relationship for continuous outcomes. Zhang and Rosenberger (2007) evaluate these procedures for survival outcomes.

In practice, we suggest extensive simulation be conducted to understand performance of different allocation proportions and randomization procedures. In the following, we present a simulation study to demonstrate how to choose one of the three procedures that target the same proportion $R\rho$ for binary outcomes. In the study, different combinations of success rates are chosen for two treatments. The sample size n is determined such that a test of null hypothesis $\eta_A - \eta_B = 0$ has power of 90% with type I error rate 0.05. The test is given by

$$\frac{\hat{\eta}_A - \hat{\eta}_B}{\sqrt{\bar{\eta}(1-\bar{\eta})n/N_A(n)/N_B(n)}},$$

where $\bar{\eta} = (\hat{\eta}_A + \hat{\eta}_B)/2$.

The result of the simulation with 10,000 replication is shown in Table 7.2. Much information is revealed from Table 7.2 regarding these three procedures. First, RVUM converges much faster than the other two procedures. This is more clearly seen when the success rate difference is large, so that the sample size n required for 90% power is small. For example, in the first row of Table 7.2 with sample size of 24, the target (theoretical) allocation proportion should be 0.63. The RVUM procedure produces an average 0.67. However, the DBCD and ERADE procedures give 0.76 and 0.81, respectively. As n gets larger, such as in rows 4 and 5, the RVUM and DBCD procedures lead to an average empirical proportion very close to the target. On the other hand, the average empirical proportion from ERADE procedure is still far away from the target value. Second, in most cases, the ERADE procedure gives rise to the smallest standard deviation of the average empirical proportion. Generally, the DBCD procedure has a smaller standard deviation of average empirical proportion than the RVUM. Third, the DBCD procedure gives the most powerful test and the ERADE procedure gives the least powerful test.

These observations clearly demonstrate the effect of the interplay between skewing to the better performing arm (or average empirical allocation proportion) and its variability on the power of the test, as shown in the template of Hu and Rosenberger (2003). The standard deviation of average empirical proportion of the ERADE procedure is smallest, but this advantage is overridden by its extreme skewing to the better treatment (due to slow convergence), leading to a significant loss of power. When selecting an optimal allocation proportion or randomization procedure, such a simulation is necessary to make an optimal choice.

7.3 LIKELIHOOD-BASED INFERENCE

The Wald test was proposed to be used following response-adaptive randomization clinical trials in Section 7.2.3. Is it appropriate to use such standard

Table 7.2 Power, Mean, and Standard Deviation (SD) of N_A/n of Three Randomization Procedures Targeting $R\rho$ for Clinical Trials with Binary Outcomes (10,000 Replications)

η_A	η_B	$R\rho$	n	RVUM		DBCD($\gamma=2$)		ERADE($\gamma=2/3$)	
				Power	Mean	SD	Power	Mean	SD
0.9	0.3	0.63	24	0.8666	0.67	0.15	0.8844	0.76	0.16
0.9	0.5	0.57	50	0.8826	0.60	0.13	0.8986	0.63	0.13
0.9	0.7	0.53	162	0.8996	0.54	0.06	0.9070	0.53	0.03
0.9	0.8	0.51	532	0.9025	0.52	0.03	0.8995	0.51	0.01
0.7	0.3	0.60	62	0.8754	0.65	0.15	0.8956	0.69	0.15
0.7	0.5	0.54	248	0.8939	0.55	0.08	0.9036	0.55	0.03
0.5	0.4	0.53	1036	0.8988	0.53	0.04	0.9018	0.53	0.01
0.3	0.1	0.63	158	0.8503	0.68	0.16	0.8833	0.74	0.17
0.2	0.1	0.59	532	0.8863	0.62	0.13	0.9040	0.63	0.11

statistical tests if adaptive randomization is used? In this section, we consider likelihood-based inference following adaptive randomization.

7.3.1 Restricted Randomization

In restricted randomization, treatment assignments do not depend on patient responses. It has been shown (Rosenberger and Lachin 2002) that if we assume responses of patients receiving the same treatments are a random sample from a homogeneous population, then the randomization mechanism is ancillary to the likelihood based on such a population model. Therefore, the usual statistical tests and estimation methods can be routinely used.

7.3.2 Covariate-Adaptive Randomization

A regression model is usually used for analysis of data from clinical trials following covariate-adaptive randomization. In such trials, treatment assignments, conditional on covariates used for stratification, do not depend on patient responses. However, care should be taken when constructing a test of the treatment effect. Shao, Yu, and Zhong (2010) investigate situations where the usual Wald-type normal test or t -test of the treatment effect is valid in the sense of preservation of type I error rate. They state that a sufficient condition for the usual tests to be valid is that they incorporate covariates used in the randomization. This implies that as long as the covariates used in randomization adaptation are also included as regressors, the usual Wald test or t -test should be still valid. In situations where covariates for randomization adaptation are not included as regressions, they find that the usual tests are conservative with a smaller type I error rate. They propose a bootstrap procedure that can correct the conservatism of the test.

7.3.3 Response-Adaptive Randomization

In response-adaptive randomization, treatment assignments do depend on available patient responses. Therefore, patient responses from the same treatment are correlated and the randomization procedure is not ancillary. In fact, from the formulation of response-adaptive randomization, there is information in the number of patients assigned to each treatment. Therefore, the properties of likelihood based analysis for clinical trials using response-adaptive randomization should be carefully examined.

First, we consider the maximum likelihood estimator (MLE) of parameters of interest. Under very mild regularity conditions, Hu and Rosenberger (2006) show that the MLE is still strongly consistent and follows an asymptotically normal distribution. These conditions are essentially a requirement that the distribution of patient responses belong to an exponential family and the empirical allocation proportion $N_A(n)/n$ strongly converges to a constant (which may be dependent on unknown parameters). As a consequence, all

Table 7.3 Type I Error Rate (Nominal $\alpha = 0.05$) of the Wald Test Following Three Randomization Procedures Targeting $R\rho$ for Clinical Trials with Binary Outcomes (10,000 Replications)

n	$\eta_A = \eta_B$	RVUM	DBCD ($\gamma = 2$)	ERADE ($\gamma = 2/3$)
200	0.9	0.0541	0.0500	0.0533
	0.7	0.0546	0.0473	0.0526
	0.5	0.0542	0.0553	0.0541
	0.3	0.0631	0.0831	0.0681
	0.1	0.0978	0.0912	0.0529
400	0.9	0.0492	0.0509	0.0465
	0.7	0.0499	0.0485	0.0493
	0.5	0.0533	0.0495	0.0500
	0.3	0.0536	0.0558	0.0509
	0.1	0.0736	0.1661	0.1061
1000	0.1	0.0537	0.0522	0.0491

response-adaptive randomization procedures we discussed so far admit an asymptotically standard normal Wald-type test.

These results are true asymptotically. To check finite sample properties, we conduct a simulation study to check whether the type I error rate ($\alpha = 0.05$) is well controlled with finite sample size. Table 7.3 presents the simulation results for three randomization procedures targeting optimal allocation proportion $R\rho$ by Rosenberger et al. (2001). In the simulation, sample sizes of $n = 200$ or 400 are used. The type I error rate is very close to the nominal level even when $n = 200$. When $n = 400$, they are closer except when success rate is low ($= 0.1$). When we increase the sample size to 1000 in the latter case, the type I error rate is very close to the nominal 0.05. We can also see that there is little difference among the three randomization procedures.

In the last section, we have shown through a simulation study that power of the standard Wald test following a response-adaptive randomization tends to be higher for the DBCD procedure. Putting these results together, if we use the DBCD procedure targeting $R\rho$ for clinical trials with binary outcomes, the design considerations should be no different than for a traditional design, with only a minor change to the allocation proportions.

Although it is shown that the MLE after a response-adaptive randomization trial is consistent and follows an asymptotically normal distribution, the estimator is biased (Coad and Ivanova 2001). They provide a method for correction that works well for small sample sizes. The bias of MLE in clinical trials with continuous or survival outcomes has not been studied in the literature to our knowledge. Confidence interval estimates have been proposed for response-adaptive randomization clinical trials with binary and survival outcomes (Coad and Woodroffe 1997, and Coad and Govindarajulu 2000). Rosenberger and Hu (1999) propose a bootstrap procedure to obtain

confidence intervals for binary outcomes. These parametric bootstrap procedures are very general and should be applicable for continuous and survival outcomes. In summary, the asymptotic theory based on likelihood following a response-adaptive randomization trial makes it possible to design and analyze the trial in a similar manner to traditional trials.

Of course, the key assumption for likelihood-based inference for restricted randomization, covariate-adaptive randomization and response-adaptive randomization is that patient responses from the same treatment form a random sample. Since clinical trials do not follow a population model with a random sampling mechanism, a randomization test is preferable, where the randomization procedure must be taken into account; we will describe randomization tests in Section 7.4.

7.3.4 Asymptotically “Best” Procedures

Suppose we use response-adaptive randomization procedures to target the same allocation proportion. Different procedures can give rise to different variability of $N_A(n)/n$. In the template of Hu and Rosenberger (2003), it has been shown that asymptotically, the larger this variance is, the less powerful is the test. Hu, Rosenberger, and Zhang (2006) define the asymptotically best procedure as one that attains a lower bound on the asymptotic variance of $N_A(n)/n$.

We use the binary case to formally define the concept. We will use $p_i, i = A, B$ to denote the success rate of patients receiving treatment i . Suppose the target proportion is ρ , a function of $\mathbf{p} = (p_A, p_B)$. Suppose the following regularity conditions hold,

1. $\mathbf{p} \in (0, 1)^2$;
2. $N_A(n)/n$ converges to ρ almost surely; and
3. $\sqrt{n}((N_A(n)/n) - \rho) \rightarrow N(0, v)$ in distribution.

Then

$$v \geq \left(\frac{\partial \rho}{\partial \mathbf{p}} \right)' I^{-1}(\mathbf{p}) \frac{\partial \rho}{\partial \mathbf{p}}, \quad (7.1)$$

where $I^{-1}(\mathbf{p})$ is the Fisher information. This definition can be readily generalized to any allocation proportion for any outcomes.

Any adaptive randomization procedure giving rise to asymptotic variance of $N_A(n)/n$ equal to the right hand side of Equation (7.1) is defined as an *asymptotically best randomization procedure* for the allocation proportion ρ . For restricted randomization, any procedure with a degenerate asymptotic variance would be asymptotically best for targeting 1/2. Efron’s biased coin

design satisfies this criterion. Among fully randomized procedures that do not have a degenerate limit, one can compare the asymptotic variances; Smith's generalized biased coin design, as γ gets larger, will have a smaller variance. Among urn models targeting $\rho = q_B/(q_A + q_B)$, Ivanova's drop-the-loser procedure is asymptotically best. Hu, Zhang, and He (2009) have shown the ERADE procedure, as a generalization of Efron's BCD, is asymptotically best for any allocation proportion if the MLE of unknown parameters in the allocation proportion is used in the course of randomization.

It is useful to remember that this property is an asymptotic one. In finite samples, an asymptotically best procedure does not always have smaller variance for $N_A(n)/n$. For example, in the simulation study in Table 7.2, the ERADE procedure led to larger power loss compared with the other two response-adaptive randomization procedures, in spite of having the smallest variance for $N_A(n)/n$.

7.4 RANDOMIZATION-BASED INFERENCE

Randomization provides a basis for inference. While this has been well-known, and has been one of the frequently cited advantages of randomization (see e.g., HK1, chapters 5 and 6), randomization tests are rarely used in practice. In this section, we describe randomization as a basis for inference and how it is related to adaptive randomization.

7.4.1 Randomization Tests

The very act of randomization provides a basis for inference. Let $\mathbf{X} = (X_1, \dots, X_n)$ be the responses based on some primary outcome variable and let \mathbf{x} be the realization. Then \mathbf{x} is a set of sufficient statistics under the null hypothesis of no treatment effect under any underlying distribution. Also, under the null hypothesis, \mathbf{x} is exchangeable with respect to treatment groups. By conditioning on the entire data set as a sufficient statistic and applying the exchangeability condition, a valid inference procedure is obtained by permuting the randomization vector $\mathbf{T} = (T_1, \dots, T_n)$, where $T_i = 1$ or 0 , in all possible ways. Each permutation carries a specific probability depending on the particular randomization sequence. A metric is established to measure the treatment effect, and a randomization p -value is obtained by adding the probabilities of sequences that yield a more extreme result than the observed sequence.

The concept of *permutation tests* is well-established and even included in some standard statistical software packages. Permutation tests rely on the same arguments of sufficiency and exchangeability, and are widely used following nonrandomized and even undesigned experiments, where the exchangeability among groups is not as clear-cut as in a randomized experiment; see Pesarin (2001). As a matter of terminology, we refer to randomization tests as permutation tests where permutations are made with respect to a particular

randomization sequence following random group assignments. While any metric depicting the treatment effect can be used with randomization-based inference, the family of *linear rank tests* provides a large class of tests with which to conduct randomization-based inference. The form of the test is $S_n = \mathbf{a}'_n \mathbf{T}$, for a score vector $\mathbf{a}_n = (a_{1n} - \bar{a}_n, \dots, a_{nn} - \bar{a}_n)'$, where a_{jn} is some function of the rank of x_j among the elements of \mathbf{x} . We can illustrate the breadth of this family with just a few score functions. For a clinical trial with binary outcomes, using *binary scores*, where $a_{jn} = 1$ or 0, yields the randomization-based equivalent of the Pearson's chi-square test. For a clinical trial with continuous outcomes, using the *simple rank scores*, given by $a_{jn} = r_{jn}/(n + 1) - 0.5$, where r_{jn} are the integer ranks, leads to the randomization-based equivalent of the Wilcoxon rank-sum test. For survival data, using *Savage scores*, given by $a_{jn} = E(X_{(j)}) - 1$, where $X_{(1)}, \dots, X_{(n)}$ are order statistics from unit exponential random variables, yields a randomization-based equivalent of the logrank test when there are no ties or censoring. One can also incorporate ties and censoring into the score function (Prentice 1978) using a censoring indicator.

The exact distribution of randomization tests is largely intractable for the sample sizes required by typical clinical trials. The simplest Monte Carlo approach would be to generate K sequences according to ϕ , the allocation function, and use the Monte Carlo frequency distribution as the probability of that sequence. Computing the test statistic with each sequence, the Monte Carlo p -value is then the proportion of sequences yielding a more extreme test statistic. This is the approach using the *unconditional reference set*. However, the unconditional reference set contains sequences that give little or no information about the treatment effect (e.g., AAAA ... AA). For that reason, the *conditional reference set* is often used, which finds probabilities conditional on $N_A = n_A$, that is, the observed number on treatment A . This presents a computationally more difficult problem. For covariate-adaptive randomization, the sequence of covariates must also be taken into account, and there is no consensus on how this should be done.

Under certain randomization procedures and certain score functions, $\mathbf{a}'_n \mathbf{T}$ is asymptotically normal under the null hypothesis of no treatment difference (e.g., HK1, chapter 6). In this case, a standard normal asymptotic test statistic can be formed as

$$\frac{\mathbf{a}'_n \mathbf{T}}{\sqrt{\mathbf{a}'_n \Sigma_T \mathbf{a}_n}}.$$

Provided Σ_T can be computed, this provides a very convenient large sample test that is simple to compute. Unfortunately, Efron's biased coin design (and its extension to Pocock–Simon's procedure) does not provide an asymptotically normal test statistic. For the conditional version of the test, let $\Sigma_{T|n_A} = \text{Var}(\mathbf{T} | N_A = n_A)$. If $\mathbf{a}'_n \mathbf{T}$, conditional on $N_A = n_A$, is asymptotically normal, then the conditional test, given by

$$\frac{\mathbf{a}'_n \mathbf{T}}{\sqrt{\mathbf{a}'_n \Sigma_{T|n_A} \mathbf{a}_n}},$$

is also asymptotically normal.

For response-adaptive randomization, \mathbf{x} is not sufficient, as the treatment assignments also contain information about the responses. Consequently, while Rosenberger (1993) demonstrates the asymptotic normality of the linear rank test under response-adaptive randomization conditioning on \mathbf{x} , the test does not have inferential validity using a sufficiency argument.

7.4.2 Monte Carlo Unconditional Tests

Suppose we generate L randomization sequences using Monte Carlo simulation. Then $S_l = \mathbf{a}' \mathbf{T}_l$ is computed for the l th generated sequence, $l = 1, \dots, L$. Let $S_{obs.}$ be the observed value of the test statistic, and let $I(\cdot)$ be the indicator function. For an unconditional test, the two-sided Monte Carlo p -value is then defined as

$$\hat{p}_u \sim \frac{\sum_{l=1}^L I(|S_l| \geq |S_{obs.}|)}{L}.$$

For restricted randomization, the key component of this computation is that disparate probabilities of sequences will be depicted by the number of identical sequences generated. To determine the number of sequences L needed for accurate estimation of p_u , we note that there are 2^n sequences each with a computed test statistic S_l , and we order them. Under H_0 , $S_{obs.}$ can be located anywhere in the ordering. Thus whether or not S_l is extreme, it is distributed as Bernoulli with underlying probability p_u , and hence \hat{p}_u is unbiased with

$$MSE(\hat{p}_u) = \frac{p_u(1-p_u)}{L}.$$

Then establishing a bound $MSE(\hat{p}_u) < \varepsilon$ implies that $L > 1/4\varepsilon$. For $\varepsilon = 0.0001$, we have $L > 2500$. This seems to be a reasonable approach to the problem.

7.4.3 Monte Carlo Conditional Tests

The conditional test is considerably more intensive than unconditional tests. We wish to condition the test on $N_A = n_{A,obs.}$, the observed number of patients assigned to treatment A . Let $N_{A,k}$ be the number of patients assigned to A in the k th generated sequence, $k = 1, \dots, K$, where $K > L$ is the number of Monte Carlo samples required. The conditional p -value is given by

$$\hat{p}_c \sim \frac{\sum_{k=1}^K I(S_k \geq S_{obs.}, N_{A,k} = n_{A,obs.})}{\sum_{k=1}^K I(N_{A,k} = n_{A,obs.})}.$$

We wish K to be selected so that there will be a sufficient number of sequences satisfying the constraint $N_{A,k} = n_{A,obs}$. In particular, we would like at least L/K of the sequences to satisfy the constraint, in order to have a reasonable MSE. The distribution of K is then negative binomial with parameters $\pi = \Pr(N_A = n_{A,obs})$ and $r = L$. Then we can compute

$$E(K) = \frac{r}{\pi} = \frac{L}{\Pr(N_A = n_{A,obs})}$$

and

$$\text{Var}(K) = \frac{L\{1 - \Pr(N_A = n_{A,obs})\}}{[\Pr(N_A = n_{A,obs})]^2} \sim \frac{L}{[\Pr(N_A = n_{A,obs})]^2}.$$

The latter approximation applies only when the denominator in the expectation is very small. Relevant quantiles can also be found. For Efron's BCD, we can compute $\Pr(N_A = n_{A,obs})$ exactly using the formula in Markaryan and Rosenberger (2010). For complete randomization,

$$\Pr(N_A = n_{A,obs}) = \frac{\binom{n}{n_{A,obs}}}{2^n},$$

and hence

$$E(K) = \frac{L2^n}{\binom{n}{n_{A,obs}}}, \quad \text{Var}(K) \sim \frac{L2^{2n}}{\binom{n}{n_{A,obs}}^2}.$$

We replace L by its lower bound $1/4\epsilon$. Suppose $n_{A,obs} = \delta n$, where $\delta \in (0, 1)$. The average value, $E(K)$, will mean that we will only have sufficient sample size on average. Hence, it would be more appropriate to find the value of K at the 95th percentile. Table 7.4 gives values for several values of n and δ when $\epsilon = 0.0001$. One can see that as n gets large, the computing requirements become substantial, especially when a fairly rare sequence is observed.

Table 7.4 Approximate 95th Percentile of K for Various $n, \delta, \epsilon = 0.0001$ (Complete Randomization)

n	$\delta = 0.45$	$\delta = 0.48$	$\delta = 0.50$
100	53,271	35,120	32,445
200	124,155	53,738	45,827
500	881,321	107,939	72,404
1000	15,243,755	227,692	102,369

It is important here to reiterate that these values refer only to complete randomization. The requisite number of Monte Carlo samples is highly dependent on the type of randomization procedure employed.

7.4.4 Expanding the Reference Set

The rationale for conditional tests is that sequences that give very little information on the treatment effect, and those that are very dissimilar in composition to the observed treatment effect, should be treated as ancillary statistics. But conditioning on sequences with exactly the same numbers of A and B assignments is overly restrictive, and often requires a very large Monte Carlo sample size, as seen in the previous section. *Provided the reference set is pre-specified before the study*, there is no reason that it cannot be expanded to include sequences that are close in composition, but not necessarily identical, to the observed sequence. This idea was probably first noted by Cox (1982) in an obscure paper, where he states “... we should take the randomization distribution not over all designs but only over those arrangements with the same or nearly the same terminal lack of balance.”

For example, we can condition on the expanded set $\{N_A \in (n_{A,obs} \pm \gamma n)\}$, where $\gamma \in (0, 1)$. Then the expected value of K can be written as

$$E(K) = \frac{L}{\Pr[N_A \in (n_{A,obs} \pm \gamma n)]}.$$

Again, this can be computed for Efron’s BCD, and is an open problem for the DBCD. Table 7.5 gives the 95th percentile of K under complete randomization when $\gamma = 0.05$, replacing L by $1/4\varepsilon$. The computational burdens are considerably reduced, even by taking sequences within only 5% of the composition of the observed. Hence, there should be no computational difficulty using this method.

7.4.5 Stratified Tests

Following stratification on known covariates, the computation of a stratified linear rank test based on the randomization distribution is straightforward by

Table 7.5 Approximate 95th Percentile of K for Various n , δ , $\varepsilon = 0.0001$ When $\gamma = 0.05$ (Complete Randomization)

n	$\delta = 0.45$	$\delta = 0.48$	$\delta = 0.50$
100	4897	3683	3490
200	4858	3167	2934
500	4939	2754	2571
1000	4989	2583	2507

summing the stratum-specific tests over independent strata. Using the Monte Carlo approach for the unconditional test, one simply needs to generate L randomization sequences and compute S_{li} , $l = 1, \dots, L$, $i = 1, \dots, I$ within stratum i , which is then summed from $i = 1, \dots, I$ across strata of the linear rank tests computed within stratum.

The conditional test is considerably more complicated. Suppose the total sample size in each stratum is given by n_1, \dots, n_I . Let N_{Ai} be the number of patients assigned to A in the i th stratum, and let $n_{Ai,obs}$ be the observed number. Then we need to generate K^* sequences such that $\{N_{A1} = n_{A1,obs}, \dots, N_{AI} = n_{AI,obs}\}$. This is a much more restrictive set than in the unconditional case, and will necessarily require a much larger Monte Carlo sample. If we take the approach in Section 7.4.2 of enlarging the set to $\{N_A \in n_{Ai,obs} \pm \gamma n_i, i = 1, \dots, I\}$, then we require

$$E(K^*) = \frac{L}{\prod_{i=1}^I \Pr\{N_{Ai} \in (n_{Ai,obs} \pm \gamma n_i)\}}.$$

As a simple example, suppose we have 8 strata with the following proportions assigned to treatment A : (10/19, 12/24, 22/48, 6/10, 10/20, 9/14, 11/21, 16/32). Note that six of the strata are close to balance, and strata 3 and 6 deviate substantially. We compute the 95th percentile of K^* for $\gamma = 0.05$ and 0.10 in Table 7.6, and then recompute when we multiply sample sizes by 2, 5, and 10. Note that by doubling γ in this context, we substantially reduce the required number of samples to an easily computable problem.

7.4.6 Regression Modeling

While a stratified test is the appropriate technique following stratified randomization, it is often desirable to adjust for covariates in the analysis if prestratification is not used, as discussed in HK1, chapter 8. One can use regression modeling to develop treatment effect metrics from residuals. We will take the approach of Gail, Tan, and Piantadosi (1988). Consider the generalized linear model given by

Table 7.6 Approximate 95th Percentile of K^* for Eight Strata and Various $n, \varepsilon = 0.0001$ When $\gamma = 0.05, 0.10$ (Complete Randomization)

n	$\gamma = 0.05$	$\gamma = 0.10$
188	1,171,802	42,555
376	734,397	27,132
940	637,005	20,659
1880	2,816,027	28,184

$$E(\mathbf{y} | \mathbf{t}, \mathbf{X}) = h(\boldsymbol{\eta}) \equiv h(\boldsymbol{\mu} + \mathbf{t}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta}),$$

where h is a link function, \mathbf{y} is a vector of patient outcomes, $\boldsymbol{\mu} = \mathbf{1}\mu$ is the intercept, $\boldsymbol{\alpha}$ is the treatment effect, \mathbf{X} is a matrix of covariates, and $\boldsymbol{\beta}$ is the covariate effect. Under the null hypothesis, $\boldsymbol{\alpha} = 0$, and we can use standard techniques for the generalized linear model under an exponential family to compute $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\beta}}_0$. From this, we can compute

$$h(\hat{\boldsymbol{\eta}}_0) = h(\hat{\boldsymbol{\mu}}_0 + \mathbf{X}\hat{\boldsymbol{\beta}}_0),$$

and then the residuals $\mathbf{r} = \mathbf{y} - h(\hat{\mathbf{h}}_0)$. The residuals can then be ranked appropriately across the n patients using score functions, and a randomization test computed based on either the unconditional or conditional reference set. There is nothing inherently different about this approach from what we have already described, except that we have a different mechanism for computing the treatment effect metric. We are imposing a model only on the covariates; the adjusted treatment effect itself is evaluated strictly by randomization-based methods.

7.4.7 Covariate-Adaptive Randomization

Simon (1979, p. 508) first describes a technique for computing randomization tests following covariate-adaptive randomization (or as he calls it *adaptive stratification*):

It is possible, though cumbersome, to perform the appropriate permutation test generated by a nondeterministic adaptive stratification design. One assumes that the patient responses, covariate values, and sequence of patient arrivals are all fixed. One then simulates on a computer the assignment of treatments to patients using the design and the treatment assignments probabilities actually employed. Replication of the simulation generates the approximate null distribution of the test statistic adopted and the significance level. One needs not make the questionable assumption that the sequence of patient arrivals is random.

The last sentence has become somewhat controversial. Ebbutt et al. (1997) present an example where results differed when the randomization test took into consideration the sequencing of patient arrivals. Senn (discussion on Atkinson 1999) concludes from this that the disease was changing in some way through the course of the trial and thus there was a time trend present.

We conclude that the appropriate techniques for randomization-based inference following covariate-adaptive randomization have yet to be determined, and this is a fertile area for future research.

7.4.8 Power

One of the criticisms of randomization tests is the conceptual difficulty of developing relevant alternative hypotheses for power computation. Since

power is computed under a local alternative that depends on parameters of a population model, standard power computations do not apply to randomization tests. We take the approach of Flyer (1998) in evaluating power through a location-shift model. The most general form of an alternative hypothesis for a randomization test is that the outcome of the j th patient is given by $y_j = x_j$ if the patient is randomized to a control treatment and as $y_j = x_j + \delta_j$ if the patient is randomized to the experimental treatment. For continuous outcomes, we can examine a simple location-shift model (Lehmann 1974, 1986), such as $\delta_j = \bar{\delta}$. For binary responses, this makes no sense, but Flyer suggests removing a random sample of experimental treatment subjects and assigning $\delta_j = -1$ to that sample; otherwise, $\delta_j = 0$. The treatment effect $\bar{\delta}$ is then chosen to produce a prespecified power for a size α randomization test, simulated over M replications of patient responses. It has become almost a part of randomization folklore that unconditional tests are more powerful than conditional tests. While this may be true in many cases, it is not difficult to construct simple counterexamples where it is not the case.

The simulation of power for an unconditional one-sided test of size 0.05 can be accomplished using the following steps:

1. Generate a data set \mathbf{x} for the control group. This can be accomplished under some stochastic model. From \mathbf{x} , compute \mathbf{a} , the score vector.
2. Generate L randomization sequences using the desired procedure and compute $S_l, l = 1, \dots, L$. Order the S_l s. Find S_c , such that 95% of the S_l s fall below S_c and 5% fall above. S_c is then the critical value.
3. Apply the location shift to \mathbf{x} to create \mathbf{y}_l for each of the K randomization sequences. From \mathbf{y}_l , compute $\mathbf{a}_l, l = 1, \dots, L$. Recompute $S_l, l = 1, \dots, L$. Power is then the proportion of the S_l s that exceed S_c .
4. Repeat steps 1–3 M times. This results in M powers, and appropriate quantiles can then be computed. This is the simulated power distribution under a randomization model.

For a conditional test, we generate $K > L$ sequences for each replication $1, \dots, M$, and order only those sequences that satisfy the requisite condition on the sample fractions.

7.5 CONCLUSIONS AND PRACTICAL CONSIDERATIONS

We have described adaptive randomization procedures, and, in particular, have discussed their impact on inference. While it is clear that response-adaptive randomization and covariate-adaptive randomization are conceptually more complex than restricted randomization, likelihood-based inference is no more complicated. Randomization-based inference is considerably more complicated, and there are open research topics associated with it. While a recent

draft document from the U.S. Food and Drug Administration (FDA 2010) on guidance for adaptive clinical trials for industry has listed adaptive randomization as an “adaptive study design whose properties are less well understood,” it should be clear that adaptive randomization procedures are very well understood, and are ready to be implemented in practice.

Clinical trials are complex multiobjective studies, and should be designed with the same scrutiny that survey samplers design complex sample surveys. Objectives, such as minimization of biases, balance, ethics, efficiency, and inference, will dictate the appropriate procedure to be used, whether it be restricted, covariate adaptive, or response adaptive. Such decisions should become standard in the design of any clinical trial.

In clinical trials using adaptive randomization other than response-adaptive randomization, the number of patients assigned to each treatment should be known with some degree of approximation in advance. However, drug supply management for response-adaptive randomization is more challenging. Although the total number of patients is known, the number of patients to be allocated to each treatment is unknown. Even if an optimal allocation proportion is used, the unknown parameters in the allocation proportion still render the quantities unknown at the beginning of the trial. However, this challenge also provides opportunities for an improvement of drug supply management. As pointed out by Quinlan and Krams (2006), “Drug supply management issues at times are perceived as an insurmountable hurdle to the implementation of adaptive designs. In our opinion, the opposite is the case: making use of the supporting infrastructure facilitating adaptive designs can also facilitate the implementation of flexible real-time supply chain management of study drug. This opens up opportunities for minimizing drug supply wastage.” In fact, even in the simplest permuted block design, it is commonplace that some blocks are not filled, causing drug wastage if half the block size of each treatment has been predeposited for the block. Therefore, a real-time drug supply chain is useful whether adaptive randomization is used or not.

Fortunately, modern technology advancement has enabled flexible drug supply management. First, interactive response system (IRS) and electronic data capture (EDC) can facilitate central randomization and drug supply, and both are critical for response-adaptive randomization. IRS is comprised of a phone-based or web-based user interface that is connected by a complex IT system to a central database. IRS can be deployed in simple settings to perform functions like central randomization, or they can be used to facilitate more complex services, such as dynamic treatment allocation. EDC guarantees high-quality clinical data to be transmitted to the center. These data are used to update the randomization probability. Moreover, most IRS vendors provide clinical supplies management services in support of studies that utilize their IRS. These services often include clinical supplies demand projection, inventory management from depot to site, tracking from depot to site to patient, and expiry date management. Clinical supplies are usually managed in real-time based on patient enrollment rate, protocol-specific information and

supply inventory levels. This type of supply management strategy works adequately for projecting need over a short time frame at the site or patient level. Since the ability to plan appropriate drug supply strategies, in support of potential treatment arm modifications and study samples size adjustments, is critical to the success of many adaptive trials, forecasting is a key enabling technology (Miller, Deleger, and Murphy 2005).

We conclude by noting that although randomization has been unanimously recognized as one of the pillars of scientific experiments, and randomized clinical trials have been the gold standard research method practiced in medical community and supported or required by regulatory agencies, controversies of application of randomization in clinical trials persist. The controversies mainly derive from the fact that in randomized clinical trials, a patient receives a treatment with a particular probability, instead of being dictated by norm of clinical care where the treatment she receives should maximize her individual benefit–risk ratio. This departure from norm of clinical care could sacrifice this patient to obtain clinical information for other patients. Some may maintain that such controversies can be settled if informed consent is given by the patient so that she is fully aware of the opportunity cost associated with enrollment in the trial. Others (Miller and Weijer 2006) contend that it violates the clinician's obligation to allow his or her patients to enter a clinical trial that may not best benefit the patient. Some (Freedman 1987) have argued that randomization is applicable only if clinical equipoise is satisfied where with all available evidence one treatment is not more favorable than the other. However, in many clinical trials, such equipoise, if it exists, can be gradually overridden by increasing evidence that favors one treatment over the other. In this sense, response-adaptive randomization is proposed to utilize the accumulative data to adjust randomization probability such that the patients' benefit is maximized with a larger probability to be assigned to a better performing treatment. Proponents (Cornfield, Halperin, and Greenhouse 1969; Weinstein 1974) of response-adaptive randomization procedures think that it at least reduces the ethical problem revolved around randomization in clinical trials. Some opponents (Royall 1991) take a dichotomous view on the ethical problem: if enough evidence favoring one treatment is found, all patients should be assigned to the treatment with certainty, not high probability. Therefore, there is no place for response-adaptive randomization. Other opponents (Simon 1991) do not think it is an attractive approach to give a patient a greater than 50% chance to get a treatment that is doing better.

ACKNOWLEDGMENT

Professor Rosenberger's research on randomization is supported by grant DMS-0904253 from the National Science Foundation under the 2009 American Reinvestment and Recovery Act.

REFERENCES

- Atkinson, A.C. (1999). Optimum biased-coin designs for sequential treatment allocation with covariate information. *Statistics in Medicine*, **18**, 1741–1752.
- Bai, Z., F. Hu, and W.F. Rosenberger (2002). Asymptotic properties of adaptive designs with delayed response. *Annals of Statistics*, **30**, 122–139.
- Biswas, A. and S. Mandal (2004). Optimal adaptive designs in phase III clinical trials for continuous responses with covariates. In: *mODa7: Advances in Model-Oriented Design and Analysis*, A. Di Buccianico, H. Lauter, and H.P. Wynn (eds.). Heidelberg, Germany: Physica-Verlag, pp. 51–58.
- Coad, D.S. and Z. Govindarajulu (2000). Corrected confidence intervals following a sequential adaptive trial with binary response. *Journal of Statistical Planning and Inference*, **91**, 53–64.
- Coad, D.S. and A. Ivanova (2001). Bias calculations for adaptive urn designs. *Sequential Analysis*, **20**, 229–239.
- Coad, D.S. and M.B. Woodroffe (1997). Approximate confidence intervals after a sequential clinical trial comparing two exponential survival curves with censoring. *Journal of Statistical Planning and Inference*, **63**, 79–96.
- Cornfield, J., M. Halperin, and S.W. Greenhouse (1969). An adaptive procedure for sequential clinical trials. *Journal of the American Statistical Association*, **64**, 759–770.
- Cox, D.R. (1982). A remark on randomization in clinical trials. *Utilitas Mathematica*, **21A**, 245–252.
- Dragalin, V. (2006). Adaptive designs: Terminology and classification. *Drug Information Journal*, **40**, 425–435.
- Ebbutt, A., R. Kay, J. McNamara, and J. Engler (1997). The analysis of trials using a minimisation algorithm. In: *Statisticians in the Pharmaceutical Industry Annual Conference Report, 1997*, PSI (ed.). London: PSI, pp. 12–15.
- Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, **58**, 403–417.
- Eisele, J. (1994). The doubly adaptive biased coin design for sequential clinical trials. *Journal of Statistical Planning and Inference*, **38**, 249–262.
- FDA (2010). Guidance for industry: Adaptive design clinical trials for drugs and biologics. Silver Spring: U.S. Food and Drug Administration (draft guidance).
- Fisher, R.A. (1935). *The Design of Experiments*. New York: Hafner Press.
- Flyer, P.A. (1998). A comparison of conditional and unconditional randomization tests for highly stratified data. *Biometrics*, **54**, 1551–1559.
- Freedman, B. (1987). Equipoise and the ethics of clinical research. *The New England Journal of Medicine*, **317**, 141–145.
- Gail, M.H., W.Y. Tan, and S. Piantadosi (1988). Tests for no treatment effect in randomized clinical trials. *Biometrika*, **75**, 57–64.
- Hu, F. and W.F. Rosenberger (2003). Optimality, variability, power: Evaluating response-adaptive randomization procedures for treatment comparisons. *Journal of American Statistical Association*, **98**, 671–678.
- Hu, F. and W.F. Rosenberger (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials*. New York: Wiley.

- Hu, F. and L.-X. Zhang (2004). Asymptotic properties of doubly adaptive biased coin design for multi-treatment clinical trials. *Annals of Statistics*, **32**, 268–301.
- Hu, F., W.F. Rosenberger, and L.-X. Zhang (2006). Asymptotically best response-adaptive randomization procedures for treatment comparisons. *Journal of Statistical Planning and Inference*, **136**, 1911–1922.
- Hu, F., L.-X. Zhang, and X. He (2009). Efficient randomized adaptive designs. *Annals of Statistics*, **37**, 2543–2560.
- Ivanova, A. (2003). A play-the-winner type urn model with reduced variability. *Metrika*, **58**, 1–13.
- Jennison, C. and B.W. Turnbull (2000). *Group Sequential Methods With Applications to Clinical Trials*. Boca Raton, FL: Chapman and Hall/CRC.
- Kuznetsova, O. and A. Ivanova (2007). Allocation in randomized clinical trials. In: *Pharmaceutical Statistics Using SAS: A Practical Guide*, A. Dmitrienko, C. Chuang-Stein, and R.B.D. Agostino (eds.). Cary, NC: SAS Institute, pp. 213–235.
- Lehmann, E.L. (1974). *Nonparametrics: Statistical Methods Based on Ranks*. Oakland, CA: Holden-Day.
- Lehmann, E.L. (1986). *Testing Statistical Hypotheses*. New York: Wiley.
- Markaryan, T. and W.F. Rosenberger (2010). Exact properties of Efron's biased coin randomization procedure. *Annals of Statistics*, **38**, 1546–1567.
- Melfi, V.F., C. Page, and M. Geraldes (2001). An adaptive randomized design with application to estimation. *The Canadian Journal of Statistics*, **29**, 107–116.
- Miller, E., S. Deleger, and J. Murphy (2005). Implementing and managing adaptive designs for clinical trials. *Pharma Supplies and News*, October 21.
- Miller, P.B. and C. Weijer (2006). Trust based obligations of the state and physician-researchers to patient-subjects. *Journal of Medical Ethics*, **32**, 542–547.
- Pesarin, F. (2001). *Multivariate Permutation Tests With Applications in Biostatistics*. Chichester, UK: Wiley.
- Pocock, S.J. and R. Simon (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, **31**, 103–115.
- Prentice, R.J. (1978). Linear rank tests with right-censored data. *Biometrika*, **65**, 167–179.
- Quinlan, J.A. and M. Krams (2006). Implementing adaptive designs: Logistical and operational considerations. *Drug Information Journal*, **40**, 437–444.
- Rosenberger, W.F. (1993). Asymptotic inference with response-adaptive treatment allocation designs. *Annals of Statistics*, **21**, 2098–2107.
- Rosenberger, W.F. and F. Hu (1999). Bootstrap methods for adaptive designs. *Statistics in Medicine*, **18**, 1757–1767.
- Rosenberger, W.F. and F. Hu (2004). Maximizing power and minimizing treatment failures in clinical trials. *Clinical Trials*, **1**, 141–147.
- Rosenberger, W.F. and J.L. Lachin (2002). *Randomization in Clinical Trials: Theory and Practice*. New York: Wiley.
- Rosenberger, W.F. and O. Sverdlov (2008). Handling covariates in the design of clinical trials. *Statistical Science*, **23**, 404–419.
- Rosenberger, W.F., N. Stallard, A. Ivanova, C. Harper, and M. Ricks (2001). Optimal adaptive designs for binary response trials. *Biometrics*, **57**, 173–177.

- Royall, R.M. (1991). Ethics and statistics in randomized clinical trials. *Statistical Science*, **6**, 52–62.
- Shao, J., X. Yu, and B. Zhong (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika*, **97**, 347–360.
- Simon, R. (1979). Restricted randomization designs in clinical trials. *Biometrics*, **35**, 503–512.
- Simon, R. (1991). A decade of progress in statistical methodology for clinical trials. *Statistics in Medicine*, **10**, 1789–1817.
- Smith, R.L. (1984). Sequential treatment allocation using biased coin designs. *Journal of the Royal Statistical Society. Series B*, **46**, 519–543.
- Wei, L.J. (1978). An application of an urn model to the design of sequential controlled clinical trials. *Journal of the American Statistical Association*, **73**, 559–563.
- Wei, L.J. and S. Durham (1978). The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association*, **73**, 840–843.
- Weinstein, M.C. (1974). Allocation of subjects in medical experiments. *New England Journal of Medicine*, **291**, 1278–1285.
- Wendler, D. (2009). The ethics of clinical research. In: *Stanford Encyclopedia of Philosophy*, E.N. Zalta (ed.). Stanford, CA: Stanford University.
- Yuan, A. and G.X. Chai (2008). Optimal adaptive generalized Pólya urn design for multi-arm clinical trials. *Journal of Multivariate Analysis*, **99**, 1–24.
- Zelen, M. (1974). The randomization and stratification of patients to clinical trials. *Journal of Chronic Disease*, **27**, 365–375.
- Zhang, L. and W.F. Rosenberger (2006). Response-adaptive randomization for clinical trials with continuous outcomes. *Biometrics*, **62**, 562–569.
- Zhang, L. and W.F. Rosenberger (2007). Response-adaptive randomization for clinical trials with survival outcomes: The parametric approach. *Journal of the Royal Statistical Society. Series C*, **53**, 153–165.
- Zhang, L. and W.F. Rosenberger (2011). Optimal response-adaptive randomization in clinical trials. In: *Handbook of Adaptive Designs for Pharmaceutical and Clinical Development*, A. Pong and S. Chow (eds.). Boca Raton, FL: CRC Press, pp. 15-1–15-13.
- Zhang, L.-X., W.S. Chan, S.H. Cheung, and F. Hu (2007). A generalized urn model for clinical trials with delayed responses. *Statistica Sinica*, **17**, 387–409.

CHAPTER 8

Search Linear Model for Identification and Discrimination

Subir Ghosh

8.1 INTRODUCTION

In designing an experiment, we often assume a model to describe the data to be collected and then find an efficient design satisfying one or more optimal properties under the assumed model. This approach works well when we are absolutely sure that the assumed model will fit the experimental data adequately. In reality, we can rarely be sure about a model in terms of its effectiveness in describing the data adequately. However, we can be sure about a set of possible models that would describe the data adequately and one of them would possibly describe the data better than the other models in the set. The pioneering work of Srivastava (1975) introduced the *search linear model* with the purpose of searching for and identifying the best model for describing the data from the set of possible models considered for this purpose. A *search design* under the search linear model has the ability to perform the task of searching for, identifying the models, and discriminating between any two possible models within this set. The purpose of this chapter is to provide an introduction to this area of search designs and search linear models with its application to factorial experiments (Bose 1947), expanding the available coverage in HK2, chapter 16.

The checking of whether a model fits the data better over another model is routinely done at the inference stage after the data have been collected. There is a large literature on alternative formal approaches for this purpose. After finding the best model for describing the data, inferences are drawn,

Design and Analysis of Experiments: Special Designs and Applications, First Edition. Edited by Klaus Hinkelmann.

© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

predictions and policy decisions are made. The issue of the *data quality* is fundamental for successfully conducting this process of model selection, drawing inferences, and making predictions and policy decisions. The *true success* depends on intuition, creativity, insight, knowledge, skill, and efficiency in the process of model selection. In many situations, it is impossible to find a model to describe the data *globally*, but it is possible to describe the data *locally*. We may implement several locally fitted models to describe the data globally. No matter how complex the process is, the data quality is the essence of the process. An efficient design of experiment is essential for improving the data quality. When we design an experiment, we do not have the privilege of having the data at hand, but incorporating the available information in designing an experiment is necessary for achieving the higher data quality. We can describe the possible models at the design stage based on the available information, as well as using our intuition, creativity, insight, knowledge, and experience.

8.2 GENERAL LINEAR MODEL WITH FIXED EFFECTS

Suppose we fit the model below to a vector of observations $\mathbf{y}(n \times 1)$:

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}, \text{Var}(\mathbf{y}) = \sigma^2\mathbf{V}, \quad (8.1)$$

where $\boldsymbol{\beta}(p \times 1)$ is a vector of fixed unknown parameters, σ^2 is an unknown constant, $\mathbf{X}(n \times p)$ is a known matrix with rank r where $r \leq p \leq n$, and $\mathbf{V}(n \times n)$ is a known or an unknown matrix with rank s where $s \leq n$. We want to draw inference on the elements of $\boldsymbol{\beta}$ and \mathbf{V} , as well as σ^2 . Also, we want \mathbf{y} and hence \mathbf{X} to be such that the above inference can be drawn efficiently. The kinds of inference that we are interested in can be diverse and can cover a broad range of paradigms. The model satisfying Equation (8.1) is called a *general linear model with fixed effects* (GLMFE). A design d is chosen in collecting the observations \mathbf{y} so that all the inferences of interest can be drawn efficiently from the data. We assume in this chapter that the matrix \mathbf{V} is an identity matrix, that is, $\mathbf{V} = \mathbf{I}$ and furthermore $r = p$. A linear function $\mathbf{l}'\boldsymbol{\beta}$ of the elements of $\boldsymbol{\beta}$ is called an *estimable function* when there is at least one linear function $\mathbf{c}'\mathbf{y}$ of the elements of \mathbf{y} such that $E(\mathbf{c}'\mathbf{y}) = \mathbf{l}'\boldsymbol{\beta}$ under model (8.1), where $\mathbf{l}'(p \times 1)$ and $\mathbf{c}'(n \times 1)$ are known vectors (Bose 1944, 1949; Scheffé 1959, p. 13, HK1, chapter 4). The model (Eq. 8.1) with $\mathbf{V} = \mathbf{I}$ is said to be *identifiable* by the design d if all the p parameters in $\boldsymbol{\beta}$ are estimable functions, and we can unbiasedly estimate σ^2 . If we want to evaluate the *lack of fit* of the fitted model to the data \mathbf{y} with respect to a bigger model, we take $n > p$. The replication in observations is needed to measure the *pure error*. When $n = p$ and the observations are not replicated, we are unable to estimate σ^2 . In this chapter, we assume the multivariate normal distribution for \mathbf{y} for testing of hypotheses. We define

$$\mathbf{R} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \quad (8.2)$$

where the matrix $\mathbf{R}(n \times n)$ is an idempotent matrix with its rank, as well as its trace (tr) equal to $(n - p)$. We have $\mathbf{X}'\mathbf{R} = \mathbf{0}$, $\mathbf{R}' = \mathbf{R}$, and $(E(\mathbf{y}))'\mathbf{R} = \mathbf{0}$. The error sum of squares is denoted by

$$SSE = \mathbf{y}'\mathbf{R}\mathbf{y}. \quad (8.3)$$

It can be checked (Rao 1973) that

$$E(SSE) = \sigma^2 \text{tr}\mathbf{R} + (E(\mathbf{y}))'\mathbf{R}(E(\mathbf{y})) = \sigma^2(n - p). \quad (8.4)$$

Let $\mathbf{D} = \{d\}$ be a class of designs with the \mathbf{X} matrices, denoted by \mathbf{X}_d , having rank $\mathbf{X}_d = p$ for all designs d within the class. The least squares estimator of $\boldsymbol{\beta}$ for a design d is $\hat{\boldsymbol{\beta}}_d = (\mathbf{X}'_d\mathbf{X}_d)^{-1}\mathbf{X}'_d\mathbf{y}_d$ with variance $\text{Var}(\hat{\boldsymbol{\beta}}_d) = \sigma^2\mathbf{V}_d$, where $\mathbf{V}_d = (\mathbf{X}'_d\mathbf{X}_d)^{-1}$. A design d^* in \mathbf{D} is said to be a D -optimum design if the determinant of \mathbf{V}_d is minimum at $d = d^*$. A design d^* in \mathbf{D} is said to be an A-optimum design if the trace of \mathbf{V}_d is minimum at $d = d^*$. A design d^* in \mathbf{D} is said to be an E-optimum design if the maximum eigenvalue of \mathbf{V}_d is minimum for $d = d^*$. The details about these three criterion functions can be found in Kiefer (1959), Fedorov (1972), Pukelsheim (1993), Atkinson, Donev, and Tobias (2007), HK2, section 1.13 among many others.

8.3 SEARCH LINEAR MODEL

Consider the search linear model introduced in Srivastava (1975)

$$E(\mathbf{y}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2, \text{Var}(\mathbf{y}) = \sigma^2\mathbf{V}, \quad (8.5)$$

where $\mathbf{X}_1 (n_1 \times p_1)$ and $\mathbf{X}_2 (n_2 \times p_2)$ are known matrices, $\boldsymbol{\beta}_1 (p_1 \times 1)$ is a vector of fixed unknown parameters, and we know that at most k elements of $\boldsymbol{\beta}_2$ are nonzero but we do not know which elements are nonzero. The goal is to search for and identify the nonzero elements of $\boldsymbol{\beta}_2$ and then estimate them along with the elements of $\boldsymbol{\beta}_1$. Such a model is called a *search linear model* (SLM). We want \mathbf{y} and hence \mathbf{X}_1 and \mathbf{X}_2 to be such that we can achieve this goal; the design d that generates \mathbf{y} is called a *search design* (SD). The SLM in Equation (8.5) becomes the general linear fixed effects model (GLMFE) in Equation (8.1) when $\boldsymbol{\beta}_2 = \mathbf{0}$, $\mathbf{X}_1 = \mathbf{X}$, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}$, and $p_1 = p$. Therefore, a possible bias from the unknown set of at most k nonzero elements of $\boldsymbol{\beta}_2$ will be present in the inferences drawn on $\boldsymbol{\beta}_1 (= \boldsymbol{\beta})$ under the GLMFE in Equation (8.1). The possibility of such bias will be eliminated in the inferences drawn on $\boldsymbol{\beta}_1 (= \boldsymbol{\beta})$ under the SLM in Equation (8.5). The search designs to generate \mathbf{y} for the SLM in Equation (8.5) provide additional features that are not present in the designs to generate \mathbf{y} for the GLMFE in Equation (8.1). Again, we assume in this chapter $\mathbf{V} = \mathbf{I}$, and, furthermore, the rank of \mathbf{X}_1 is p_1 , and for all $(n \times k)$

submatrices \mathbf{X}_{2i} and $\mathbf{X}_{2i'}$ of \mathbf{X}_2 , $i, i' = 1, \dots, \binom{p_2}{k}$, $i \neq i'$, the rank of $[\mathbf{X}_1, \mathbf{X}_{2i}, \mathbf{X}_{2i'}]$ is $p_1 + 2k$.

To explain the SLM in Equation (8.5), we consider first the $\binom{p_2}{k}$ models for $i = 1, \dots, \binom{p_2}{k}$

$$\mathbf{E}(\mathbf{y}) = \mathbf{X}_1\beta_1 + \mathbf{X}_{2i}\beta_{2i}, \text{Var}(\mathbf{y}) = \sigma^2\mathbf{I}, \quad (8.6)$$

where β_{2i} is a $(k \times 1)$ sub-vector of β_2 . For an SD, all the elements of β_1 and β_{2i} are estimable functions for $i = 1, \dots, \binom{p_2}{k}$ under Equation (8.6). An SD with replications or without replications permits the estimation of σ^2 . In other words, an SD allows to identify all the $\binom{p_2}{k}$ models in Equation (8.6). There is an additional role of SD in discriminating between two models i and i' in Equation (8.6) for all $i \neq i'$. To explain this discrimination between two models in all possible pairs of models in Equation (8.6), we consider next the $\binom{\binom{p_2}{k}}{2}$ models (Srivastava 1975) for $i, i' = 1, \dots, \binom{p_2}{k}$, $i \neq i'$,

$$\mathbf{E}(\mathbf{y}) = \mathbf{X}_1\beta_1 + \mathbf{X}_{2i}\beta_{2i} + \mathbf{X}_{2i'}\beta_{2i'}, \text{Var}(\mathbf{y}) = \sigma^2\mathbf{I}. \quad (8.7)$$

We note that both models i and i' in Equation (8.6) are nested within the model (i, i') in Equation (8.7). For an SD, all the elements of β_1 , β_{2i} , and $\beta_{2i'}$ are estimable functions for $i, i' = 1, \dots, \binom{p_2}{k}$, $i \neq i'$, under Equation (8.7). This property of SD permits the SLM to discriminate between two models (Srivastava 1975) in all possible pairs of models in Equation (8.6). We now state the fundamental theorem (Srivastava 1975) of search linear model.

Theorem 8.1 The Fundamental Theorem of the Search Linear Model. (Srivastava 1975)

- (a)** A necessary condition for the identification of all and the discrimination between the $\binom{p_2}{k}$ models for $i = 1, \dots, \binom{p_2}{k}$ in Equation (8.6) is that

$$\text{rank}(\mathbf{X}_1, \mathbf{X}_{2i}, \mathbf{X}_{2i'}) = p_1 + 2k, i, i' = 1, \dots, \binom{p_2}{k}, i \neq i',$$

- (b)** If $\sigma^2 = 0$, then the above rank condition is also sufficient. In this case, the identification and discrimination can be done with probability one. If $\sigma^2 > 0$, the identification and discrimination can be done with probability less than 1.

We define

$$\begin{aligned} \mathbf{X}^{(i)} &= [\mathbf{X}_1, \mathbf{X}_{2i}], \\ \mathbf{R}_i &= \mathbf{I} - \mathbf{X}^{(i)}(\mathbf{X}^{(i)\prime}\mathbf{X}^{(i)})^{-1}\mathbf{X}^{(i)\prime}. \end{aligned} \quad (8.8)$$

The error sum of squares for model i in Equation (8.6) is

$$\text{SSE}_i = \mathbf{y}' \mathbf{R}_i \mathbf{y}. \quad (8.9)$$

It can be checked (Rao 1973) that

$$E(\text{SSE}_i) = \sigma^2 \text{tr} \mathbf{R}_i + (E(\mathbf{y}))' \mathbf{R}_i (E(\mathbf{y})), \text{tr} \mathbf{R}_i = n - p_1 - k. \quad (8.10)$$

All the $\binom{p_2}{k}$ models for $i = 1, \dots, \binom{p_2}{k}$ in Equation (8.6) have the elements of $\boldsymbol{\beta}_1$ as the common parameters and the remaining elements in $\boldsymbol{\beta}_{2i}$ may or may not have some common parameters. For the best model from the $\binom{p_2}{k}$ models, the sum of squares of error (SSE) is minimum (Srivastava 1975). For a fixed value of k , all $\binom{p_2}{k}$ models are fitted to the data, and the search procedure selects the model with the smallest SSE as the best model for describing the data. Two popular criteria for model comparisons, namely minimizing the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) (Akaike 1969, 1973; Sawa 1978), turn out to be exactly equivalent to the minimization of SSE in this search procedure for the comparison of $\binom{p_2}{k}$ models. When we compare the best models for different values of k , they may not be equivalent.

Cheng, Deng, and Tang (2002), Cheng, Steinberg, and Sun (1999), Cheng and Mukerjee (1998) implemented the concept of *estimation capacity* (see also Chapter 9) of a design d for the models in Equation (8.6) in the context of a fractional factorial experiment as the number of i s out of the $\binom{p_2}{k}$ possibilities for which all the elements of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_{2i}$ are estimable functions under Equation (8.6). A design is said to have a *full estimation capacity* if it has the estimation capacity $\binom{p_2}{k}$. We denote the determinant of $\mathbf{X}^{(i)'} \mathbf{X}^{(i)}$ by $\det(\mathbf{X}^{(i)'} \mathbf{X}^{(i)})$. The indicator function $I(\det(\mathbf{X}^{(i)'} \mathbf{X}^{(i)}))$ is now defined as

$$I(\det(\mathbf{X}^{(i)'} \mathbf{X}^{(i)})) = \begin{cases} 1 & \text{if } \det(\mathbf{X}^{(i)'} \mathbf{X}^{(i)}) > 0 \\ 0 & \text{if } \det(\mathbf{X}^{(i)'} \mathbf{X}^{(i)}) = 0. \end{cases} \quad (8.11)$$

Clearly, for a design d with the full estimation capacity, we have $\det(\mathbf{X}^{(i)'} \mathbf{X}^{(i)}) > 0$ for all $i = 1, \dots, \binom{p_2}{k}$. A general definition of the estimation capacity $E_k(d)$ can now be given as

$$E_k(d) = \sum_{i=1}^{\binom{p_2}{k}} I(\det(\mathbf{X}^{(i)'} \mathbf{X}^{(i)})). \quad (8.12)$$

Thus $E_k(d) = \binom{p_2}{k}$ for a design d with the full estimation capacity.

We define $\boldsymbol{\beta}^{(i)} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_{2i})$. The least squares estimator of $\boldsymbol{\beta}^{(i)}$ for a design d is $\hat{\boldsymbol{\beta}}_d^{(i)} = (\mathbf{X}_d^{(i)'} \mathbf{X}_d^{(i)})^{-1} \mathbf{X}_d^{(i)'} \mathbf{y}_d$ with variance $\text{Var}(\hat{\boldsymbol{\beta}}_d^{(i)}) = \sigma^2 \mathbf{V}_d^{(i)}$, where $\mathbf{V}_d^{(i)} = (\mathbf{X}_d^{(i)'} \mathbf{X}_d^{(i)})^{-1}$. Let $\mathbf{D} = \{d\}$ be a class of search designs (SDs). We consider the six criterion functions (Läuter 1974; Srivastava 1977) below to compare the SDs in \mathbf{D} :

$$\begin{aligned}
\text{AD} &= \text{Arithmetic mean of the determinants of } \mathbf{V}_d^{(i)}, i = 1, \dots, \binom{P_2}{k}, \\
\text{AT} &= \text{Arithmetic mean of the traces of } \mathbf{V}_d^{(i)}, i = 1, \dots, \binom{P_2}{k}, \\
\text{AMEV} &= \text{Arithmetic mean of the maximum eigenvalues of } \mathbf{V}_d^{(i)}, i = 1, \dots, \binom{P_2}{k}, \\
\text{GD} &= \text{Geometric mean of the determinants of } \mathbf{V}_d^{(i)}, i = 1, \dots, \binom{P_2}{k}, \\
\text{GT} &= \text{Geometric mean of the traces of } \mathbf{V}_d^{(i)}, i = 1, \dots, \binom{P_2}{k}, \\
\text{GMEV} &= \text{Geometric mean of the maximum eigenvalues of } \mathbf{V}_d^{(i)}, i = 1, \dots, \binom{P_2}{k}.
\end{aligned} \tag{8.13}$$

An SD d^* in \mathbf{D} is optimum with respect to (wrt) a criterion function if the value of the criterion function is minimum for d^* . When $\beta_2 = 0$, the criterion functions AD and GD become the D -optimality criterion function, AT and GT become the A -optimality criterion function, and AMEV and GMEV become the E -optimality criterion function.

8.4 APPLICATIONS

We now consider the application of the SLM in fractional factorial experiments where we normally assume that the lower order effects are important and the higher order effects are all negligible. In a main effect plan, we assume that the factors do not interact, or in other words, the interaction effects are all negligible. Such an assumption may or may not be true in reality because of a possible presence of a few significant non-negligible interaction effects. The GLMFEs cannot identify these non-negligible effects using a small number of runs or treatments considerably smaller than the total number of possible runs for an experiment. This motivates the use of SDs under the SLM in searching for and identifying the non-negligible effects. We explain this first with fractional factorial plans or designs for 2^m factorial experiments and then explain it with fractional factorial plans or designs for 3^m factorial experiments.

8.4.1 2^m Factorial Designs

For a 2^m factorial experiment, we denote a run v by (a_1, a_2, \dots, a_m) , where a_u takes the value 1 when the factor u is at the high level and the value -1

when the factor u is at the low level in the run v for $u = 1, 2, \dots, m$. We obtain an $((\binom{m}{w} \times m)$ matrix $\mathbf{S}_v^{(w)}$, whose rows are runs obtained from the run v by multiplying any w elements by (-1) and the remaining $(m - w)$ by 1. We can choose these w elements of the run v in $\binom{m}{w}$ ways giving the rows of the matrix $\mathbf{S}_v^{(w)}$, with $0 \leq w \leq m$.

Example 8.1. We consider a 2^4 factorial experiment, that is, $m = 4$. For the run v , (a_1, a_2, a_3, a_4) , the matrices $\mathbf{S}_v^{(w)}$, $w = 0, 1, 2, 3, 4$ are given below:

$$\begin{aligned}\mathbf{S}_v^{(0)} &= (a_1, a_2, a_3, a_4) = -\mathbf{S}_v^{(4)}, \\ \mathbf{S}_v^{(1)} &= \begin{bmatrix} -a_1 & a_2 & a_3 & a_4 \\ a_1 & -a_2 & a_3 & a_4 \\ a_1 & a_2 & -a_3 & a_4 \\ a_1 & a_2 & a_3 & -a_4 \end{bmatrix} = -\mathbf{S}_v^{(3)}, \\ \mathbf{S}_v^{(2)} &= \begin{bmatrix} -a_1 & -a_2 & a_3 & a_4 \\ -a_1 & a_2 & -a_3 & a_4 \\ -a_1 & a_2 & a_3 & -a_4 \\ a_1 & -a_2 & -a_3 & a_4 \\ a_1 & -a_2 & a_3 & -a_4 \\ a_1 & a_2 & -a_3 & -a_4 \end{bmatrix}. \end{aligned} \tag{8.14}$$

The run v can be any of the 16 possible runs. Moreover, the total number of runs in $\mathbf{S}_v^{(0)}, \mathbf{S}_v^{(1)}, \mathbf{S}_v^{(2)}, \mathbf{S}_v^{(3)}$, and $\mathbf{S}_v^{(4)}$ is 16, representing the 16 possible runs for the 2^4 factorial design.

Example 8.2. For a 2^4 factorial experiment, we consider the GLMFE in Equation (8.1) with the parameters in the vector β as the general mean and the main effects. The design with five runs: the run v and four runs in $\mathbf{S}_v^{(3)}$, is optimum wrt the A -, D -, and E -optimality criterion functions. If we replicate each treatment $r (>1)$ times, we can then identify the model in Equation (8.1) in the sense that the parameters in β are estimable functions, and moreover we can unbiasedly estimate σ^2 . We can achieve the same with unequal replicates of runs as well. The 16 possible choices for the run v give 16 such designs, which are all isomorphic to one another by renaming of the levels of factors. We present one design with five runs (Ghosh and Tian 2006):

$$\begin{aligned}v &= (1, 1, 1, 1), \\ \mathbf{S}_v^{(3)} &= \begin{bmatrix} -1 & -1 & -1 & 1 \\ -1 & -1 & 1 & -1 \\ -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}. \end{aligned} \tag{8.15}$$

Example 8.3 A Near Foldover SD. For a 2^4 factorial experiment, we consider the model in Equation (8.6) with the parameters in the vector β_1 as the general mean and the main effects and the vector β_{2i} as 3 two-factor interactions ($k = 3$). The design with 9 runs: the run v , 4 runs in $\mathbf{S}_v^{(3)}$, and 4 runs in $\mathbf{S}_v^{(1)}$, is optimum wrt the AD, AT, AMEV, GD, GT, and GMEV optimality criterion functions. The 16 possible choices for the run v give 16 such designs that are all isomorphic to one another by the renaming of the levels of factors. We present one design d with 9 runs (Ghosh and Tian 2006):

$$v = (1, 1, 1, 1) \text{ and } \mathbf{S}_v^{(3)} \text{ as in Equation (8.15),}$$

$$\mathbf{S}_v^{(1)} = \begin{bmatrix} -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{bmatrix}. \quad (8.16)$$

The design d in Equation (8.16) is a near foldover SD because the runs in $\mathbf{S}_v^{(1)}$ are folded over into the runs in $\mathbf{S}_v^{(3)}$ as $\mathbf{S}_v^{(1)} = -\mathbf{S}_v^{(3)}$. The d is an SD under the SLM in Equation (8.5) with the vector β_1 as the general mean and the main effects, the vector β_2 as two-factor interactions, and $k = 1$. Moreover, d has the full estimation capacity for the models in Equation (8.7).

Example 8.4 A Foldover SD. For a 2^5 factorial experiment, we consider the model in Equation (8.6) with the parameters in the vector β_1 as the general mean and the main effects and the vector β_{2i} as three two-factor interactions ($k = 3$). The design with 10 runs: 5 runs in $\mathbf{S}_v^{(1)}$ and 5 runs in $\mathbf{S}_v^{(4)}$ for a run v is optimum wrt the AD, AT, AMEV, GD, GT, and GMEV optimality criterion functions. The runs v and $-v$ generate the same design. The 32 possible choices for the run v give 16 distinct designs that are all isomorphic to one another by renaming of the levels of factors. We present one design d (Ghosh and Tian 2006) with $v = (1, 1, 1, 1, 1)$:

$$\mathbf{S}_v^{(1)} = -\mathbf{S}_v^{(4)} = \begin{bmatrix} -1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 \\ 1 & 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 \end{bmatrix}. \quad (8.17)$$

The design d in Equation (8.17) is a foldover SD because the runs in $\mathbf{S}_v^{(1)}$ are folded over into the runs in $\mathbf{S}_v^{(4)}$ as $\mathbf{S}_v^{(1)} = -\mathbf{S}_v^{(4)}$. The d is an SD under the SLM in Equation (8.5) with the vector β_1 as the general mean and the main effects, the vector β_2 as two-factor interactions, and $k = 1$. Moreover, d has the full estimation capacity for the models in Equation (8.7).

Example 8.5 A Series of SDs. For a 2^m factorial experiment, we consider the SLM in Equation (8.5) with the vector β_1 as the general mean, the main effects, and the two-factor interactions, and the vector β_2 as three-factor and higher-order interactions, and $k = 1$. The design with $(1 + 2m + \binom{m}{2})$ runs: the run v , the m runs in $\mathbf{S}_v^{(1)}$, the m runs in $\mathbf{S}_v^{(m-1)}$, and the $\binom{m}{2}$ runs in $\mathbf{S}_v^{(2)}$, is an SD. The 2^m choices for the run v give 2^m such designs that are all isomorphic to one another by renaming of the levels of factors. A design with $v = (-1, -1, \dots, -1)$ is given in Srivastava and Ghosh (1976) with the complete proof. These SDs are called the *resolution V plus one plans* (Srivastava and Ghosh 1976, HK2, chapter 16).

8.4.2 3^m Factorial Experiments

When the factors are at three levels: “high,” “medium,” and “low,” we denote the levels by 2 for high, 1 for medium, and 0 for low. Considering all quantitative factors, we denote a run v by (a_1, a_2, \dots, a_m) , where a_u takes the values 0, 1, or 2. The main effect of a factor has two components: the linear and quadratic effects. The interaction effect of two factors has four components: the linear \times linear, linear \times quadratic, quadratic \times linear, and quadratic \times quadratic effects (HK1, chapter 11). We continue this way for the higher order interaction effects.

Example 8.6 A Factor Screening SD. For a 3^5 factorial experiment, we consider the SLM in Equation (8.5) with $k = 1$ and 2 and the vector β_1 as the general mean and the vector β_2 as the main effects. The design given in Equation (8.18) (Ghosh and Burns 2002) is able to search and identify at most two factors out of five factors with only seven runs. The design is optimum wrt the AD, AT, AMEV, GD, GT, and GMEV optimality criterion functions for both $k = 1$ and 2.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 1 & 2 \\ 0 & 2 & 1 & 2 & 1 \\ 1 & 0 & 2 & 2 & 2 \\ 2 & 0 & 1 & 1 & 1 \\ 1 & 2 & 0 & 1 & 2 \\ 2 & 1 & 0 & 2 & 1 \end{bmatrix}. \quad (8.18)$$

Example 8.7 Another Factor Screening SD. For a 3^{10} factorial experiment, we consider the SLM in Equation (8.5) with the vector β_1 as the general mean and the vector β_2 as the main effects, and $k = 1$ and 2. The design given in Equation (8.19) (Ghosh and Burns 2002) is able to search and identify at most 2 factors out of 10 factors with only 19 runs. The design is optimum wrt the

AD, AT, AMEV, GD, GT, and GMEV optimality criterion functions for both $k = 1$ and 2.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 1 & 2 & 2 & 1 & 2 & 1 & 2 \\ 0 & 2 & 1 & 2 & 1 & 1 & 2 & 1 & 2 & 1 \\ 1 & 0 & 2 & 2 & 2 & 1 & 2 & 1 & 2 & 1 \\ 2 & 0 & 1 & 1 & 1 & 2 & 1 & 2 & 1 & 2 \\ 1 & 2 & 0 & 1 & 2 & 2 & 1 & 2 & 1 & 2 \\ 2 & 1 & 0 & 2 & 1 & 1 & 2 & 1 & 2 & 1 \\ 1 & 1 & 1 & 0 & 2 & 1 & 2 & 1 & 2 & 1 \\ 2 & 2 & 2 & 0 & 1 & 2 & 1 & 2 & 1 & 2 \\ 1 & 2 & 1 & 2 & 0 & 2 & 1 & 2 & 1 & 2 \\ 2 & 1 & 2 & 1 & 0 & 1 & 2 & 1 & 2 & 1 \\ 1 & 1 & 2 & 1 & 2 & 0 & 2 & 1 & 2 & 1 \\ 2 & 2 & 1 & 2 & 1 & 0 & 1 & 2 & 1 & 2 \\ 1 & 2 & 2 & 2 & 2 & 1 & 0 & 2 & 1 & 2 \\ 2 & 1 & 1 & 1 & 1 & 2 & 0 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 & 2 & 2 & 1 & 0 & 2 & 1 \\ 2 & 2 & 2 & 2 & 1 & 1 & 2 & 0 & 1 & 2 \\ 1 & 2 & 1 & 2 & 2 & 1 & 2 & 1 & 0 & 2 \\ 2 & 1 & 2 & 1 & 1 & 2 & 1 & 2 & 0 & 1 \end{bmatrix}. \quad (8.19)$$

Interestingly, the design in Equation (8.18) is a submatrix, with the first five rows and the first five columns of the design in Equation (8.19). In other words, the design in Equation (8.18) is nested within the design in Equation (8.19).

Example 8.8 A Series of SDs. For a 3^m factorial experiment with $m \geq 3$, we consider the SLM in Equation (8.5) with the vector β_1 as the general mean and the main effects and the vector β_2 as the two-factor interactions, and $k = 1$. Let \mathbf{I}_m be the $(m \times m)$ identity matrix and \mathbf{J}_m be the $(m \times m)$ matrix with all elements being 1. We define

$$\mathbf{T}_1 = -2\mathbf{I}_m + 2\mathbf{J}_m, \quad \mathbf{T}_2 = -\mathbf{I}_m + 2\mathbf{J}_m, \quad \mathbf{T}_3 = -\mathbf{I}_m + \mathbf{J}_m. \quad (8.20)$$

The m runs as the rows of \mathbf{T}_1 have the i^{th} factor at the level 0 and the other factors at level 2, $i = 1, \dots, m$, the m runs as the rows of \mathbf{T}_2 have the i^{th} factor at the level 1 and the other factors at level 2, $i = 1, \dots, m$, and the m runs as the rows of \mathbf{T}_3 have the i^{th} factor at the level 0 and the other factors at level 1, $i = 1, \dots, m$. The design d with $(1 + 3m)$ runs: the run $(2, 2, \dots, 2)$, m runs

in \mathbf{T}_1 , m runs in \mathbf{T}_2 , and m runs in \mathbf{T}_3 , is an SD which can search for one nonzero two-factor interaction ($k = 1$) and estimate it along with the general mean and main effects (Ghosh and Zhang 1987).

Example 8.9 Another Series of SDs. For a 3^m factorial experiment with $m \geq 4$, we consider the SLM in Equation (8.5) with the vector β_1 as the general mean, the main effects and two-factor interactions and the vector β_2 as the three-factor interactions, and $k = 1$. Let $\mathbf{T}_4(u, v)$ be a $((\frac{m}{2}) \times m)$ matrix whose rows represent the pairs (i, j) , $i, j = 1, \dots, m$, $i < j$. The $(\frac{m}{2})$ runs as the rows of $\mathbf{T}_4(u, v)$ have the i^{th} factor at the level u , the j^{th} factor at the level v and the other factors at level 2. The design d with $(1 + 3m + 4(\frac{m}{2}))$ runs: the run $(2, 2, \dots, 2)$ with all factors at level 2, m runs in \mathbf{T}_1 , m runs in \mathbf{T}_2 , m runs in \mathbf{T}_3 , $(\frac{m}{2})$ runs in $\mathbf{T}_4(0, 0)$, $(\frac{m}{2})$ runs in $\mathbf{T}_4(0, 1)$, $(\frac{m}{2})$ runs in $\mathbf{T}_4(1, 0)$, and $(\frac{m}{2})$ runs in $\mathbf{T}_4(1, 1)$, is an SD that can search for one nonzero three-factor interaction ($k = 1$) and estimate it along with the general mean, main effects, and two-factor interactions (Ghosh and Zhang 1987). The number of runs in d can be expressed as $(1 + m + 2m^2)$. Minimal resolution III, V, and VII plans require $(1 + 2m)$, $(1 + 2m^2)$, and $(1 + 2m^2 + 8(\frac{m}{3}))$ runs.

Many other examples of SDs can be found in Shirakura and Ohnishi (1985), Shirakura, Suetsugu, and Tsuji (2002), Srivastava and Gupta (1979), Chatterjee (1990), Chatterjee and Mukerjee (1993). A review article by Ghosh, Shirakura, and Srivastava (2007) presents many details.

8.5 EFFECTS OF NOISE IN PERFORMANCE COMPARISON

The fundamental theorem of the SLM states that the discrimination between the competitive models within a class can be performed to identify the best model with probability 1 if $\sigma^2 = 0$ and with probability less than 1 if $\sigma^2 > 0$. In reality, we always have $\sigma^2 > 0$. The search procedure with the data collected using an SD may or may not identify the best model correctly when $\sigma^2 > 0$ (Srivastava and Mallenby 1984; Shirakura, Takahashi, and Srivastava 1996; Ghosh and Teschmacher 2002). The performance of an SD d_1 is said to be better than the performance of another SD d_2 if the proportion of times the search procedure correctly identifies the best model with the data for d_1 is higher than the proportion with the data for d_2 .

We explain the idea of performance comparison of competing designs for a factorial experiment with an illustrative example. We present in Table 8.1 four fractional factorial plans $D1$, $D2$, $D3$, and $D4$ with 11 runs for a 2^5 factorial experiment. These designs are SDs with the ability to identify and discriminate between the models in Equation (8.6) with the parameters in the vector β_1 as the general mean and the main effects and the vector β_{2i} as one two-factor interaction ($k = 1$). Since $(\frac{5}{2}) = 10$, we have in total 10 models in (Eq. 8.6),

Table 8.1 The Designs $D1, D2, D3$, and $D4$

D1 and D2									
D1					D2				
A	B	C	D	E	A	B	C	D	E
+1	+1	+1	+1	-1	+1	+1	+1	+1	-1
+1	+1	+1	-1	+1	+1	+1	+1	-1	+1
+1	+1	-1	+1	+1	+1	+1	-1	+1	+1
+1	-1	+1	+1	+1	+1	-1	+1	+1	+1
-1	+1	+1	+1	+1	-1	+1	+1	+1	+1
+1	-1	-1	-1	-1	+1	-1	-1	-1	-1
-1	+1	-1	-1	-1	-1	+1	-1	-1	-1
-1	-1	+1	-1	-1	-1	-1	+1	-1	-1
-1	-1	-1	+1	-1	-1	-1	-1	-1	+1
-1	-1	-1	-1	-1	-1	-1	+1	+1	+1
D3 and D4									
D3					D4				
A	B	C	D	E	A	B	C	D	E
+1	+1	+1	+1	-1	+1	+1	+1	+1	-1
+1	+1	+1	-1	+1	+1	+1	+1	-1	+1
+1	+1	-1	+1	+1	+1	+1	-1	+1	+1
+1	-1	+1	+1	+1	-1	+1	+1	+1	+1
-1	+1	+1	-1	-1	-1	-1	-1	-1	-1
-1	+1	-1	+1	-1	-1	-1	-1	-1	+1
-1	+1	-1	-1	+1	-1	-1	-1	+1	-1
-1	-1	+1	+1	-1	-1	-1	+1	-1	-1
-1	-1	+1	-1	+1	-1	+1	-1	-1	-1
-1	-1	-1	+1	+1	+1	-1	-1	-1	-1
-1	-1	-1	-1	+1	+1	+1	+1	+1	+1

labeled by (u, v) , with $u, v = 1, \dots, 5, u < v$. The designs $D1, D2, D3$, and $D4$ are able to discriminate between two models in all $\binom{10}{2} = 45$ pairs of models. We now consider the class $\mathbf{D} = \{D\}$ consisting of all fractional factorial designs D with 11 runs that can identify and discriminate between the models in Equation (8.6) (Ghosh, Deng, and Luan 2007). We denote the variance-covariance matrix of the least squares estimators of the 7 β -parameters in Equation (8.6) under model (u, v) for a design D in \mathbf{D} as $\mathbf{V}_D^{(u,v)}$, $u, v = 1, \dots, 5, u < v$. We consider the six criterion functions in Equation (8.13) for comparing fractional factorial designs in \mathbf{D} . Ghosh and Tian (2006) demonstrated that $D1$ is optimum with respect to (AMCR, GMCR), and $D2$ is optimum with respect to (AD, AT, GD, GT).

We now generate the simulated data sets for $y(x_1, x_2, x_3, x_4, x_5)$ using the designs $D1, D2, D3$, and $D4$ when the true model is model (1, 2) with the additional normality assumption, i.e.,

$$y(x_1, x_2, x_3, x_4, x_5) \text{ is } N(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_{12} x_1 x_2, \sigma^2), \quad (8.21)$$

with $\beta_0 = 30.3, \beta_1 = 5.6, \beta_2 = 16.9, \beta_3 = 5.4, \beta_4 = -0.4, \beta_5 = 0.3$, and $\beta_{12} = 3.4$. For each value of $\sigma^2 = 0.05, 0.5, 2.5, 5, 7.5, 10, 12.5, 15, 17.5$, and 20 , we simulate 10,000 data sets satisfying Equation (8.21). To each of the 10,000 simulated data sets for a fixed σ^2 , we fit 10 models (u, v) with the additional distributional assumption, that is,

$$y(x_1, x_2, x_3, x_4, x_5) \text{ is } N(\beta_0 + \beta_1 x_1 + \dots + \beta_5 x_5 + \beta_{uv} x_u x_v, \sigma^2), \quad (8.22)$$

where $u, v = 1, \dots, 5, u < v$. We define for $(u, v), (u', v')$ with $(u, v) \neq (u', v')$ and a fixed σ^2 ,

$$\begin{aligned} W_{(u,v)} &= \text{number of data sets with } MSE_{(u,v)} \leq \min_{(u',v')} MSE_{(u',v')}, \\ W_{(u,v)}^* &= \frac{W_{(u,v)}}{10,000}. \end{aligned} \quad (8.23)$$

The larger the value of $W_{(u,v)}$ or $W_{(u,v)}^*$ relative to $W_{(u',v')}$ or $W_{(u',v')}^*$ means that model (u, v) is stronger than model (u', v') in describing the data. The model giving the largest W and W^* values is in fact the best model in describing the data. Note that $0 \leq W_{(u,v)} \leq 10,000$, and $0 \leq W_{(u,v)}^* \leq 1$. We calculate $W_{(u,v)}$ and $W_{(u,v)}^*$ for 10 models (u, v) .

Figure 8.1 displays the $W_{(1,2)}^*$ values for $D1, D2, D3$, and $D4$. In Figure 8.1, we have plotted $W_{(1,2)}^*$ against the values of σ^2 for all four designs. The designs $D1$ and $D2$ perform very close to each other, and they are the best of the four designs considered, performing better than the design $D3$ and the design $D4$ with respect to their W or W^* values, the criteria in Equation (8.23).

Figure 8.2 displays the difference between the $W_{(1,2)}$ values for $D2$ and $D1$ for $\sigma^2 = 0.5, 2.5, 5, 7.5, 10, 12.5, 15, 17.5$, and 20 . In Figure 8.2, we have plotted the difference between the values of $W_{(1,2)}$ in $D2$ and $D1$ ($= W_{(1,2)} \text{ in } D2 - W_{(1,2)} \text{ in } D1$) against the values of σ^2 .

We observe that the numerical value of $W_{(1,2)}^*$ is greater than or equal to 0.7984 for $\sigma^2 \leq 10$ for both $D1$ and $D2$. The designs $D1$ and $D2$ are comparable with each other for $\sigma^2 \leq 7.5$. The design $D1$ performs slightly better than $D2$ for $0.5 \leq \sigma^2 < 6.1875$, the design $D2$ performs better than $D1$ for $\sigma^2 > 6.1875$. The design $D1$ is indistinguishable from $D2$ for $\sigma^2 \leq 0.5$ and $\sigma^2 = 6.1875$.

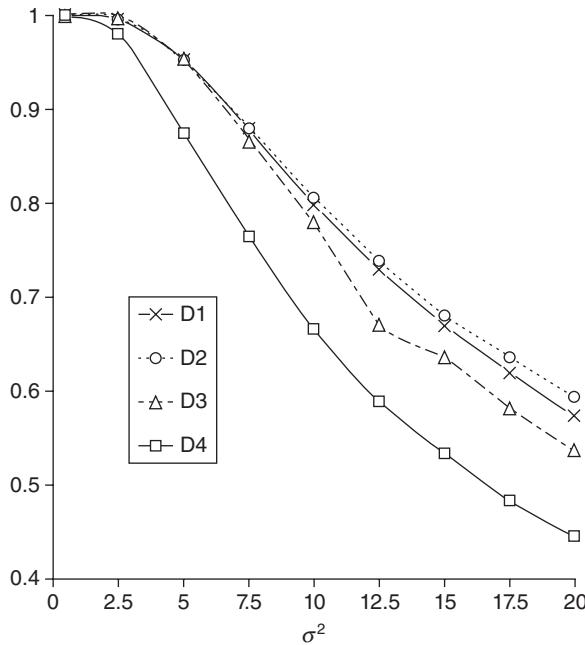


Figure 8.1 Comparison between $D1$, $D2$, $D3$, and $D4$ with respect to their $W_{(1,2)}^*$ values for different σ^2 .

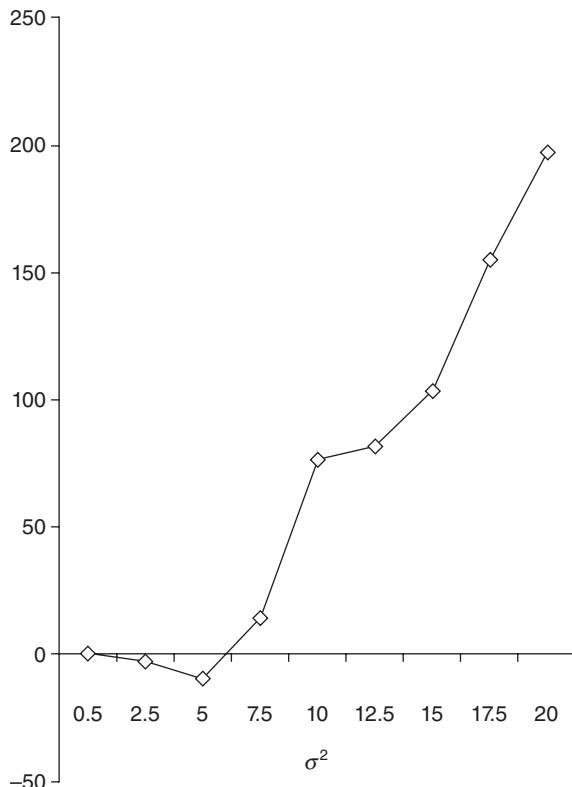


Figure 8.2 Comparison between $D1$ and $D2$ with respect to the difference between their $W_{(1,2)}$ values for different σ^2 .

REFERENCES

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: *Second International Symposium on Information Theory*, B.N. Petrov and F. Csaki (eds.). Budapest: Akadémiai Kiadó, pp. 267–281.
- Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Stat. Math.*, **21**, 243–247.
- Atkinson, A.C., A.N. Donev, and R.D. Tobias (2007). *Optimum Experimental Designs, with SAS*. New York: Oxford University Press.
- Bose, R.C. (1944). The fundamental theorem of linear estimation. In: *Proceedings of the 31st Indian Science Congress*, pp. 2–3 (Abstract).
- Bose, R.C. (1947). Mathematical theory of the symmetrical factorial designs. *Sankhyā*, **8**, 107–166.
- Bose, R.C. (1949). *Least Squares Aspects of Analysis of Variance*, Inst. Stat. Mimeo, Ser. 9, Chapel Hill: University of North Carolina.
- Chatterjee, K. (1990). Search designs for searching for one among the two- and three-factor interaction effects in the general symmetric and asymmetric factorials. *Ann. Inst. Stat. Math.*, **42**, 783–803.
- Chatterjee, K. and R. Mukerjee (1993). Search designs for estimating main effects and searching several two-factor interactions in general factorials. *J. Stat. Plann. Inference*, **37**, 385–392.
- Cheng, C.-S. and R. Mukerjee (1998). Regular fractional factorial designs with minimum aberration and maximum estimation capacity. *Ann. Stat.*, **26**(6), 2289–2300.
- Cheng, C.-S., D.M. Steinberg, and D.X. Sun (1999). Minimum aberration and model robustness for two-level fractional factorial designs. *J. R. Stat. Soc. B*, **61**, 85–93.
- Cheng, C.-S., L.-Y. Deng, and B. Tang (2002). Generalized minimum aberration and design efficiency for nonregular fractional factorial designs. *Stat. Sin.* **12**, 991–1000.
- Fedorov, V.V. (1972). *Theory of Optimal Experiments*. New York: Academic Press.
- Ghosh, S. and C. Burns (2002). Comparison of four new general classes of search designs. *Aust. N. Z. J. Stat.*, **44**(3), 357–366.
- Ghosh, S. and L. Teschmacher (2002). Comparison of search designs using search probabilities. *J. Stat. Plann. Inference*, **104**, 439–458.
- Ghosh, S. and Y. Tian (2006). Optimum fractional factorial plans for model identification and discrimination. *J. Multivariate Anal.*, **97**, 1437–1450.
- Ghosh, S. and X.D. Zhang (1987). Two new series of search designs for 3^m factorial experiments. *Utilitas Math.*, **32**, 245–254.
- Ghosh, S., H. Deng, and Y. Luan (2007). Effects of noise in performance comparisons of designs for model identification and discrimination. *J. Stat. Plann. Inference*, **137**, 3871–3881.
- Ghosh, S., T. Shirakura, and J.N. Srivastava (2007). Model identification using search linear models and search designs. In: *Entropy, Search, Complexity: Bolyai Society Mathematical Studies*, 16:In: I. Csiszár, G.O.H. Katona, and G. Tardos (eds.). Budapest: Springer, pp. 85–112.
- Kiefer, J. (1959). Optimum experimental designs (with discussion). *J. R. Stat. Soc. B*, **21**, 272–319.

- Läuter, E. (1974). Experimental design in a class of models. *Math. Oper. Stat. Ser. Stat.*, **5**, 379–398.
- Pukelsheim, F. (1993). *Optimum Design of Experiments*. New York: Wiley.
- Rao, C.R. (1973). *Linear Statistical Inference and its Applications* (2nd ed.). New York: Wiley.
- Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica*, **46**, 1273–1282.
- Scheffé, H. (1959). *The Analysis of Variance*. New York: Wiley.
- Srivastava, J.N. (1975). Designs for searching non-negligible effects. In: *A Survey of Statistical Design and Linear Models*, J.N. Srivastava (ed.). Amsterdam: North-Holland, Elsevier Science B. V., pp. 507–519.
- Srivastava, J.N. (1977). Optimal search designs, or designs optimal under bias-free optimality criteria. In: *Proc. International Symposium on Decision Theory*, S.S. Gupta and D.S. Moore (eds.). New York: Academic Press, pp. 375–409.
- Srivastava, J.N. and S. Ghosh (1976). A series of 2^m factorial designs of resolution V which allow search and estimation of one extra unknown effect. *Sankhyā Ser. B*, **38**, 280–289.
- Srivastava, J.N. and B.C. Gupta (1979). Main effect plan for 2^m factorials which allow search and estimation of one unknown effect. *J. Stat. Plann. Inference*, **3**, 259–265.
- Srivastava, J.N. and D.M. Mallenby (1984). On a decision rule using dichotomies for identifying non-negligible parameters in certain linear models. *J. Multivariate Anal.*, **16**, 318–334.
- Shirakura, T. and T. Ohnishi (1985). Search designs for 2^m factorials derived from balanced arrays of strength $2(l+1)$ and AD-optimal search designs. *J. Stat. Plann. Inference*, **11**, 247–258.
- Shirakura, T., T. Suetsgu, and T. Tsuji (2002). Construction of main effect plus two plans for 2^m factorials. *J. Stat. Plann. Inference*, **105**, 405–415.
- Shirakura, T., T. Takahashi, and J.N. Srivastava (1996). Searching probabilities for nonzero effects in search designs for the noisy case. *Ann. Stat.*, **24**(6), 2560–2568.

CHAPTER 9

Minimum Aberration and Related Criteria for Fractional Factorial Designs

Hegang H. Chen and Ching-Shui Cheng

9.1 INTRODUCTION

Fractional factorial designs have a long history of successful use in scientific investigations and industrial experiments. This important subject was treated in HK2, chapter 13. Several criteria for choosing fractional factorial designs, including the popular criterion of minimum aberration, were briefly presented in HK2, section 13.3.4, and the issue of optimal blocking of fractional factorial designs was discussed in section 13.8.3. These criteria were proposed for choosing designs with better capability of estimating lower-order effects, albeit with different interpretations of such capability. The differences in the interpretations sometimes lead to inconsistencies or even contradictions among the different criteria, and one should not expect any criterion to work in all circumstances. In this chapter, we give a more comprehensive and in-depth discussion of these criteria, including clarifications of their relationship. We also extend optimal blocking to cover models with random block effects.

For simplicity, we only consider two-level designs, though many of the results can be extended easily to the case where the number of factor levels is a prime power. Suppose there are n two-level treatment factors. Then there are a total of 2^n factor-level combinations, called treatment combinations. A complete factorial requires 2^n runs. A 2^{-m} th fraction, referred to as a 2^{n-m} fractional factorial design, consists of 2^{n-m} of the 2^n treatment combinations. We mainly focus on fractional factorial designs that can be constructed by

Design and Analysis of Experiments: Special Designs and Applications, First Edition. Edited by Klaus Hinkelmann.

© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

using defining relations as discussed in HK2, Section 13.3. Such designs are called *regular fractional factorial designs*. Nonregular designs will be briefly discussed near the end of this chapter.

We first review some notations and basic concepts. Each treatment factor is represented by a letter such as A, B, C, \dots , and each of the $2^n - 1$ factorial effects (main effects and interactions) is represented by a string of letters, called a *word*, consisting of the letters associated with the factors that are involved. For example, the main effect of factor A is also denoted by A , and the interaction of factors A, B , and D is denoted by ABD . The number of letters in a word is called its *length*.

A regular 2^{n-m} fractional factorial design is defined by a set of m independent interactions and is constructed by solving equations such as equation (13.6) in HK2. The words that represent the m independent interactions are called *independent defining words*. Products of independent defining words, subject to the rule that even powers of any letter are deleted, are called *defining words*. There are a total of $2^m - 1$ defining words, and the corresponding factorial effects are called *defining effects* of the fraction. The defining effects constitute the *defining relation*, and cannot be estimated. The other $2^n - 2^m$ factorial effects are partitioned into $2^{n-m} - 1$ alias sets each of size 2^m . Each of these effects is estimable if all its aliases are assumed to be negligible.

Example 9.1. Consider the 2^{6-2} design d defined by two independent defining effects $ABCE$ and $BCDF$. Then $ADEF = (ABCE)(BCDF)$ is also a defining effect. The defining relation is written as

$$I = ABCE = BCDF = ADEF.$$

The 60 factorial effects other than the three defining effects are partitioned into 15 alias sets each of size 4, where the aliases of each factorial effect can be obtained by multiplying all the defining words by the word representing that effect:

$$A = BCE = ABCDF = DEF,$$

$$B = ACE = CDF = ABDEF,$$

$$C = ABE = BDF = ACDEF,$$

$$D = ABCDE = BCF = AEF,$$

$$E = ABC = BCDEF = ADF,$$

$$F = ABCEF = BCD = ADE,$$

$$AB = CE = ACDF = BDEF,$$

$$AC = BE = ABDF = CDEF,$$

$$AD = BCDE = ABCF = EF,$$

$$\begin{aligned}
&\underline{AE} = \underline{BC} = \underline{ABCDEF} = \underline{DF}, \\
&AF = BCEF = ABCD = DE, \\
&BD = ACDE = CF = ABEF, \\
&BF = ACEF = CD = ABDE, \\
&\underline{ABD} = \underline{CDE} = \underline{ACF} = \underline{BEF}, \\
&\underline{ABF} = \underline{CEF} = \underline{ACD} = \underline{BDE}.
\end{aligned}$$

Note that three alias sets are underlined. This is not relevant here and is to be discussed in Example 9.8 (Section 9.10) when we present blocked fractional factorial designs.

An important property of a 2^{n-m} fractional factorial design is its *resolution*. Box and Hunter (1961) defined the resolution of a regular fractional factorial design to be the length of the shortest defining word. We say that d is a 2_r^{n-m} design if it is a 2^{n-m} fractional factorial design of resolution r . Under a design of resolution r , no s -factor interaction is aliased with any other effect involving less than $r-s$ factors. Under the hierarchical assumption that lower-order effects are more important than higher-order effects and that effects of the same order are equally important, the experimenter may prefer a design that has the highest possible resolution. The design in Example 9.1 is a resolution IV design; all the main effects are aliased with some three-factor and higher-order interactions.

However, not all 2^{n-m} designs of maximum resolution are equally good. Fries and Hunter (1980) introduced the *minimum aberration* criterion for further discriminating between 2^{n-m} designs of the same resolution. For each regular 2^{n-m} fractional factorial design d , let $A_i(d)$ be the number of defining words of length i and $W(d)$ be the vector

$$W(d) = (A_1(d), A_2(d), \dots, A_n(d)).$$

Then $W(d)$ is called the *wordlength pattern* of d . Given two 2^{n-m} fractional factorial designs d_1 and d_2 , d_1 is said to have less aberration than d_2 if $A_s(d_1) < A_s(d_2)$, where s is the smallest integer such that $A_s(d_1) \neq A_s(d_2)$. A 2^{n-m} design has minimum aberration if no other 2^{n-m} design has less aberration. In other words, the criterion of minimum aberration sequentially minimizes $A_1(d)$, $A_2(d)$, \dots , and so on. Chen and Hedayat (1996) proposed a weaker version of minimum aberration. A regular 2^{n-m} design of maximum resolution r_{max} is said to have *weak minimum aberration* if it minimizes the number of words of length r_{max} among all the designs of resolution r_{max} . Intuitively, minimum and weakly minimum aberration designs are expected to produce less aliasing among lower-order effects, a desirable feature under the hierarchical assumption. The design in Example 9.1 has minimum aberration among all the 2^{6-2} designs.

For any 2^{n-m} design d , let \tilde{d} be the design obtained from d by switching the two levels. Then the treatment combinations in d and \tilde{d} together form a regular $2^{n-(m-1)}$ design, called the *foldover* of d . If d is a resolution III design, then the foldover of d has resolution IV, and all its defining words are of even lengths (see HK2, section 13.6.4). Designs that only have defining words of even lengths are called even designs. One can also add a factor at constant level to d . Then the foldover of the resulting $2^{n+1-(m+1)}$ design is a $2^{(n+1)-m}$ even design of resolution IV. More generally, the foldover of a design of odd resolution r has resolution $r+1$.

Throughout this chapter, we also denote the run size 2^{n-m} of a 2^{n-m} fractional factorial design by N . To prevent aliasing among the main effects, we only consider designs of resolution III or higher. Such designs, called *resolution III+ designs*, must have $n \leq N-1$. Resolution III designs with $n = N-1$ are unique up to isomorphism, and are called *saturated designs*. It can be shown that designs of resolution IV+ must have $n \leq N/2$. Resolution IV designs with $n = N/2$ are also unique up to isomorphism. Such a design can be constructed by folding over a saturated design of run size $N/2$ that is supplemented by a factor at constant level, as described in the previous paragraph. Therefore, resolution IV designs with $n = N/2$ are even designs. It can be shown that every even design of size N is a foldover design and can be constructed by deleting factors from a resolution IV design with $N/2$ factors if $n < N/2$. Therefore, we call resolution IV designs with $N/2$ factor *maximal even designs*.

In Section 9.2, we discuss projections of regular fractional factorial designs onto subsets of factors, and show that minimum aberration designs have good projection properties. In Section 9.3, we provide a better understanding of minimum aberration by investigating alias structures of minimum aberrations designs. This is crucial for justifying minimum aberration as a good surrogate for the criterion of maximum estimation capacity under model uncertainty. Two other criteria, number of clear two-factor interactions and estimation index, are presented in Sections 9.4 and 9.5, respectively. Relationship between aberration, estimation capacity, and estimation index is addressed in Section 9.6. In Section 9.7, we present a method of constructing minimum aberration designs via their complementary designs. After a review of some properties of orthogonal arrays as well as extending the concepts of estimation capacity and clear two-factor interactions to nonregular designs in Section 9.8, a brief introduction to generalized minimum aberration, a natural extension of the minimum aberration criterion to nonregular designs, is presented in Section 9.9. The last section of this chapter is devoted to optimal blocking.

9.2 PROJECTIONS OF FRACTIONAL FACTORIAL DESIGNS

In factor screening experiments, among a large number of factors to be examined, typically only a few are expected to be active. Therefore, information about the design when it is restricted to a small number of factors is valuable.

Let d be a 2^{n-m} design. For any k of the n factors, the design obtained by dropping the other $n-k$ factor is called a k -dimensional projection of d .

Let d be a 2_r^{n-m} design. Since the defining words of a lower-dimensional projection of d are also defining words of d , all lower-dimensional projections of d are of resolution r or higher. We first consider the r -dimensional projections. Box and Hunter (1961) pointed out that any k -dimensional projection with $k < r$ is a replicated complete 2^k design. Projections onto r factors are replicated fractional factorial designs of resolution r if the r factors form a defining word, and are full factorials or replicated full factorials otherwise. Specifically, each defining word of length r gives an r -dimensional projection that consists of $2^{n-m-r+1}$ replicates of a 2_r^{r-1} design. There are $A_r(d)$ such r -dimensional projections. Each of the remaining $\binom{n}{r} - A_r(d)$ projections consists of 2^{n-m-r} copies of a full 2^r factorial.

Example 9.2. Let d be a 2_V^{8-2} design with the following defining relation and wordlength pattern:

$$\begin{aligned} I = ABCDG = ABFH = CDEFGH, \\ W(d) = (0, 0, 0, 0, 2, 1, 0, 0). \end{aligned}$$

Then all the k -dimensional projections of d with $k < 5$ are 2^{6-k} replicates of a full 2^k factorial. Projections onto the two subsets of five factors $\{A, B, C, D, G\}$ and $\{A, B, E, F, H\}$, which form the two defining words of length 5, respectively, are replicated 2_V^{5-1} designs, and each of the other five-dimensional projections consists of two replicates of a full 2^5 factorial.

Chen (1998) studied the relationship between projections of a fractional factorial design and its wordlength pattern, and completely characterized the projections of a 2_r^{n-m} design onto $r+1$ to $r+[(r-1)/2]$ dimensions.

Theorem 9.1. (Chen 1998) Let d be a 2_r^{n-m} design. For $r+1 \leq k \leq r+[(r-1)/2]$, each k -dimensional projection of d is a possibly replicated 2^k design, or a replicated 2^{k-1} design. For $j = r, r+1, \dots, k, \binom{n-j}{k-j} A_j(d)$ of the $\binom{n}{k}$ k -dimensional projections consist of $2^{n-m-k+1}$ replicates of a 2_j^{k-1} design, and the other $\binom{n}{k} - \sum_{j=r}^k \binom{n-j}{k-j} A_j(d)$ k -dimensional projections consist of 2^{n-m-k} copies of a 2^k design.

In Example 9.2, we have discussed the k -dimensional projections of a 2_V^{8-2} design for all $k \leq 5$. Theorem 9.1 can be used to determine the six- and seven-dimensional projections. For $k = 6$, six projections consist of two copies of a 2_6^{6-1} design, one projection consists of two replicates of a 2_{VI}^{6-1} design, and the remaining projections are 2^6 designs. Among the eight seven-dimensional projections, six are 2_V^{7-1} designs, and two are 2_{VI}^{7-1} designs.

Theorem 9.1 shows that for $r+1 \leq k \leq r + [(r-1)/2]$, the number of k -dimensional projections that have resolution j , $r \leq j \leq k$, is proportional to $A_j(d)$. Therefore by sequentially minimizing the $A_j(d)$ s, minimum aberration designs have good projection properties in that they produce fewer projections of low resolutions.

Example 9.3. There are three nonisomorphic 2_{IV}^{7-2} designs:

$$\begin{aligned} d_1 &: I = ABCF = BCDG = ADFG, \\ &\quad W(d_1) = (0, 0, 0, 3, 0, 0, 0), \\ d_2 &: I = ABCF = ADEG = BCDEFG, \\ &\quad W(d_2) = (0, 0, 0, 2, 0, 1, 0), \\ d_3 &: I = DEFG = ABCDF = ABCEG, \\ &\quad W(d_3) = (0, 0, 0, 1, 2, 0, 0). \end{aligned}$$

It can be shown that the maximum attainable resolution for a 2^{7-2} design is IV, and d_3 has minimum aberration. Design d_3 has only one replicated 2_{IV}^{4-1} among its four-dimensional projections, while d_1 and d_2 have three and two such projections, respectively. Among the 21 five-dimensional projections of d_3 , three are replicated 2_{IV}^{5-1} designs, and two consist of two copies of a 2_V^{5-1} design, whereas d_1 has nine projections that are replicated 2_{IV}^{5-1} designs, and d_2 has six such projections.

9.3 ESTIMATION CAPACITY

In Section 9.2, we showed that minimum aberration designs have good projection properties. In this section, we introduce the criterion of estimation capacity, and provide another justification of minimum aberration by showing that it is a good surrogate for maximum estimation capacity. We first investigate the alias structures of minimum aberration designs.

Let $g = 2^{n-m} - 1$. Then under a 2^{n-m} design d of resolution III+, $g-n$ of the g alias sets do not contain main effects. Let $f = g-n$, and, without loss of generality, assume that the first f alias sets do not contain main effects. For $1 \leq i \leq g$, let $m_i(d)$ be the number of two-factor interactions in the i th alias set.

From each defining word of length 3, say ABC , we can identify three two-factor interactions, AB , AC , and BC , that are aliased with main effects. It follows that under any design d of resolution III+, the number of two-factor interactions that are not aliased with main effects is equal to

$$\sum_{i=1}^f m_i(d) = \binom{n}{2} - 3A_3(d). \quad (9.1)$$

Cheng, Steinberg, and Sun (1999) further showed that

$$A_4(d) = \frac{1}{6} \left\{ \sum_{i=1}^g [m_i(d)]^2 - \binom{n}{2} \right\}. \quad (9.2)$$

It follows from Equations (9.1) and (9.2) that a minimum aberration design of resolution III+ maximizes $\sum_{i=1}^f m_i(d)$, and minimizes $\sum_{i=1}^g [m_i(d)]^2$ among those which maximize $\sum_{i=1}^f m_i(d)$. The second step tends to make the $m_i(d)$ s as equal as possible. Then since $\sum_{i=1}^f m_i(d)$ is equal to the number of two-factor interactions that are not aliased with main effects, one can conclude that a minimum aberration design of resolution III+ maximizes the number of two-factor interactions that are not aliased with main effects, and tend to distribute these two-factor interactions very uniformly over the alias sets that do not contain main effects. Likewise, a minimum aberration design of resolution V+ maximizes the number of three-factor interactions that are not aliased with two-factor interactions among the resolution V+ designs, and tend to distribute these three-factor interactions very uniformly over the alias sets that do not contain main effects and two-factor interactions, and so on. This property is important for understanding the statistical meaning of minimum aberration and relating it to the criterion of estimation capacity, which we now define.

Estimation capacity was introduced by Sun (1993) as a measure of the capability of a design d to handle and estimate different potential models involving interactions. For simplicity, assume that the main effects are of primary interest and their estimates are required. Furthermore, all the three-factor and higher-order interactions are assumed to be negligible. For any $1 \leq k \leq \binom{n}{2}$, let $E_k(d)$ be the number of models containing all the main effects and k two-factor interactions such that all the effects in the model are estimable under d , where k can be thought of as the number of active two-factor interactions. It is desirable to have $E_k(d)$ as large as possible. A design d_1 is said to dominate another design d_2 if $E_k(d_1) \geq E_k(d_2)$ for all k , with strict inequality for at least one k . We say that d has *maximum estimation capacity* if it maximizes $E_k(d)$ for all k . See also Chapter 8.

It turns out that $E_k(d)$ is a function of $\mathbf{m}(d) = (m_1(d), \dots, m_f(d))$. It is easy to see that $E_k(d) = 0$ for $k > f$, and

$$E_k(d) = \sum_{1 \leq i_1 < \dots < i_k \leq f} \prod_{j=1}^k m_{i_j}(d), \quad \text{if } k \leq f. \quad (9.3)$$

In other words, $E_k(d)$ is the k th elementary symmetric function of the $m_i(d)$ s.

Table 9.1 $m_1(d), \dots, m_f(d)$ for Minimum Aberration $2^{n-(n-5)}$ Designs with $9 \leq n \leq 29$

n	r	f	$m_1(d), \dots, m_f(d)$
9	4	22	1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,4,0
10	4	21	2,5
11	4	20	3,3,3,3,4,4,4,4,4,4,4,4,4,4,4,0,0,0,0,0
12	4	19	4,4,4,4,4,4,4,4,4,5,5,5,5,5,5,5,0,0,0,0
13	4	18	5,5,5,5,5,5,5,5,5,5,5,6,6,0,0,0
14	4	17	6,6,6,6,6,6,6,6,6,6,6,6,6,6,7,0,0
15	4	16	7,7,7,7,7,7,7,7,7,7,7,7,7,7,7,7,7,0
16	4	15	8,8,8,8,8,8,8,8,8,8,8,8,8,8,8,8,8
17	3	14	8,8,8,8,8,8,8,8,8,8,8,8,8,8,8,8,8
18	3	13	8,8,8,8,8,8,8,8,8,8,8,8,8,8,8,9
19	3	12	8,8,8,8,8,8,8,8,9,9,9,9
20	3	11	8,8,8,8,9,9,9,9,9,9,9,9
21	3	10	9,9,9,9,9,9,9,9,9,9,9
22	3	9	8,8,10,10,10,10,10,10,10,11
23	3	8	8,11,11,11,11,11,11,11,11
24	3	7	12,12,12,12,12,12,12
25	3	6	12,12,12,12,12,12
26	3	5	12,12,12,12,13
27	3	4	12,13,13,13
28	3	3	14,14,14
29	3	2	14,14

r denotes resolution.

Cheng, Steinberg, and Sun (1999) argued that a design d^* has large estimation capacity if it (1) maximizes $\sum_{i=1}^f m_i(d)$, and (2) the $m_i(d^*)$ s are as equal as possible. This is because, by proposition F.1 on page 78 of Marshall and Olkin (1979), $E_k(d)$ as given in Equation (9.3) is a Schur concave function of $\mathbf{m}(d) = (m_1(d), \dots, m_f(d))$ and is nondecreasing in each component of $\mathbf{m}(d)$. By the discussion in the paragraph following Equation (9.2), minimum aberration is a good surrogate for maximum estimation capacity.

Table 9.1 shows the values of $m_i(d)$ s, $1 \leq i \leq f$, for 32-run minimum aberration designs with $9 \leq n \leq 29$. For $16 \leq n \leq 21$ and $24 \leq n \leq 29$, the $m_i(d)$ values of minimum aberration $2^{n-(n-5)}$ designs differ from one another by at most one. This distribution is the most uniform possible. It follows from the Schur-concavity of $E_k(d)$ that these designs maximize $E_k(d)$ for all k ; a little extra work shows that it is also true for $n = 22$ and 23. That is, for $n \geq 16$, minimum aberration 32-run designs have maximum estimation capacity (Cheng, Steinberg, and Sun 1999). However, for $n < 16$, minimum aberration designs typically do not maximize $E_k(d)$ for all k , but they still maximize $E_k(d)$ unless k is too large. Note that for $n < 16$, the minimum aberration designs are of resolution IV. This points to an important difference between resolution III and IV designs. We shall return to this in Sections 9.5 and 9.6.

Cheng and Mukerjee (1998) obtained some general results on the construction of designs with maximum estimation capacity, in particular, those with the $m_i(d)$ s differing from one another by at most 1.

9.4 CLEAR TWO-FACTOR INTERACTIONS

It is prudent to include the main effects and a set of judiciously selected two-factor interactions in a first approximation to the response model. Therefore, resolution III and IV designs are important for factor screening experiments. Wu and Chen (1992) classified two-factor interactions into three categories—ineligible, eligible, and clear. A two-factor interaction is called *ineligible* if it is aliased with at least one main effect, *eligible* if it is not aliased with any main effect, and *clear* if it is neither aliased with main effects nor aliased with other two-factor interactions. Clearly, a two-factor interaction is ineligible if it is part of a defining word of length 3, eligible if it is not part of a defining word of length 3, and clear if it is not part of a defining word of length 3 or 4. It is obvious that if estimates of all the main effects are required, then ineligible two-factor interactions are not estimable. On the other hand, if we can assume that the two-factor and higher order interactions aliased with an eligible two-factor interaction are negligible, then that eligible two-factor interaction can be estimated. Under the assumption of negligible three-factor and higher order interactions, all clear two-factor interactions are estimable. We use two examples to illustrate these concepts.

Example 9.4. There are five nonisomorphic 2^{9-5}_{III} designs (see Chen, Sun, and Wu 1993) with the following independent defining words and wordlength patterns:

$$\begin{aligned}
 d_1 : \quad & I = ABCDE = ABF = BCG = CDH = ABCJ \\
 & W(d_1) = (0, 0, 7, 9, 6, 6, 3, 0, 0) \\
 d_2 : \quad & I = ABCDE = ABF = ACG = BCH = ABCJ \\
 & W(d_2) = (0, 0, 8, 10, 4, 4, 4, 1, 0) \\
 d_3 : \quad & I = ABCDE = ABF = ACG = ADH = ABCJ \\
 & W(d_3) = (0, 0, 6, 10, 8, 4, 2, 1, 0) \\
 d_4 : \quad & I = ABCDE = ABF = ADG = BCH = CDJ \\
 & W(d_4) = (0, 0, 6, 9, 9, 6, 0, 0, 1) \\
 d_5 : \quad & I = ABCDE = ABF = ACG = ADH = BCDJ \\
 & W(d_5) = (0, 0, 4, 14, 8, 0, 4, 1, 0).
 \end{aligned}$$

In this case, the maximum resolution is three, and there are no clear two-factor interactions under any of the five resolution III designs. The minimum

aberration design d_5 has four defining words of length 3 and 14 defining words of length 4:

$$\begin{aligned} I &= ABF = ACG = ADH = AEJ \\ &= BCFG = BDFH = BEFJ = CDGH = CEGJ = DEHJ = BCDJ \\ &= BCEJ = BDEG = BGHJ = CDEF = CFHJ = DFGJ = EFGH = \dots \end{aligned}$$

The 12 two-factor interactions $AB, AC, AD, AE, AF, AG, AH, AJ, BF, CG, DH, EJ$ are ineligible, and the other 24 two-factor interactions are eligible.

The following example shows that when both resolution III and IV designs exist, designs of higher resolution do not necessarily have more clear two-factor interactions.

Example 9.5. There are three 2_{III}^{6-2} resolution III designs and one 2_{IV}^{6-2} resolution IV design (Chen, Sun, and Wu 1993). Consider the following two designs, where d_1 is of resolution III, and d_2 is of resolution IV:

$$\begin{aligned} d_1 : \quad I &= ABC = ADEF = BCDEF \\ W(d_1) &= (0, 0, 1, 1, 1, 0) \\ d_2 : \quad I &= ABCE = BCDF = ADEF \\ W(d_2) &= (0, 0, 0, 3, 0, 0). \end{aligned}$$

Note that d_2 is the design in Example 9.1. From the alias sets given in Example 9.1, it can be seen that under d_2 , there is no clear two-factor interaction. However, there are six clear two-factor interactions, $\{BD, BE, BF, CD, CE, CF\}$, under the resolution III design d_1 .

Not only designs of higher resolution may have fewer clear two-factor interactions, Chen, Sun, and Wu (1993) listed several examples of minimum aberration designs that are not the best in terms of the number of clear two-factor interactions. This seems surprising at the first look, since “intuitively, minimum aberration designs should be the best designs with respect to the estimation of two-factor interactions” (HK2, p. 522). However, it is not surprising once one realizes that minimum aberration designs tend to distribute the two-factor interactions that are not aliased with main effects nearly uniformly over the alias sets, as demonstrated in the previous section. A two-factor interaction is clear if and only if it is the only two-factor interaction in an alias set that does not contain any main effect. Thus, the number of clear two-factor interactions is equal to the number of $m_i(d)$ s, $1 \leq i \leq f$, that are equal to 1. For a given run size, unless the number of factors is small, a minimum aberration design, due to nearly uniform distribution of the two-factor interactions over the alias sets not containing main effects, will have more than one two-factor interaction in each of such alias sets, resulting in no clear two-factor interac-

tion. This shows that minimum aberration typically runs counter to large numbers of clear two-factor interactions. We have demonstrated in the previous section that minimum aberration is a good surrogate for maximum estimation capacity, a criterion under model uncertainty. Indeed, as Fries and Hunter (1980) indicated, minimum aberration was for “situations in which prior knowledge is diffuse concerning the possible greater importance of certain effects.” On the other hand, to take advantage of clear two-factor interactions, one needs to have some knowledge about which two-factor interactions might be important. The two criteria are really designed for different objectives.

Example 9.4 shows that there is no clear two-factor interaction under any 2^{9-5}_{III} design. Chen and Hedayat (1998) provided a complete characterization of the existence of clear two-factor interactions under 2^{n-m} designs of resolution III or IV, and revealed their structures. For a fixed number of runs $N = 2^{n-m}$, the maximum resolution of any 2^{n-m} fractional factorial design is equal to III when $N/2 < n \leq N-1$. It is natural to ask whether in this case a 2^{n-m} resolution III design contains any clear two-factor interaction. The following theorem gives a negative answer to this question.

Theorem 9.2. When $N/2 < n \leq N-1$, there is no clear two-factor interaction under any 2^{n-m} design of resolution III.

Although when the maximum resolution is III, no resolution III design can have clear two-factor interactions, there may be eligible two-factor interactions. As shown in the previous section, for any design d , the number of eligible two-factor interactions is equal to $\binom{n}{2} - 3A_3(d)$. By minimizing $A_3(d)$, resolution III designs with (weak) minimum aberration maximize the number of eligible two-factor interactions. In Example 9.4, we observed that there is no clear two-factor interaction under any 2^{9-5} resolution III design. However, the minimum aberration design d_5 yields the maximum number (24) of eligible two-factor interactions.

The following two theorems from Chen and Hedayat (1998) characterize the existence of clear two-factor interactions under resolution IV designs.

Theorem 9.3. When $N/4 + 1 < n \leq N/2$, there is no clear two-factor interaction under any 2^{n-m} design of resolution IV, but there exist resolution III 2^{n-m} designs with clear two-factor interactions.

Let $n(k)$ be the maximum number of factors n that can be accommodated in a $2^{n-(n-k)}$ design of resolution V+. Then for $n(k) < n \leq 2^{k-1}$, the maximum resolution of a $2^{n-(n-k)}$ design is IV.

Theorem 9.4. For $n(k) < n \leq N/4 + 1$, where $N = 2^k$, there exist $2^{n-(n-k)}$ resolution IV designs that have clear two-factor interactions.

It follows from Theorems 9.2, 9.3, and 9.4 that for 32 runs, no regular design with $16 < n \leq 31$ can have clear two-factor interactions; for $10 \leq n \leq 16$, no

resolution IV design can have clear two-factor interactions, but there are resolution III designs with clear two-factor interactions; for $6 < n \leq 9$, there are resolution III and resolution IV designs that have clear two-factor interactions.

Some lower and upper bounds for the number of clear two-factor interactions were derived in Tang et al. (2002). Wu and Wu (2002) developed an approach to show whether a given design has the maximum number of clear two-factor interactions. Zhang et al. (2008) defined a general minimum lower-order confounding (GMC) criterion, which can be viewed as a refined version of the criterion of maximizing the number of clear two-factor interactions.

9.5 ESTIMATION INDEX

In a factorial experiment, there may be a large number of interactions that are potentially important. Then we should choose a design that allows the estimation of as many such interactions as possible. For example, in a robust design experiment (see Chapter 13) to study the effects of control and noise factors on certain responses of a product or process, we prefer a design that can be used to entertain models that contains all the main effects of control and noise factors and as many of their interactions as possible.

Example 9.6. Suppose we wish to perform an experiment with 16 runs and six two-level factors, labeled A, B, C, D, E , and F , where A, B , and C are control factors, and the others are noise factors. The estimability of main effects and control-noise interactions is the primary concern. There are four nonisomorphic 2^{6-2} designs, among which the design in Example 9.1 (d_2 in Example 9.5) has minimum aberration. Again, from the alias sets given in Example 9.1 we can see that under d_2 , we are able to estimate at most seven two-factor interactions, and therefore, it is not possible to entertain all nine control-noise interactions. This is because two-factor interactions appear in only seven of the nine alias sets that do not contain main effects. On the other hand, under the more aberration resolution III design d_1 in Example 9.5, none of the main effects and control-noise two-factor interactions are aliased among themselves. Therefore, all the main effects and control-noise two-factor interactions are estimable if the other interactions are negligible. Although as a design of lower resolution, d_1 causes aliasing of some main effects and two-factor interactions, in many circumstances, an experimenter may have prior knowledge that certain interactions are negligible. If the goal of the experiment is to explore as many interaction effects as possible, then selection decisions based on the wordlength pattern alone may not be sufficient. One should also examine the alias structure of the design.

One can see from Table 9.1 that for all the 32-run minimum aberration designs with $n \geq 16$, all the $m_i(d)$ s, $1 \leq i \leq f$, are positive. That is, there is at least one two-factor interaction in every alias set that does not contain main effects.

Under such designs, one can entertain models that contain all the main effects and up to $f = 2^{n-m} - n - 1$ two-factor interactions, using up all the available degrees of freedom. On the contrary, most of the minimum aberration resolution IV designs with $n \leq 15$ have some zero $m_i(d)$ s. For example, under the minimum aberration 2^{11-6} design, 15 of the $m_i(d)$ s are nonzero and 5 are zero. Such a design can be used to entertain only up to 15 two-factor interactions, even though there are 20 degrees of freedom that are not aliased with main effects. In Example 9.6, the minimum aberration design d_2 has 7 nonzero $m_i(d_2)$ s with $1 \leq i \leq 9$, but for d_1 , all the 9 $m_i(d_1)$ s with $1 \leq i \leq 9$ are not equal to zero.

We define the length of an alias set to be the length of the shortest word in the set. The *estimation index* of a design d , denoted by $\rho(d)$, is defined as the largest length of all the alias sets.

For a resolution III+ design, clearly, a necessary and sufficient condition for $m_i(d) > 0$ for all $1 \leq i \leq f$, (i.e., there is at least one two-factor interaction in each alias set that does not contain main effects) is that $\rho(d) = 2$. So each resolution III+ design with estimation index 2 can be used to entertain some models containing all the main effects and up to $2^{n-m} - n - 1$ two-factor interactions, the largest number possible, if all the other interactions are negligible. Such designs are called second-order saturated designs by Block and Mee (2003). If the estimation index is greater than 2, then fewer than $2^{n-m} - n - 1$ two-factor interactions can be entertained. As Example 9.6 shows, for resolution IV designs, estimation index 2 may not be achievable.

Example 9.6 (Revisited). In Example 9.6, the maximum attainable resolution is IV. The minimum aberration resolution IV design d_2 has estimation index 3 since it has six alias sets of length 1, seven alias sets of length 2, and two alias sets of length 3. The lower resolution design d_1 has estimation index 2, since it has six alias sets of length 1 and nine alias sets of length 2.

From the definitions of resolution and alias sets, it is not difficult to establish the following relationship between resolution and estimation index.

Proposition 9.1. Let d be a 2^{n-m} design of resolution r and estimation index ρ . Then

$$\rho \geq [(r-1)/2], \quad (9.4)$$

where $[x]$ is the largest integer less than or equal to x .

Under a design with $\rho = [(r-1)/2]$, one can estimate all the effects involving at most $(r-1)/2$ factors if the higher-order interactions are negligible. Except for a few cases, equality in Equation (9.4) usually does not hold. However, a 2^{n-m} design with $\rho = [(r-1)/2] + 1$ should be fairly good. For $r = 3$ or 4, $[(r-1)/2] + 1 = 2$.

The estimation index of a saturated design ($n = 2^{n-m} - 1$) is equal to 1; this is because under such a design, every alias set contains one main effect. All

the other resolution III+ designs have $\rho(d) \geq 2$. We have seen that $\rho(d) = 2$ if and only if $m_i(d) > 0$ for all $1 \leq i \leq f$. An interesting fact is that this lower bound 2 is always achieved as long as $N/2 < n < N - 1$. Note that this is when the maximum possible resolution is three.

Theorem 9.5. (Chen and Cheng 2004) If $N/2 < n < N - 1$, then all resolution III 2^{n-m} designs achieves the minimum possible estimation index 2.

In Table 9.1, for all the 32-run minimum aberration designs of resolution III (those with $n > 16$), we do have $m_i(d) > 0$ for all $1 \leq i \leq f$.

Resolution IV designs exist when $n \leq N/2$. In this case, the following result holds.

Theorem 9.6. (Chen and Cheng 2004) If $N/4 + 1 \leq n \leq N/2$, then any resolution IV 2^{n-m} design has estimation index at most 3.

The resolution IV designs covered by Theorem 9.6 can have estimation indices 2 or 3. Unlike resolution III designs, in general, most of these resolution IV designs have estimation index equal to 3. For 32-run designs, resolution IV designs with estimation index 2 exist only when $n = 9, 10$ and 16 (see Table 9.1). This is a consequence of some results from coding theory and finite projective geometry, which have important implications in the structure and construction of resolution IV designs. A brief account is given below. The readers are referred to Chen and Cheng (2004, 2006) for more detailed discussions.

A design of resolution IV+ is called *maximal* if and only if its resolution reduces to three whenever a factor is added. Maximal designs are important since all nonmaximal designs of resolution IV can be obtained from maximal ones by deleting some factors, that is, they are projections of maximal designs.

We state two theorems, which are translations of results from coding theory and finite projective geometry (Davydov and Tombak 1990; Bruen, Haddad, and Wehlau 1998; Bruen and Wehlau 1999) into design language.

Theorem 9.7. For any regular design d of resolution IV+, the following conditions are equivalent:

- (a) d has estimation index 2;
- (b) d is maximal;
- (c) d is second-order saturated.

Theorem 9.7 shows that nonmaximal resolution IV designs are not second-order saturated and have estimation index greater than 2. One intuitive explanation is as follows. The nonmaximal designs must be constructed by deleting factors from maximal ones. When some factors are dropped, it does not change the alias sets where interactions of the other factors are located. Thus, the number of nonzero $m_i(d)$ s does not increase even though extra degrees of freedom become available due to elimination of some main effects. For

example, a maximal 2^{16-11} design of resolution IV has all the two-factor interactions distributed over 15 alias sets. A 2^{11-6} design of resolution IV can be constructed by deleting five factors from the maximal 2^{16-11} design of resolution IV. It still has 15 nonzero $m_i(d)$ s, but after deleting five factors from a 2^{16-11} design, we have five more, that is, 20, alias sets that do not contain main effects.

Suppose the two levels of each factor are denoted by 1 and -1 , and each fractional factorial design d is represented by an $N \times n$ matrix $\mathbf{X}(d)$, with each row corresponding to a run and each column corresponding to a factor. Then the matrix

$$\begin{bmatrix} \mathbf{X}(d) & \mathbf{X}(d) \\ \mathbf{X}(d) & -\mathbf{X}(d) \end{bmatrix},$$

represents a design with $2N$ runs and $2n$ factors. We call this design the double of d .

Theorem 9.8. For $N/4 + 1 \leq n \leq N/2$, where $N = 2^k$, $k \geq 4$, maximal designs of resolution IV+ exist if and only if $n = N/2$ or $n = (2^i + 1)N/2^{i+2}$ for some integer i such that $2 \leq i \leq k - 2$. A maximal design with $n = (2^i + 1)N/2^{i+2}$ can be obtained by repeatedly doubling a maximal regular resolution IV+ design with 2^{i+2} runs and $2^i + 1$ factors $k - i - 2$ times.

By Theorems 9.7 and 9.8, for $N = 32$, maximal designs of resolution IV+ exist only for $n = 9, 10$, and 16 , as noted before. For all the other n values, no resolution IV 2^{n-m} design is maximal, and hence all the resolution IV 2^{n-m} designs must have estimation indices greater than two. By Theorem 9.6, they are equal to 3. Also, the maximal resolution IV+ designs in Theorem 9.8 can be constructed by repeatedly doubling maximal designs of smaller run sizes. For example, all maximal designs with $n = 5N/16$ can be constructed by repeatedly doubling the 2^{5-1} design defined by $I = ABCDE$.

Remark 9.1. Besides folding over a saturated design, the maximal even design can also be constructed by repeatedly doubling the 2^2 complete factorial

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 1 \\ -1 & -1 \end{bmatrix}.$$

Saturated designs can be constructed by repeatedly doubling

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

followed by deletion of the first column.

Some important applications of Theorems 9.7 and 9.8 will be given in Section 9.7. Theorem 9.4 shows that $N/4 + 1$ is the maximum number of factors for resolution IV designs to have clear two-factor interactions. One class of resolution IV designs with $n = N/4 + 1$ was shown by Wu and Wu (2002) to have the maximum number of clear two-factor interactions. It can be seen that they are maximal designs.

9.6 ESTIMATION INDEX, MINIMUM ABERRATION, AND MAXIMUM ESTIMATION CAPACITY

Minimum aberration was shown in Section 9.3 to be a good surrogate for maximum estimation capacity, a criterion for choosing designs that can be used to entertain the largest number of models containing a given number of two-factor interactions. However, minimum aberration and maximum estimation capacity do not always coincide. We have already pointed out in Section 9.3 that for 32 runs, all minimum aberration designs of resolution III (those with $n > 16$) have maximum estimation capacity, but minimum aberration designs of resolution IV (those with $n \leq 16$), except for $n = 16$, maximize $E_k(d)$ unless k is too large, but may not maximize $E_k(d)$ for larger k 's. This is closely related to the phenomenon observed in the previous section that for $N/2 < n < N - 1$, all the resolution III designs have estimation index 2, but for $n < N/2$, most of the resolution IV designs have estimation index greater than 2.

When $N/2 < n < N - 1$, since none of the $m_i(d)$'s is constrained to be equal to zero, we can say that the $m_i(d)$'s for a minimum aberration resolution III design are truly nearly equal. In this case, we do expect high consistency between minimum aberration and maximum estimation capacity. We wonder whether all minimum aberration 2^{n-m} designs with $N/2 < n < N - 1$ maximize $E_k(d)$'s for all k . This is known to be true for 16- and 32-run designs (Cheng, Steinberg, and Sun 1999).

Under a resolution IV design with $n = N/2$, the maximal even design, the $\binom{n}{2}$ two-factor interactions are distributed uniformly over the $2^{n-m-1} - 1$ alias sets that do not contain main effects, i.e., each of these alias sets contains the same number of two-factor interactions. This design has minimum aberration, estimation index 2, and also has maximum estimation capacity over all designs.

For $n < N/2$, in most cases, the minimum aberration designs have estimation indices greater than 2. For example, we noted earlier that under the minimum aberration 2^{11-6} design, 15 of the $m_i(d)$'s are nonzero, and 5 are zero. Such a design can be used to entertain only up to 15 two-factor interactions. As a result, $E_k(d) = 0$ for all $k > 15$. On the other hand, there exist 2^{11-6} resolution III designs with estimation index 2. Such designs have positive $E_k(d)$'s for $16 \leq k \leq 20$, and due to the continuity of $E_k(d)$ as a function of k , they also have larger $E_k(d)$'s than the minimum aberration design for the k 's that are not

much smaller than 16. Although the nonzero $m_i(d)$ s for a minimum aberration design are nearly equal, the constraint that some of them must be zero prevents the design from maximizing $E_k(d)$ for larger k 's.

For 32 runs, by the discussion in the paragraph preceding the previous one, the maximal design with $n = 16$ has minimum aberration and maximum estimation capacity. We note that the maximal design with $n = 10$ also has minimum aberration. It can be shown that it maximizes $E_k(d)$ for all k as well. The fact that it has estimation index 2 plays an important role for it to have maximum estimation capacity. The readers are referred to Chen and Cheng (2004) for details. The third maximal design, with $n = 9$, however, does not have minimum aberration. In this case, the minimum aberration design can be obtained by deleting one factor from the maximal design with $n = 10$, and therefore has estimation index 3.

In general, if a minimum aberration design of resolution III+ has estimation index 2, then it is expected to have large, if not maximum, $E_k(d)$ for all k s. On the other hand, if a minimum aberration design has estimation index 3, but another design has estimation index 2, then the minimum aberration design tends to be optimal for smaller k s, but not for larger k s. In this case, if the number of active two-factor interactions is expected to be large, then one may want to use a design that has minimum aberration among those with estimation index 2.

9.7 COMPLEMENTARY DESIGN THEORY FOR MINIMUM ABERRATION DESIGNS

Construction of minimum aberration designs has been studied by many authors. In this section, we review a useful technique of constructing minimum aberration designs via complementary designs.

Let d be a regular 2^{n-m} design of resolution III+. Then d can be constructed by deleting $f = 2^{n-m} - 1 - n$ factors from the $2^{n-m} - 1$ factors of a saturated design. The deleted factors form another regular design, denoted by \bar{d} . We call it the *complementary design* of d .

Tang and Wu (1996) derived identities that relate the wordlength pattern of d to that of \bar{d} . Chen and Hedayat (1996) independently derived such identities for defining words of lengths 3 and 4:

$$A_k(d) = C_0 + \sum_{j=3}^{k-1} C_j A_j(\bar{d}) + (-1)^k A_k(\bar{d}), \quad 3 \leq k \leq n, \quad (9.5)$$

where the C_j 's are constants not depending on d . It follows from Equation (9.5) that sequentially minimizing $A_k(d)$ is equivalent to sequentially minimizing $(-1)^k A_k(\bar{d})$, that is, maximizing $A_3(\bar{d})$, followed by minimizing $A_4(\bar{d})$, and then maximizing $A_5(\bar{d})$, etc. Therefore, the determination of a minimum

aberration design can be done via the selection of its complementary design. This is particularly useful when d is nearly saturated, in which case \bar{d} has only a few factors, and it is much easier to determine the wordlength pattern of \bar{d} than that of d . This result was extended to regular s^{n-m} designs, where s is a prime power, by Suen, Chen, and Wu (1997). Chen and Hedayat (1996) showed that the complementary designs of minimum aberration designs in the saturated designs have the same structure as long as they have the same number of factors, and determined all minimum aberration 2^{n-m} designs with $f \leq 16$.

For $n = N/2$, we already know that the maximal even designs have minimum aberration, and that they can be constructed by applying the method of foldover to saturated designs of size $N/2$ or by repeatedly doubling the complete 2^2 factorial.

When $n < N/2$, the complementary design has more factors than the original design. Therefore, the method of complementary designs described above is not useful. In this case, the minimum aberration designs must have resolution IV+, and can be constructed by deleting factors from certain maximal designs of resolution IV+, instead of the saturated designs. Can a similar complementary design theory be developed, with maximal resolution IV+ designs as the universes?

By Theorem 9.8, there is no maximal design with $5N/16 < n < N/2$. The following is an important consequence of this observation.

Theorem 9.9. All the regular resolution IV designs with $5N/16 < n < N/2$ are projections of the maximal even design of size N .

It follows from Theorem 9.9 that all regular resolution IV designs with $5N/16 < n < N/2$ are foldover and even designs. For example, to construct a minimum aberration 32-run design with 14 factors, we only have to determine which two factors to drop from the 16 factors of the 32-run maximal even design. This substantially simplifies the problem. Such a complementary design theory was provided by Butler (2003a); see also Chen and Cheng (2009). The following theorem is due to Butler (2003a).

Theorem 9.10. For $5N/16 < n < N/2$, a regular 2^{n-m} design d has minimum aberration if and only if it is an n -dimensional projection of the maximal even design of size N , and the design obtained by deleting the factors of d from the maximal even design has minimum aberration among all $(N/2 - n)$ -dimensional projections of the maximal even design.

Chen and Cheng (2009) obtained explicit identities that relate the wordlength pattern of an even design to that of its complement in the maximal even design. These identities were used to further investigate the structures of (weak) minimum aberration designs of resolution IV.

For $9N/32 < n \leq 5N/16$, a minimum aberration 2^{n-m} design is either a projection of the maximal even design or the maximal design with $5N/16$ factors.

We need to compare the best projections of these two maximal designs with respect to the minimum aberration criterion. Chen and Cheng (2006) showed that the best projection of the maximal even design has more aberration than the best projection of the maximal design with $5N/16$ factors. It follows that minimum aberration designs with $9N/32 < n \leq 5N/16$ are n -dimensional projections of the maximal design with $5N/16$ factors; therefore, they are not even (foldover) designs. In particular, the maximal design with $5N/16$ factors is a minimum aberration design. They also showed that the maximal design with $n = 9N/32$ does not have minimum aberration. Instead, the minimum aberration design for $n = 9N/32$ is also a projection of the maximal design with $5N/16$ factors.

By Theorem 9.8, for $17N/64 < n < 9N/32$, minimum aberration designs can be projections of maximal designs with $9N/32$, $5N/16$, or $N/2$ factors. Again, one needs to compare the best n -dimensional projections of these three maximal designs. In general, for $n \geq N/4 + 1$, to determine a minimum aberration 2^{n-m} design, one needs to compare the best n -dimensional projections of the maximal designs with at least n factors. A complementary design theory for the maximal designs would be useful for determining their best projections. By Theorem 9.8, all the maximal designs with $N/4 + 1 \leq n \leq N/2$ can be obtained by repeatedly doubling smaller maximal designs. Xu and Cheng (2008) developed a general complementary design theory for repeated doubles. It can be applied to saturated designs as well, since the saturated designs can also be constructed by the method of doubling; see Remark 9.1. By applying the general result to saturated designs and maximal even designs, one obtains the results of Chen and Hedayat (1996), Tang and Wu (1996) and Butler (2003a) as special cases. As an application, Xu and Cheng (2008) obtained the following improvement of Chen and Cheng's (2006) result mentioned in the previous paragraph.

Theorem 9.11. For $N = 32 \cdot 2^t$, $t \geq 0$, and $17N/64 \leq n \leq 5N/16$, a minimum aberration design with N runs and n factors must be a projection of the maximal regular design of resolution IV with $5N/16$ factors.

As a numerical illustration, for 256 runs, Xu and Cheng (2008) determined the best n -factor projections of the maximal design with 80 factors for $69 \leq n \leq 79$. This corresponds to complementary designs with 1–11 factors. Block (2003) considered the construction of minimum aberration designs from this maximal design by deleting one column at a time. The theoretical result in Xu and Cheng (2008) confirmed that the designs with $69 \leq n \leq 79$ obtained by Block (2003) indeed have minimum aberration except for $n = 71$.

9.8 NONREGULAR DESIGNS AND ORTHOGONAL ARRAYS

Two-level regular designs are easy to construct and analyze, but the run sizes must be powers of two. On the other hand, nonregular designs are more

flexible in terms of run sizes. To pave the way for extending the minimum aberration criterion to nonregular designs in the next section, we first review some basic properties of orthogonal arrays. We also briefly discuss extension of the concepts of estimation capacity and clear two-factor interactions to nonregular designs.

Suppose the two levels of each factor are denoted by 1 and -1 . Then, as in the paragraph preceding Theorem 9.8 and elsewhere in this chapter, each N -run design d for n two-level factors can be represented by an $N \times n$ matrix $\mathbf{X}(d)$, where each column corresponds to one factor and each row represents a factor-level combination. Let the i th column of $\mathbf{X}(d)$ be $\mathbf{x}_i(d)$. Then, $\mathbf{x}_i(d)$ is the column of the model matrix that corresponds to the main effect of the i th factor. For each $1 \leq k \leq n$ and any subset $S = \{i_1, \dots, i_k\}$ of $\{1, \dots, n\}$, let $\mathbf{x}_S(d) = \mathbf{x}_{i_1}(d) \odot \dots \odot \mathbf{x}_{i_k}(d)$, where for any two vectors \mathbf{x} and \mathbf{y} , $\mathbf{x} \odot \mathbf{y}$ is the componentwise product of \mathbf{x} and \mathbf{y} . Then $\mathbf{x}_S(d)$ is the column of the model matrix that corresponds to the interaction of factors i_1, \dots, i_k .

Tang (2001) defined $J_S(d)$, called a J -characteristic, as the sum of all the entries of $\mathbf{x}_S(d)$:

$$J_S(d) = \sum_{l=1}^N x_{li_1}(d) \cdots x_{li_k}(d),$$

where $x_{li}(d)$ is the (l, i) th entry of $\mathbf{X}(d)$, that is, the level of the i th factor at the l th run.

Definition 9.1. An orthogonal array $\text{OA}(N, s^n, t)$ is an $N \times n$ matrix with s distinct symbols in the i th column, $1 \leq i \leq n$, such that in each $N - t$ submatrix, all combinations of the symbols appear equally often as row vectors. The positive integer t is called the strength of the orthogonal array.

For any subset S of $\{1, \dots, n\}$, let $|S|$ be the cardinality (size) of S . If $\mathbf{X}(d)$ is an orthogonal array of strength t , then it is easy to see that for any S such that $|S| \leq t$, $\mathbf{x}_S(d)$ contains the same number of 1s and -1 s. It follows that $J_S(d) = 0$ for all S such that $|S| \leq t$. It can be shown that the converse is also true:

Theorem 9.12. A fractional factorial design d is an orthogonal array of strength t if and only if $J_S(d) = 0$ for all S such that $|S| \leq t$.

Under a regular design, if the k -factor interaction involving factors i_1, \dots, i_k appears in the defining relation, then $\sum_{l=1}^N x_{li_1}(d) \cdots x_{li_k}(d)$ is equal to either N or $-N$; otherwise, $J_S(d) = 0$; in particular $J_S(d) = 0$ for all S such that $|S| < r$. This and Theorem 9.12 together imply the following result.

Theorem 9.13. Each regular fractional factorial design of resolution r is an orthogonal array of strength $r - 1$.

For any two subsets S and T of the factors, we say that their corresponding factorial effects are orthogonal if $\mathbf{x}_S(d)^T \mathbf{x}_T(d) = 0$. We have

$$\mathbf{x}_S(d)^T \mathbf{x}_T(d) = J_{S\Delta T}(d),$$

where $S\Delta T = (S \cup T) \setminus (S \cap T)$. For regular designs, $J_{S\Delta T}(d)/N$ is equal to 0, 1 or -1 ; that is, any two factorial effects are either orthogonal or completely aliased. For nonregular designs, however, $J_{S\Delta T}(d)/N$ can have an absolute value strictly between 0 and 1, indicating partial aliasing of factorial effects. It follows from Theorem 9.12 that if $\mathbf{X}(d)$ is an orthogonal array of strength t , then for any two subsets S and T of factors, the corresponding factorial effects are orthogonal if $|S\Delta T| \leq t$; in particular, this is true if $|S| + |T| \leq t$. Thus, for example, under an orthogonal array of strength 2, all the main effects are mutually orthogonal, and under an orthogonal array of strength 3, not only the main effects are mutually orthogonal, they are also orthogonal to two-factor interactions.

The Plackett–Burman designs discussed in HK2, section 14.2.4, are orthogonal arrays of strength 2. Such designs are constructed from Hadamard matrices. A Hadamard matrix of order N is an $N \times N$ matrix \mathbf{H} with all the entries equal to 1 or -1 such that $\mathbf{H}^T \mathbf{H} = N\mathbf{I}$, where \mathbf{I} is the identity matrix. Without loss of generality, we may assume that all the entries in the first column of \mathbf{H} are equal to 1. Then, the last $N - 1$ columns of \mathbf{H} form an orthogonal array of strength 2.

As in the case of regular designs, the foldover of a two-level orthogonal array of even strength t is an orthogonal array of strength $t + 1$. In particular, the foldover of \mathbf{H} , with the first column of 1s included, is a strength 3 orthogonal array with N factors and $2N$ runs.

Tang (2006) generalized the concept of clear two-factor interactions to orthogonal arrays. A two-factor interaction is called clear if it is orthogonal to all the main effects and other two-factor interactions. In addition to construction results for orthogonal arrays with clear two-factor interactions, the following generalizations of Theorems 9.2 and 9.3 were established:

Theorem 9.14. A necessary condition for an orthogonal array of strength 2 to have a clear two-factor interaction is that (1) N is a multiple of 8 and (2) $n \leq N/2$.

Theorem 9.15. A necessary condition for an orthogonal array of strength 3 to have a clear two-factor interaction is that (1) N is a multiple of 16 and (2) $n \leq N/4 + 1$.

We now turn to estimation capacity. Recall that estimation capacity was defined by counting the number of estimable models in a class of candidate models. For nonregular designs, it is not enough to count the number of estimable models since different estimable models may have different efficiencies. Sun (1993) proposed the use of average efficiency (based on the D -criterion)

over the candidate models, called information capacity. This idea was adopted by Li and Nachtsheim (2000) in their search of model-robust designs. The use of average efficiency over a set of candidate models to select factorial designs under model uncertainty was also proposed by Tsai, Gilmour, and Mead (2000) in their investigation of three-level designs. While Sun (1993) and Li and Nachtsheim (2000) considered the D-criterion and models that contain all the main effects and a certain number of two-factor interactions, Tsai, Gilmour, and Mead (2000) used an approximation to the A_s -criterion and examined average performances of designs over lower-dimensional projections. They developed a Q -criterion, and extended it to a more general Q_B -criterion in Tsai, Gilmour, and Mead (2007) for incorporating prior information about model uncertainty. Applications and comparisons of the Q_B -criterion with other criteria in many different situations, including mixed-level quantitative/qualitative factors, were explored in Tsai and Gilmour (2010).

9.9 GENERALIZED MINIMUM ABERRATION

Deng and Tang (1999) extended resolution and minimum aberration to generalized resolution and generalized minimum aberration for comparing and assessing nonregular designs. A variant of generalized minimum aberration, called *minimum G_2 -aberration*, was proposed in Tang and Deng (1999). In this section, we shall concentrate on minimum G_2 -aberration.

Let

$$B_k(d) = N^{-2} \sum_{S:|S|=k} [J_S(d)]^2.$$

Then the minimum G_2 -aberration criterion proposed by Tang and Deng (1999) is to sequentially minimize $B_1(d), B_2(d), \dots, B_n(d)$. We shall call $(B_1(d), B_2(d), \dots, B_n(d))$ the *generalized wordlength pattern*.

We have pointed out in the previous section that if the interaction of the factors in S appears in the defining relation of a regular design, then $J_S(d)$ is equal to either N or $-N$; otherwise, $J_S(d) = 0$. Therefore $[J_S(d)]^2/N^2$ is equal to 1 or 0, depending on whether the interaction of the factors in S appears in the defining relation. It follows that for regular designs, $B_k(d) = A_k(d)$, the number of defining words of length k . Therefore, the minimum G_2 -aberration criterion reduces to minimum aberration when they are applied to regular designs. In view of the discussion in the previous section, the quantities $B_k(d)$ can be viewed as overall measures of aliasing among the factorial effects.

For regular designs, the resolution is at least r if and only if $A_k(d) = 0$ for all $k < r$. This extends to the following result for the general case.

Theorem 9.16. A fractional factorial design d is an orthogonal array of strength t if and only if $B_k(d) = 0$ for all $k \leq t$.

This is just a rephrase of Theorem 9.12, since $B_k(d) = 0$ if and only if $J_S(d) = 0$ for all S such that $|S| = k$.

Tang and Deng (1999) also justified the minimum G_2 -aberration criterion from the viewpoint of minimizing the contamination of nonnegligible interactions on the estimation of main effects in the order of importance given by the hierarchical assumption. This approach was used by Cheng and Tang (2005) to develop a general theory of minimum aberration for regular designs with a given set of effects whose estimates are required and a specified hierarchical ordering of the other effects.

It was shown in Section 9.3 that minimum aberration is a good surrogate for maximum estimation capacity. Cheng, Deng, and Tang (2002) showed that minimum G_2 -aberration is a good surrogate for maximum information capacity under the same class of candidate models considered in Cheng, Steinberg, and Sun (1999). Tsai and Gilmour (2010) demonstrated that the Q_B -criterion is closely related to generalized minimum aberration.

Using connections with coding theory, Xu and Wu (2001) extended minimum G_2 -aberration to a generalized minimum aberration criterion for mixed-level designs. Ma and Fang (2001) independently extended it to q -level designs with $q > 2$. Cheng and Ye (2004) defined a generalized minimum aberration criterion for quantitative factors.

Some theoretical results for the determination of minimum G_2 -aberration designs were derived in Butler (2003b) for two-level designs with $n \geq N/2 - 2$, Butler (2004) for two-level designs with $n \leq N/2$, and Butler (2005) for the case of more than two levels. Minimum G_2 -aberration designs for many n s were obtained for $N = 16, 24, 32, 48, 64$, and 96. For example, it was shown in Butler (2004) that all the 24-run two-level minimum G_2 -aberration designs with $n \leq 12$ are foldovers of projections of a 12×12 Hadamard matrix. One useful tool to show this is an analogue of Theorem 9.9 for nonregular designs. By Theorem 9.9, all regular resolution IV designs with $5N/16 < n \leq N/2$ are foldover designs. Butler (2007) extended this to the nonregular case and showed that foldover designs are the only strength-three orthogonal arrays with $N/3 \leq n \leq N/2$. A key technical result in Butler (2003b, 2004, 2005), as well as Butler (2003a) cited in Section 9.7, relates the generalized wordlength pattern to the moments of the Hadamard distances between the runs (rows). The former measures aliasing of factorial effects, and the latter measures similarity between the runs. This idea was also used by Xu (2003) to formulate a row-based minimum moment aberration criterion, which was shown to be equivalent to minimum G_2 -aberration for two-level designs and generalized minimum aberration for the more general case considered by Xu and Wu (2001). This equivalence, in addition to theoretical implications, provides a powerful tool for algorithmic construction of generalized minimum aberration designs (Xu 2002).

We refer the readers to Xu, Phoa, and Wong (2009) for additional references and an extensive review of recent development in generalized minimum aberration designs.

9.10 OPTIMAL FRACTIONAL FACTORIAL BLOCK DESIGNS

Blocking of fractional factorial designs was presented in HK2, section 13.8, and optimal blocking based on a criterion proposed by Chen and Cheng (1999) was briefly discussed in HK2, section 13.8.3. In this section, we provide more details, and extend it to the case where the block effects are random.

To construct a 2^{n-m} regular design in 2^l blocks of size 2^{n-m-l} , we need $m+l$ independent defining effects: m of these effects are used to define the fraction, and the other l effects are used to divide the 2^{n-m} treatment combinations into 2^l blocks.

Example 9.8. For the case $n = 6$ and $m = l = 2$, consider the resolution IV design in Example 9.1 defined by the two defining words $ABCE$ and $BCDF$. Suppose we use ACD and ABD to do blocking. The first six columns of the following matrix give the levels of factors A, B, C, D, E , and F in the 16 treatment combinations defined by $I = ABCE = BCDF (= ADEF)$. The seventh column is the componentwise product of the columns corresponding to A, C , and D , and the last column is the componentwise product of the columns corresponding to A, B , and D . Consider the entries in these two columns as the levels of two additional factors denoted by b_1 and b_2 , respectively. Then we write $b_1 = ACD$ and $b_2 = ABD$. The 16 treatment combinations are divided into four blocks of size 4 according to the values of b_1 and b_2 . For example, those with $b_1 = b_2 = -1$ are in the same block, and those with $b_1 = 1$ and $b_2 = -1$ are in one block, etc.

The design in Example 9.8 can be viewed as a 2^{8-4} fractional factorial design with independent defining words $ABCE$, $BCDF$, b_1ACD , and b_2ABD for the eight factors A, B, C, D, E, F, b_1 , and b_2 , where A, B, C, D, E, F are treatment factors, and b_1, b_2 are called *block factors*. In general, a regular 2^{n-m} design in 2^l blocks can be viewed as a $2^{(n+l)-(m+l)}$ design with n treatment factors and l block factors. Then there are two types of defining words: those which do not contain block factors and those which contain block factors. The former are called *treatment defining words* and the latter are called *block defining words*.

When viewed as a 2^{8-4} design, the design in Example 9.8 has defining relation

$$\begin{aligned} I &= ABCE = BCDF = ACDb_1 = ABD_b_2 = ADEF = BDEb_1 = CEFb_1 = ABFb_1 \\ &= ACFb_2 = BEFb_2 = CDEb_2 = BCb_1b_2 = AEb_1b_2 = DFb_1b_2 = ABCDEFb_1b_2 \end{aligned}$$

The 15 alias sets are as given in Example 9.1. The three underlined alias sets are those that are confounded with blocks. We can also determine the effects that are confounded with blocks from the 12 block defining words: the treatment letters in each block defining word represent a treatment interaction that is confounded with blocks. The two types of defining words play quite different roles. The treatment defining words can be used to determine aliasing among treatment factorial effects, while the block-defining words provide information about which treatment factorial effects are confounded with blocks. A naive minimum aberration criterion based on the lengths of all the defining words is not appropriate, and it is necessary to treat the two types of defining words differently. Let $A_{i,0}(d)$ be the number of treatment defining words of length i and $A_{i,1}(d)$ be the number of block defining words that contain i treatment letters. Several different wordlength patterns based on these wordlengths have been proposed in the literature for defining minimum aberration blocked fractional factorial designs (Sitter, Chen, and Feder 1997; Chen and Cheng 1999; Cheng and Wu 2002). Xu (2006) and Xu and Lau (2006) constructed all minimum-aberration blocked two-level fractional factorial designs with 8, 16, and 32 runs and those with 64 runs and up to 32 factors with respect to these wordlength patterns, and found that in most cases, they lead to the same minimum aberration design. We pay special attention to the wordlength patterns W_{CC} proposed by Chen and Cheng (1999) and W_1 proposed by Cheng and Wu (2002):

$$W_{CC} = (3A_{3,0} + A_{2,1}, A_{4,0}, 10A_{5,0} + A_{3,1}, A_{6,0}, \dots),$$

$$W_1 = (A_{3,0}, A_{4,0}, A_{2,1}, A_{5,0}, A_{6,0}, A_{3,1} \dots).$$

The wordlength pattern W_{CC} was based on the consideration of estimation capacity. Under models with fixed block effects, the factorial effects that are confounded with blocks cannot be estimated. Therefore, we only consider the designs under which none of the treatment main effects is aliased with other main effects or confounded with blocks. Again, we define $E_k(d)$ as the number

of models containing all the main effects and k two-factor interactions that can be estimated by a design d . Then Equation (9.3) holds with f replaced by the number of alias sets that neither contain main effects nor are confounded with blocks, say h . Thus a design has large estimation capacity if (1) $\sum_{i=1}^h m_i(d)$ (= the number of two-factor interactions that are neither aliased with main effects nor confounded with blocks) $= \binom{n}{2} - 3A_{3,0}(d) - A_{2,1}(d)$ is maximized, and (2) the $m_i(d)$ s are as equal as possible. Similar to the discussions in Section 9.3, in this case, maximizing $\sum_{i=1}^h m_i(d)$ is equivalent to minimizing $3A_{3,0}(d) + A_{2,1}(d)$, and a good surrogate for (2) is to minimize $A_{4,0}(d)$. This leads to the first two terms of W_{CC} . The rest can be derived by considering higher-order effects. Therefore, as in Section 9.3, minimum aberration with respect to W_{CC} is a good surrogate for maximum estimation capacity in the present context.

Example 9.8 (Revisited). The design in Example 9.8 has minimum aberration with respect to W_{CC} . It has $A_{3,0}(d) = 0$ and $A_{2,1}(d) = 3$. Under this design, no two-factor interaction is aliased with main effects, and three are confounded with blocks. It maximizes the number of two-factor interactions that are neither aliased with main effects nor confounded with blocks. Each of the six alias sets that do not contain main effects and are not confounded with blocks contains two two-factor interactions, a uniform distribution. Therefore the design maximizes $E_k(d)$ for all k .

Example 9.9. Consider the case $n = 13$, $m = 8$, and $l = 3$, that is, 32-run designs for 13 factors with the 32 runs grouped into eight blocks of size 4 each. Let d_1 and d_2 be two designs with the following sets of independent defining words:

$$d_1: \{1236, 1247, 1348, 2349, 125t_{10}, 135t_{11}, 235t_{12}, 145t_{13}, 13b_1, 14b_2, 15b_3\},$$

$$d_2: \{126, 137, 148, 2349, 1234t_{10}, 235t_{11}, 245t_{12}, 345t_{13}, 23b_1, 24b_2, 15b_3\},$$

where t_{10}, \dots, t_{13} are factors 10, ..., 13. Then d_1 is a resolution IV design with $A_{3,0}(d_1) = 0$, $A_{4,0}(d_1) = 55$, and $A_{2,1}(d_1) = 36$, and d_2 is a resolution III design with $A_{3,0}(d_2) = 4$, $A_{4,0}(d_2) = 39$, and $A_{2,1}(d_2) = 22$. Design d_1 has minimum aberration with respect to W_1 , and d_2 has minimum aberration with respect to W_{CC} :

$$W_{CC}(d_1) = (36, 55, \dots),$$

$$W_{CC}(d_2) = (34, 39, \dots).$$

Under d_1 , there are $\binom{13}{2} - 36 = 42$ two-factor interactions that are neither aliased with main effects nor confounded with blocks. These 42 interactions are partitioned into eight alias sets, two of which are of size 6 and six are of size 5. On the other hand, under d_2 , there are 44 two-factor interactions that are neither aliased with main effects nor confounded with blocks, and they are partitioned into 11 alias sets each of size 4. Not only does d_2 produce more

two-factor interactions that are neither aliased with main effects nor confounded with blocks, but also they are distributed among the alias sets in the most uniform fashion. It follows that d_2 has maximum estimation capacity over all the 2^{13-8} designs in eight blocks. Later we shall compare d_1 and d_2 again under models with random block effects.

Under models with random block effects, there is information for the treatment interactions that are confounded with blocks, but they are estimated with larger variances than those that are not confounded with blocks. The variance of the estimator of an interaction that is confounded with blocks depends on the interblock variance ξ_1 , and that of the estimator of an interaction orthogonal to the blocks depends on the intrablock variance ξ_2 . Typically, $\xi_2 < \xi_1$. Cheng and Tsai (2009) investigated how to incorporate ξ_1 and ξ_2 into a criterion for design selection. In the spirit of Cheng, Steinberg, and Sun (1999), the objective is to find good designs under model uncertainty, which is the rationale behind minimum aberration. However, due to different precisions for effects estimated in the inter- and intrablock strata, again counting the number of estimable models among the candidate models is not enough, since different estimable models may be estimated with different precisions, as in the case of nonregular designs discussed in Section 9.8.

The approach of information capacity was adopted in Cheng and Tsai (2009). Since it is assumed that none of the main effects are confounded with blocks, we can concentrate on the efficiencies of the estimators of two-factor interactions, measured by $|D|^{1/k}$, where $|D|$ is the determinant of the information matrix for the two-factor interactions, and k is the number of two-factor interactions in the model. For a given design d , the information capacity $I_k(d)$ is defined as the average of its efficiencies under this determinant-based criterion over the models containing all the main effects and k two-factor interactions.

Among the f alias sets that are not aliased with main effects, $2^l - 1$ are confounded with blocks. Without loss of generality, suppose that the effects in each of the first $2^l - 1$ alias sets are confounded with blocks, those in the next $f - (2^l - 1)$ alias sets are neither confounded with blocks nor aliased with main effects, and each of the last n alias sets contains one main effect. Cheng and Tsai (2009) showed that a good surrogate for maximizing $I_k(d)$ is to maximize

$$\xi_1^{-1/k} \sum_{i=1}^{2^l-1} m_i(d) + \xi_2^{-1/k} \sum_{i=2^l}^f m_i(d),$$

and minimize

$$\xi_1^{-2/k} \sum_{i=1}^{2^l-1} [m_i(d)]^2 + \xi_2^{-2/k} \sum_{i=2^l}^f [m_i(d)]^2,$$

among those that maximize

$$\xi_1^{-1/k} \sum_{i=1}^{2^l-1} m_i(d) + \xi_2^{-1/k} \sum_{i=2^l}^f m_i(d).$$

They further showed that a sufficient condition for a design to be optimal with respect to this surrogate for all ξ_1 and ξ_2 such that $\xi_2 < \xi_1$ is that it (1) maximizes $\sum_{i=1}^f m_i(d)$ and minimizes $\sum_{i=1}^f [m_i(d)]^2$ among those maximizing $\sum_{i=1}^f m_i(d)$, and (2) maximizes $\sum_{i=2^l}^f m_i(d)$, and minimizes $\sum_{i=2^l}^f [m_i(d)]^2$ among those maximizing $\sum_{i=2^l}^f m_i(d)$. This provides a way of checking the strong optimality of a design without having to know ξ_1 and ξ_2 . When no design can achieve both (1) and (2), this result can be used to eliminate inferior designs. If a design d_1 is at least as good as another design d_2 with respect to both (1) and (2), and is better than d_2 under either (1) or (2), then d_2 is an *inadmissible* design that can be eliminated. Then we can concentrate on admissible designs and compare them based on knowledge about ξ_1 and ξ_2 or other information. Note that by the discussion in Section 9.3, (1) assures that d is a good unblocked design, and (2) assures that it is a good design when only intrablock information is used.

Example 9.9 (Revisited). For $n = 13$, $m = 8$ and $l = 3$, d_1 and d_2 are the only two admissible designs. Neither dominates the other, but they together dominate all the other designs. The W_{CC} minimum aberration design is better than the W_1 minimum aberration design if and only if the intrablock variance ξ_2 is sufficiently smaller than the interblock variance ξ_1 .

The approach described above can be applied to split-plot designs, with whole-plots playing the same role as blocks. One important difference is that in a split-plot experiment, some main effects are confounded with whole-plots. The readers are referred to Cheng and Tsai (2009) for detailed discussions, comparisons with the results in Bingham and Sitter (2001), and additional references. Under a block design with random block effects or a split-plot design, there are two error terms, or we say two error strata. Cheng and Tsai (2011) investigated multistratum fractional factorial designs for experiments with multiple error terms, such as those studied in Miller (1997), Mee and Bates (1998), McLeod and Brewster (2004), Bingham et al. (2008), and Vivacqua and Bisgaard (2009).

REFERENCES

- Bingham, D.R. and R.R. Sitter (2001). Design issues in fractional factorial split-plot experiments. *J. Qual. Technol.*, **33**, 2–15.
 Bingham, D.R., R.R. Sitter, E. Kelly, L. Moore, and J.D. Olivas (2008). Factorial designs with multiple levels of randomization. *Stat. Sin.*, **18**, 493–513.

- Block, R.M. (2003). Theory and construction methods for large regular resolution IV designs. PhD dissertation, University of Tennessee, Knoxville.
- Block, R.M. and R.W. Mee (2003). Second order saturated resolution IV designs. *J. Stat. Theory Appl.*, **2**, 96–112.
- Box, G.E.P. and J.S. Hunter (1961). The 2^{k-p} fractional factorial designs I. *Technometrics*, **3**, 311–351.
- Bruen, A., L. Haddad, and L. Wehlau (1998). Binary codes and caps. *J. Comb. Des.*, **6**, 275–284.
- Bruen, A. and D. Wehlau (1999). Long binary linear codes and large caps in projective space. *Des. Codes Cryptogr.*, **17**, 37–60.
- Butler, N.A. (2003a). Some theory for constructing minimum aberration fractional factorial designs. *Biometrika*, **90**, 233–238.
- Butler, N.A. (2003b). Minimum aberration construction results for nonregular two-level fractional factorial designs. *Biometrika*, **90**, 891–898.
- Butler, N.A. (2004). Minimum G_2 -aberration properties of two-level foldover designs. *Stat. Probab. Lett.*, **67**, 121–132.
- Butler, N.A. (2005). Generalised minimum aberration construction results for symmetrical orthogonal arrays. *Biometrika*, **92**, 485–491.
- Butler, N.A. (2007). Results for two-level fractional factorial designs of resolution IV or more. *J. Stat. Plann. Inference*, **137**, 317–323.
- Chen, H. (1998). Some projective properties of fractional factorial designs. *Stat. Probab. Lett.*, **40**, 185–188.
- Chen, H. and C.-S. Cheng (1999). Theory of optimal blocking of 2^{n-m} designs. *Ann. Stat.*, **27**, 1948–1973.
- Chen, H.H. and C.-S. Cheng (2004). Aberration, estimation capacity and estimation index. *Stat. Sin.*, **14**, 203–215.
- Chen, H.H. and C.-S. Cheng (2006). Doubling and projection: A method of constructing two-level designs of resolution IV. *Ann. Stat.*, **34**, 546–558.
- Chen, H.H. and C.-S. Cheng (2009). Some results on 2^{n-m} designs of resolution IV with (weak) minimum aberration. *Ann. Stat.*, **37**, 3600–3615.
- Chen, H. and A.S. Hedayat (1996). 2^{n-l} designs with weak minimum aberration. *Ann. Stat.*, **24**, 2536–2548.
- Chen, H. and A.S. Hedayat (1998). 2^{n-m} designs with resolution III or IV containing clear two-factor interactions. *J. Stat. Plann. Inference*, **75**, 147–158.
- Chen, J., D.X. Sun, and C.F.J. Wu (1993). A catalogue of two-level and three-level fractional factorial designs with small runs. *Internat. Stat. Rev.*, **61**, 131–145.
- Cheng, C.-S. and R. Mukerjee (1998). Regular fractional factorial designs with minimum aberration and maximum estimation capacity. *Ann. Stat.*, **26**, 2289–2300.
- Cheng, C.S. and B. Tang (2005). A general theory of minimum aberration and its applications. *Ann. Stat.*, **33**, 944–958.
- Cheng, C.-S. and P.-W. Tsai (2009). Optimal two-level regular fractional factorial block and split-plot designs. *Biometrika*, **96**, 83–93.
- Cheng, C.-S. and P.-W. Tsai (2011). Multistratum fractional factorial designs. *Stat. Sin.*, **21**, 1001–1021.

- Cheng, C.-S., D.M. Steinberg, and D.X. Sun (1999). Minimum aberration and maximum estimation capacity. *J. R. Stat. Soc. B*, **61**, 85–94.
- Cheng, C.S., L.Y. Deng, and B. Tang (2002). Generalized minimum aberration and design efficiency for nonregular fractional factorial designs. *Stat. Sin.*, **12**, 991–1000.
- Cheng, S.W. and C.F.J. Wu (2002). Choice of optimal blocking schemes in two-level and three-level designs. *Technometrics*, **44**, 269–277.
- Cheng, S.W. and K.Q. Ye (2004). Geometric isomorphism and minimum aberration for factorial designs with quantitative factors. *Ann. Stat.*, **32**, 2168–2185.
- Davydov, A.A. and L.M. Tombak (1990). Quasiperfect linear binary codes with distance 4 and complete caps in projective geometry. *Probl. Inform. Transm.*, **25**, 265–275.
- Deng, L.Y. and B. Tang (1999). Generalized resolution and minimum aberration criteria for Plackett-Burman and other nonregular factorial designs. *Stat. Sin.*, **9**, 1071–1082.
- Fries, A. and W.G. Hunter (1980). Minimum aberration 2^{k-p} designs. *Technometrics*, **22**, 601–608.
- Li, W. and C.J. Nachtsheim (2000). Model-robust factorial designs. *Technometrics*, **42**, 245–252.
- Ma, C.X. and K.T. Fang (2001). A note on generalized aberration in factorial designs. *Metrika*, **53**, 85–93.
- Marshall, A.W. and I. Olkin (1979). *Inequalities: Theory of Majorization and Its Applications*. New York: Academic Press.
- McLeod, R.G. and J.F. Brewster (2004). The design of blocked fractional factorial split-plot experiments. *Technometrics*, **46**, 135–146.
- Mee, R.W. and R.L. Bates (1998). Split-lot designs: Experiments for multi-stage batch processes. *Technometrics*, **40**, 127–140.
- Miller, A. (1997). Strip-plot configurations of fractional factorials. *Technometrics*, **39**, 153–161.
- Sitter, R.R., J. Chen, and M. Feder (1997). Fractional resolution and minimum aberration in blocking factorial designs. *Technometrics*, **39**, 382–390.
- Suen, C.-Y., H. Chen, and C.F.J. Wu (1997). Some identities on q^{n-m} designs with application to minimum aberrations. *Ann. Stat.*, **25**, 1176–1188.
- Sun, D.X. (1993). Estimation capacity and related topics in experimental designs. PhD dissertation, University of Waterloo.
- Tang, B. (2001). Theory of J-characteristics for fractional factorial designs and projection justification of minimum G_2 -aberration. *Biometrika*, **88**, 401–407.
- Tang, B. (2006). Orthogonal arrays robust to nonnegligible two-factor interactions. *Biometrika*, **93**, 137–146.
- Tang, B. and L.Y. Deng (1999). Minimum G_2 -aberration for Nonregular Fractional Factorial designs. *Ann. Stat.*, **27**, 1914–1926.
- Tang, B. and C.F.J. Wu (1996). Characterization of minimum aberration 2^{n-m} designs in terms of their complementary designs. *Ann. Stat.*, **24**, 2549–2559.
- Tang, B., F.S. Ma, D. Ingram, and H. Wang (2002). Bounds on the maximum numbers of clear two-factor interactions for 2^{n-m} designs of resolution III and IV. *Can. J. Stat.*, **30**, 127–136.

- Tsai, P.-W. and S.G. Gilmour (2010). A general criterion for factorial designs under model uncertainty. *Technometrics*, **52**, 231–242.
- Tsai, P.-W., S.G. Gilmour, and R. Mead (2000). Projective three-level main effects designs robust to model uncertainty. *Biometrika*, **87**, 467–475.
- Tsai, P.W., S.G. Gilmour, and R. Mead (2007). Three-level main-effects designs exploiting prior information about model uncertainty. *J. Stat. Plann. Inference*, **137**, 619–627.
- Vivacqua, C. and S. Bisgaard (2009). Post-fractionated strip-block designs. *Technometrics*, **51**, 47–55.
- Wu, C.F.J. and Y. Chen (1992). A graph-aided method for planning two-level experiments when certain interactions are important. *Technometrics*, **34**, 162–175.
- Wu, H. and C.F.J. Wu (2002). Clear two-factor interactions and minimum aberration. *Ann. Stat.*, **30**, 1496–1511.
- Xu, H. (2002). An algorithm for constructing orthogonal and nearly-orthogonal arrays with mixed levels and small runs. *Technometrics*, **44**, 356–368.
- Xu, H. (2003). Minimum moment aberration for nonregular designs and supersaturated designs. *Stat. Sin.*, **13**, 691–708.
- Xu, H. (2006). Blocked regular fractional factorial designs with minimum aberration. *Ann. Stat.*, **34**, 2534–2553.
- Xu, H. and C.-S. Cheng (2008). A complementary design theory for doubling. *Ann. Stat.*, **36**, 445–457.
- Xu, H. and S. Lau (2006). Minimum aberration blocking schemes for two- and three-level fractional factorial designs. *J. Stat. Plann. Inference*, **136**, 4088–4118.
- Xu, H. and C.F.J. Wu (2001). Generalized minimum aberration for asymmetrical fractional factorial designs. *Ann. Stat.*, **29**, 1066–1077.
- Xu, H., F.K.H. Phoa, and W.K. Wong (2009). Recent developments in nonregular fractional factorial designs. *Stat. Surv.*, **3**, 18–46.
- Zhang, R.C., P. Li, S.L. Zhao, and M.Y. Ai (2008). A general minimum lower-order confounding criterion for two-level regular designs. *Stat. Sin.*, **18**, 1689–1705.

C H A P T E R 10

Designs for Choice Experiments for the Multinomial Logit Model

Deborah J. Street and Leonie Burgess

10.1 INTRODUCTION

People make choices all the time; some of these are minor, like deciding what clothes to wear to work today, but some are major and of interest to governments and businesses. Governments might be interested in modeling demand for health services in the future, for instance, or in assessing the likely impact on the electorate of a decision to allow mining or logging in national parks. Businesses want to predict the likely market for new goods and services.

To get information about products or services that do not yet exist, an experimental approach is appropriate. Such experiments are called “stated preference” or “stated choice” experiments. This chapter provides an overview of the best way to design generic stated preference choice experiments from a mathematical perspective. A more extensive discussion appears in Street and Burgess (2007).

Throughout this chapter, we assume that all of the options in each choice set are described by several attributes, and that each attribute has two or more levels. We will assume that all the choice sets in a particular experiment have the same number of options. We assume that a multinomial logit (MNL) model, the workhorse model for analyzing choice experiments, will be used to analyze the results of the stated choice experiment. Good designs for the MNL model under the null hypothesis have shown themselves to be robust and to perform well for other choice models and at other values of the unknown

parameters. Indeed Ferrini and Scarpa (2007) performed a simulation study comparing various design strategies and concluded, “However, if good quality *a priori* information is lacking, and if there is strong uncertainty about the real DGP (data generating process)—conditions which are quite common in environmental valuation—then practitioners might be better off with shifted designs built from conventional fractional factorial designs for linear models.” We focus on the construction of such designs in this chapter.

In the next section, we describe a typical stated choice experiment, give references to several published choice experiments, and list the designs that we will use as building blocks to construct choice experiments, referring to relevant results in the first two volumes.

In Section 10.3, we discuss the multinomial logit model at some length, and in Section 10.4, we discuss ways of comparing designs. Sections 10.5 and 10.6 discuss the use of conventional designs in the construction of choice experiments, Section 10.7 looks at the use of prior values and at Bayesian approaches in general, and Section 10.8 briefly considers the design of choice experiments when both the best and worst option in each choice set are indicated. We close with a brief discussion of other work in the area.

10.2 DEFINITIONS

Stated choice experiments are easy to describe. A *stated choice experiment* consists of a set of choice sets. Each choice set consists of two or more options (or alternatives). Each respondent (also called subject) is shown each choice set in turn and asked to choose the option they think is best (or worst) from among the options presented in the choice set. The number of options in a choice set is called the *choice set size*. A stated choice or stated preference choice experiment is often called a *discrete choice experiment*, and the abbreviations *SP experiment* and *DCE* are very common. We will focus on the design of choice experiments for the simplest stated preference situation in this chapter, the so-called *generic* stated preference choice experiment, in which we assume that all options in each choice set, other than the “none of these” option, if present, are described by the same set of attributes, and each of these attributes can take one level from a finite set of possible levels.

Example 10.1. In a discrete choice experiment to investigate the effect of various attributes of energy efficient light bulbs on the decision to purchase such a bulb, there were six attributes: *quality of light* with four levels, *lifetime of bulb* with four levels, *recycling available* with two levels, *dimmable* with two levels, *time to reach full brightness* with two levels, and *cost* with four levels. The attributes and levels are given in Table 10.1. There were three options in each of the choice sets, and a sample choice set is given in Table 10.2.

Table 10.1 Attributes and Levels for the Light Bulb Study

Attribute	Attribute Levels			
	0	1	2	3
Quality of light	Daylight white	Cool white	White	Warm white
Lifetime of bulb	6000 hours	9000 hours	12,000 hours	15,000 hours
Recycling	Shopping mall	Kerbside		
Dimmable	No	Yes		
Time to full brightness	5 seconds	60 seconds		
Cost	\$3.50	\$7	\$10.50	\$14

Table 10.2 A Sample Choice Set

Attribute	Option 1	Option 2	Option 3
Quality of light	Daylight white	White	Cool white
Lifetime of bulb	9000 hours	15,000 hours	12,000 hours
Recycling	Shopping mall	Shopping mall	Kerbside
Dimmable	Yes	Yes	No
Time to full brightness	60 seconds	60 seconds	5 seconds
Cost	\$14	\$7	\$3.50

Which of these three light bulbs would you choose? (*tick one only*)
 Option 1 Option 2 Option 3

This example illustrates the fact that in many choice experiments people are forced to choose one of the options presented. We call such an experiment a *forced choice* experiment. Sometimes every possible option is presented in the choice set but mostly a forced choice experiment is used even though the list of options presented is not exhaustive. This is done to try to find out how respondents “trade-off” the different characteristics of the options presented. A simple example is to offer a cheap flight with restrictive check-in times or a more expensive flight where there are fewer restrictions on check-in times. In reality, there might be intermediate choices, but these are not offered in the choice set.

DCEs are used extensively in a variety of areas, including marketing, transport, health economics, and environmental evaluation, among others.

For instance, Chakraborty, Ettenson, and Gaeth (1994) describe a choice experiment to investigate how consumers choose health insurance. As well as the actual company offering the insurance, 23 other attributes were used to describe health insurance plans. Respondents were presented with choice sets with four options in each and asked to indicate their preferred plan from each choice set.

Hanley, Mourato, and Wright (2001) describe a stated preference study to investigate demand for climbing in Scotland. Each choice set contained two possible climbs and a “neither of these” option. They also give a table with details of about 10 other studies that used DCEs to investigate questions in environmental evaluation.

Ryan, Gerard, and Amaya-Amaya (2008) discuss many aspects of the use of DCEs in health economics, and Guttmann, Castle, and Fiebig (2009) give many examples of the use of DCEs in the health economics context between 2001 and 2007.

10.2.1 Standard Designs

We will represent the items under consideration in a DCE by a list of factors or *attributes*, each of which is presented at one particular value or *level* from a finite set of possible levels. This suggests a natural correspondence between the treatment combinations in a complete factorial design and the complete set of items available for use when constructing a DCE.

Thus we will assume that the options in a DCE are described by k attributes and that attribute q has ℓ_q levels, represented by coded levels $0, 1, \dots, \ell_q - 1$. If $\ell_1 = \ell_2 = \dots = \ell_k$, then we say the design is *symmetric* otherwise it is *asymmetric*. In Example 10.1, $k = 6$ and $\ell_1 = \ell_2 = \ell_6 = 4$ and $\ell_3 = \ell_4 = \ell_5 = 2$ and, the sample choice set in Table 10.2 is represented by (010113, 230111, 121000).

As usual, we call the set of all $L = \prod_q \ell_q$ level combinations the *complete* factorial design and refer to a subset of the level combinations as a *fractional* factorial design (FFD). In a symmetric factorial design with ℓ_q equal to a prime or a prime power, one easy way to get a fractional factorial design is to use the solutions to a set of independent equations over the finite field $GF[\ell_q]$. Such a fractional factorial design is said to be *regular*. These are described further in HK1 (chapter 11) and HK2 (chapters 7–14).

Fractions of a factorial design are closely related to orthogonal arrays and orthogonal main effects plans, and we recall the relevant definitions below.

An *orthogonal array* $OA[R, k, \ell, t]$ is a $R \times k$ array with elements from a set of ℓ symbols such that any $R \times t$ subarray has each t -tuple appearing as a row R/ℓ^t times. Often R/ℓ^t is called the *index* of the array, t the *strength* of the array, k is the *number of constraints*, and ℓ is the *number of levels*. It is worth bearing in mind that an orthogonal array of strength t is a fractional factorial design of resolution $t + 1$.

This gives us a definition that we can easily generalize to asymmetric factorials. The estimability properties of these asymmetric orthogonal arrays are the same as those of symmetric orthogonal arrays of the same strength; see Hedayat, Sloane, and Stufken (1999) for a formal proof.

An *asymmetric orthogonal array* $OA[R; \ell_1, \ell_2, \dots, \ell_k; t]$ is a $R \times k$ array with elements from a set of ℓ_q symbols in column q such that any $R \times t$ subarray has each t -tuple appearing as a row an equal number of times. Such an array is said to have *strength* t .

We will usually use “orthogonal array” for either an asymmetric or a symmetric array.

Orthogonal arrays of strength two are a subset of the class of orthogonal main effect plans. We let n_{xq} be the number of times that level x appears in column q of the array. A k factor, R run, ℓ_q -level, $1 \leq q \leq k$, *orthogonal main effect plan* (OMEP) is an $R \times k$ array with symbols $0, 1, \dots, \ell_q - 1$ in column q such that for any pair of columns q and p , the number of times that the ordered pair (x,y) appears in the columns is $n_{xq}n_{yp}/R$. In an ordinary least squares analysis, it can be shown that the main effects can be estimated independently from the results of an OMEP; see Dey (1985) and HK2 (chapter 14).

Sometimes, several factors will have the same number of levels, and this is often indicated by powers. So an OA[32;2,2,2,4,4;4] is written as OA[32;2³,4²;4]. Another common notation for an OA or an OMEP is to use $\ell_1 \times \ell_2 \times \dots \times \ell_k // R$ for an OA[R; $\ell_1, \ell_2, \dots, \ell_k; t$], most often when $t = 2$, or the fact that $t > 2$ is not relevant.

10.3 THE MNL MODEL

We assume that each subject chooses, from each choice set, the option that is “best” for them. The researcher knows which options have been compared in each choice set and which option has been selected, but has no idea how the subject has decided the relative value of each option. However, the researcher assumes that these relative values are a function of the levels of the attributes of the options under consideration, some of which have been deliberately varied by the researcher.

The aim of this section is to recall the MNL model, which is commonly used when analyzing DCEs, and the corresponding information matrix, since often the information matrix is used to compare designs. Readers who want more details are referred to Train (2003) or to chapter 3 of Street and Burgess (2007). Other models are outlined briefly in Section 10.9.

The MNL model is the model that is most often used to analyze DCEs and for which most design results exist. Following Train (2003), we define *utility* as “the net benefit derived from taking some action”; in a choice experiment, we assume that each subject chooses the option that has maximum utility from the ones available in each choice set. Thus, each subject assigns some utility to each option in a choice set and then the subject chooses the option with the maximum utility. If we let $U_{j\alpha}$ be the utility assigned by subject α to option j , $j = 1, 2, \dots, m$, in a choice set with m options, then option i is chosen if and only if $U_{i\alpha} > U_{j\alpha} \forall j \neq i$. The researcher does not see the utilities but only the options offered and the choice made (from each of the choice sets). These options are usually described by levels of several attributes. The systematic component of the utility that the researcher captures will be denoted by $V_{j\alpha}$, and we assume that $U_{j\alpha} = V_{j\alpha} + \varepsilon_{j\alpha}$, where $\varepsilon_{j\alpha}$ includes all the things that affect the utility that have not been included in $V_{j\alpha}$. Thus, the $\varepsilon_{j\alpha}$ are random terms,

and we get different choice models depending on the assumptions that we make about the distribution of the $\varepsilon_{j\alpha}$. Train (2003) shows how a choice process that bases choices on the principle of random utility maximization can result in the MNL model if the $\varepsilon_{j\alpha}$ are assumed to be independent extreme value type 1 random variables.

Suppose that each item is represented by a k -tuple (x_1, x_2, \dots, x_k) , where x_q is the level of attribute q . Order the items lexicographically so $(0, 0, \dots, 0)$ is item 1, $(0, 0, \dots, 1)$ is item 2, and so on, and let $\pi_i = e^{V_i}$, where V_i is the deterministic part of the utility associated with item i and is assumed to depend only on the item and not on the particular respondent. We call π_i the *merit* of item i . We will write $V_i = \beta' \mathbf{x}_i$, where \mathbf{x}_i is a column vector that includes all the x terms needed to fit the model of interest, and β is a column vector that contains the parameters to be estimated. So in the case of a main effects term for attribute q , the level x_q would be replaced by a set of $\ell_q - 1$ contrasts (see HK1, chapter 7), and a two-factor interaction term involving attributes q_1 and q_2 would be represented by the $(\ell_{q_1} - 1)(\ell_{q_2} - 1)$ component-wise products of the corresponding main effects terms. Similar representations for higher-order interactions are possible. The next example illustrates these concepts.

Example 10.2. Recall the light bulb study of Example 10.1. There, $k = 6$ and $\ell_1 = \ell_2 = \ell_6 = 4$ and $\ell_3 = \ell_4 = \ell_5 = 2$. There is one contrast for the two-level attributes, and we will use coefficients -1 and 1 . There are three contrasts for the four-level attributes. We could use the linear, quadratic and cubic orthogonal polynomial contrasts derived in HK1 (chapter 7), or we could use the (non-orthogonal) contrasts that compare each of levels 1, 2, and 3 with level 4. We will adopt this second approach, which is often called *effects coding*, here. Thus, the contrasts are $(1, 0, 0, -1)$, $(0, 1, 0, -1)$, and $(0, 0, 1, -1)$. Obviously, any other full rank coding could be used, and, for quantitative factors, it is perhaps more natural to use the coding from orthogonal polynomials; see Street and Burgess (2007). It should be noted that any coding will result in the same estimated probabilities from a given choice set.

Thus, if we are estimating main effects only, each light bulb described by the six attributes has three entries for each of the four-level attributes, the relevant coefficients of the contrasts associated with the four-level attribute, and one entry for each two-level attribute in \mathbf{x}_i . For instance, the light bulb in option 1 of the sample choice set in Table 10.2 is described by attribute levels $(0, 1, 0, 1, 1, 3)$ and has $\mathbf{x}_{(0,1,0,1,1,3)} = (1, 0, 0, 0, 1, 0, 1, -1, -1, -1, -1, -1)'$, since level 0 of a four-level attribute has coefficients 1, 0, and 0 in the three contrasts, for example, and level 3 of a four-level attribute has coefficients $-1, -1$, and -1 . Similarly, options 2 and 3 in Table 10.2 have $\mathbf{x}_{(2,3,0,1,1,1)} = (1, 0, 1, -1, -1, -1, 1, -1, -1, 0, 1, 0)'$ and $\mathbf{x}_{(1,2,1,0,0,0)} = (0, 1, 0, 0, 0, 1, -1, 1, 1, 1, 0, 0)'$, respectively. The β vector for estimating main effects only is given by

$$\beta = (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{20}, \beta_{21}, \beta_{22}, \beta_{30}, \beta_{40}, \beta_{50}, \beta_{60}, \beta_{61}, \beta_{62})',$$

where β_{qj} represents the j th effect for attribute q .

Consider an experiment in which there are N choice sets of m options, of which n_{i_1, i_2, \dots, i_m} compare the specific options $T_{i_1}, T_{i_2}, \dots, T_{i_m}$, where

$$n_{i_1, i_2, \dots, i_m} = \begin{cases} 1 & \text{if } (T_{i_1}, T_{i_2}, \dots, T_{i_m}) \text{ is a choice set,} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$N = \sum_{i_1 < i_2 < \dots < i_m} n_{i_1, i_2, \dots, i_m}.$$

Given a choice set that contains m options $T_{i_1}, T_{i_2}, \dots, T_{i_m}$ the probability that option T_{i_1} is preferred to the other $m - 1$ options in the choice set is

$$P(T_{i_1} > T_{i_2}, \dots, T_{i_m}) = \frac{\pi_{i_1}}{\sum_{j=1}^m \pi_{i_j}},$$

and we get a similar probability for each of the items in the choice set (assuming that all options in each choice set are distinct). We also assume that choices made in one choice set do not affect choices made in any other choice set.

Hence, we can evaluate the likelihood function associated with a set of choice sets, transform to the parameters $V_i = \ln(\pi_i) = \beta' \mathbf{x}_i$ and evaluate Fisher's information matrix for the V_i s, $\Lambda(\boldsymbol{\pi})$, say. We define $\Lambda(\boldsymbol{\pi})$ to be the matrix of expectations of the second derivatives of the log-likelihood function, where $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_L)$. Since this matrix depends on the parameters that we are trying to estimate, we must make some assumptions about the values of these parameters in order to construct and compare designs. The most common assumption is that $\boldsymbol{\pi} = \mathbf{j}_L$, where \mathbf{j}_L is a vector of L ones, equivalent to an assumption that all the β parameters are equal to 0. The general form of the $\Lambda(\boldsymbol{\pi})$ matrix for generic forced choice DCEs is given in Section 10.5.1, and for extensions of this model in Section 10.5.2.

Often, we are only interested in some contrasts of the V_i , typically those that correspond to the main effects or the main effects and two-factor interactions, and so to specific β_i s. Thus, we define a contrast matrix, \mathbf{B} , which has the contrasts of interest as its rows. Then the information matrix for \mathbf{BV} is $\mathbf{BAB}' = \mathbf{C}$. This transformation does not change the dependence of the information matrix on the unknown parameters, of course, but usually reduces the rank of the information matrix and hence the size of the design required to estimate the effects of interest.

10.4 DESIGN COMPARISONS

In the univariate, ordinary least squares, normal errors situation, the performance of estimates is often judged by the width of the resulting confidence interval. Estimates that are minimum variance unbiased are often deemed to be the best. In the multivariate, ordinary least squares, normal errors situation, the performance of estimates is often judged by some property of the asymptotic variance–covariance matrix. For a DCE, we find the variance–covariance matrix as a function of the parameters and define the optimal design to be the design with the “best” variance–covariance matrix, dealing with the inherent circularity by either evaluating the variance–covariance matrix at a specific prior value for the parameters, often $\mathbf{0}$, or by taking a Bayesian approach (see Section 10.7). Possible ways of defining “best” are given below.

Using the variance–covariance matrix, \mathbf{C}^{-1} , ensures that we have a method to compare different choice experiments objectively by using some function of the eigenvalues of the matrix. We could also compare designs using structural properties of either the variance–covariance matrix or of the designs.

For the variance–covariance matrix, the structural property that is of most interest is whether or not the parameters have been independently estimated. This is so if the matrix is diagonal. When we estimate main effects, for an attribute with ℓ_q levels, there are $\ell_q - 1$ contrasts, and these are not uniquely defined. So, in fact, we really only require that \mathbf{C}^{-1} be block diagonal for the effects from different attributes to be independently estimated.

Burgess (2007) provides software to calculate the information matrix and variance–covariance matrix of any set of choice sets, under a prior assumption that the parameters are all 0, for generic forced choice designs analyzed using the MNL model.

For the designs themselves, desirable structural properties typically arise from a link between the structure of the design and the properties of the resulting estimates. Structural properties of interest could include the frequency with which each level of each attribute appears in the design or the relationship between the options in each choice set.

We discuss these different approaches below.

10.4.1 Optimality

For the estimation of a single parameter, the estimated variance of the parameter estimate is a natural statistic to use to compare competing estimates, although other properties, such as linearity, may also play a part in estimator comparison. When a vector of parameters is to be estimated, the natural analogue of the estimated variance is the estimated variance–covariance matrix.

The D -, A -, and E -optimality measures are appropriate to our situation, and we now define these; see Pukelsheim (1993) or Atkinson, Donev, and Tobias (2007) for an extensive collection of results on optimal designs for the linear model for a variety of optimality criteria. Both books briefly discuss some of

the issues in the construction of optimal designs for models in which the covariance matrix depends on the unknown parameters (as it does for the MNL model).

A design is *D-optimal* if it minimizes the generalized variance of the parameter estimates, that is, $\det(\mathbf{C}^{-1})$ is as small as possible for the *D*-optimal designs.

A design is *A-optimal* if it minimizes the average variance of the parameter estimates, that is, $\text{tr}(\mathbf{C}^{-1})$ is as small as possible for the *A*-optimal designs.

A design is *E-optimal* if it minimizes the variance of the least well-estimated parameter, that is, the largest eigenvalue of \mathbf{C}^{-1} is as small as possible for the *E*-optimal designs.

These three measures focus on the properties of the parameter estimates. Kessels, Goos, and Vandebroek (2006) have argued that since the role of a choice experiment is to make precise predictions, it makes more sense to focus on designs that are *G* and *V* optimal.

A design is *G-optimal* if it minimizes the maximum prediction variance over all possible choice sets of a given size.

A design is *V-optimal* if it minimizes the average prediction variance over all possible choice sets of a given size.

10.4.2 Structural Properties

There has been a tradition in the design of experiments to try and identify structural properties of designs that are linked with desirable statistical properties. Then, useful designs can be found merely by considering these structural properties. Although some structural properties have been shown to be linked with desirable properties in choice experiments, to date, the results have not been as clear cut as in the linear models setting.

Huber and Zwerina (1996) describe a set of features that they believe are characteristic of optimal choice designs. These features are:

1. *Level Balance*. All the levels of each attribute occur with equal frequency over all options in all choice sets (often called equireplicate in the statistical literature).
2. *Orthogonality*. The levels of each attribute vary independently of each other. This means that for any two attributes, all combinations of pairs of levels appear with proportional frequencies.
3. *Minimal Overlap*. Thus, the difference between the number of times that any two levels of an attribute are replicated within each choice set should be as small as possible, ideally 0, and at most 1.
4. *Utility Balance*. Options within a choice set should be equally attractive to subjects.

While appealing, these principles are unfortunately neither necessary nor sufficient to guarantee that a DCE satisfying them is optimal; see Street and Burgess (2007) for details.

10.5 OPTIMAL DESIGNS FOR DCEs

10.5.1 Generic Forced Choice DCEs

In a generic choice experiment, the same k attributes are used to describe the items in each of the m alternatives. Since respondents would never be asked to choose between identical items, any pair of alternatives in a choice set must differ in the levels of at least one attribute.

While most applications of DCEs have employed FFDs in some way to represent the options in the choice sets, a number of different approaches have been adopted to place the treatment combinations from the FFD into the choice sets. The most common early approach was to obtain an OMEP and then to randomly place these treatment combinations into the choice sets (see Ryan and Gerard 2003, for example). Another approach is to take several FFDs and to randomly choose the first alternative from the first FFD, the second alternative from the second FFD, and so on, until there are m alternatives. Another method is to construct an OMEP with $m \times k$ attributes in total and use the first k columns for the first alternative, the second k columns for the second alternative, and so on up to the last k columns for the m th alternative. This final construction method has been called the L^{MA} approach (see Louviere, Hensher, and Swait 2000). Bunch, Louviere, and Anderson (1996) introduced the idea of *shifted designs* in which a set of initial options is chosen for each of the N choice sets in an experiment and the subsequent option(s) in each choice set are obtained by using modular arithmetic to “shift each combination of initial attribute levels by adding a constant that depends on the number of levels.” Another common approach has been to construct designs that satisfy the four criteria in Huber and Zwerina (1996) as discussed in Section 10.4.2. For all of these approaches, nothing was known on the optimality of the designs for the general case.

Burgess and Street (2005) established the form of the optimal design under the null hypothesis for the estimation of main effects only, and optimal designs for main effects plus two-factor interactions when all attributes have two levels. They also give a method for constructing good DCEs. These results are given in more detail in Street and Burgess (2007). There is software available to construct these designs and that can check any proposed design; see Burgess (2007). We now summarize their results on optimal designs for generic forced choice DCEs.

The order of the treatment combinations in the choice sets is arbitrary, so we consider choice sets such as $(T_{i_1}, T_{i_2}, T_{i_3})$ and $(T_{i_2}, T_{i_1}, T_{i_3})$ to be the same. As before, let π_{ij} be the merit of the treatment combination T_{ij} . Given a choice set that contains m options $T_{i_1}, T_{i_2}, \dots, T_{i_m}$, the probability that option T_{i_1} is preferred to the other $m - 1$ options in the choice set is

$$P(T_{i_1} > T_{i_2}, \dots, T_{i_m}) = \frac{\pi_{i_1}}{\sum_{j=1}^m \pi_{ij}},$$

for $i_j = 1, 2, \dots, L$ (assuming that all options in each choice set are distinct).

We use the method in Burgess and Street (2003) (and Street and Burgess 2007) to derive the form of the entries in Λ .

Let w_{i_1, i_2, \dots, i_m} be an indicator variable where

$$w_{i_1, i_2, \dots, i_m} = \begin{cases} 1 & \text{if } T_{i_1} > T_{i_2}, T_{i_3}, \dots, T_{i_m}, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\mathcal{E}(w_{i_1, i_2, \dots, i_m}) = \frac{\pi_{i_1}}{\sum_{j=1}^m \pi_{i_j}} \quad \text{and} \quad \text{Var}(w_{i_1, i_2, \dots, i_m}) = \frac{\pi_{i_1} \sum_{j=2}^m \pi_{i_j}}{\left(\sum_{j=1}^m \pi_{i_j}\right)^2}.$$

We let w_{i_1} be the number of times option T_{i_1} is preferred to the other options available in any choice set in which T_{i_1} appears. Then

$$w_{i_1} = \sum_{i_2 < i_3 < \dots < i_m} w_{i_1, i_2, \dots, i_m},$$

where the summation is over $i_2 < i_3 < \dots < i_m$ and $i_j \neq i_1$ for $j = 2, \dots, m$.

It follows that

$$\mathcal{E}(w_{i_1}) = \pi_{i_1} \sum_{i_2 < i_3 < \dots < i_m} \frac{1}{\sum_{j=1}^m \pi_{i_j}},$$

$$\text{Var}(w_{i_1}) = \pi_{i_1} \sum_{i_2 < i_3 < \dots < i_m} \frac{\sum_{j=2}^m \pi_{i_j}}{\left(\sum_{j=1}^m \pi_{i_j}\right)^2},$$

$$\text{and} \quad \text{Cov}(w_{i_1}, w_{i_2}) = \sum_{i_3 < \dots < i_m} \frac{-\pi_{i_1} \pi_{i_2}}{\left(\sum_{j=1}^m \pi_{i_j}\right)^2}.$$

The likelihood function of the DCE is given by

$$L(\pi) = \frac{\pi_1^{w_1} \pi_2^{w_2} \dots \pi_L^{w_L}}{\prod_{c=1}^L \binom{m}{n_c} \left(\sum_{j=1}^m \pi_{c_j}\right)^{n_c}},$$

where n_c indicates whether or not the corresponding m -subset of the L treatments is a choice set or not. Differentiating the log likelihood function with respect to the merits, π_i , gives

$$\frac{\partial \ln(L(\pi))}{\partial \pi_i} = \frac{w_i}{\pi_i} - \sum_c \frac{n_c}{\sum_{j=1}^m \pi_{c_j}}.$$

Thus, the Fisher information matrix for $\boldsymbol{\pi}$ has entries

$$\mathcal{E} \left(\left(\frac{w_i}{\pi_i} - \sum_c \frac{n_c}{\sum_{j=1}^m \pi_{c_j}} \right) \left(\frac{w_h}{\pi_h} - \sum_c \frac{n_c}{\sum_{j=1}^m \pi_{c_j}} \right) \right).$$

We know that $V_i = \ln(\pi_i)$, so $\partial V_i / \partial \pi_i = 1/\pi_i$. Thus, the information matrix of \mathbf{V} , $\mathcal{I}_{\mathbf{V}}$, satisfies $\mathcal{I}_{\mathbf{V}} = \mathbf{P} \mathcal{I}_{\boldsymbol{\pi}} \mathbf{P}$, where \mathbf{P} is a diagonal matrix with i th entry equal to π_i . We let $\Lambda(\boldsymbol{\pi}) = \mathcal{I}_{\mathbf{V}}$.

We define $\lambda_{i_1, i_2, \dots, i_m} = n_{i_1, i_2, \dots, i_m} / N$, where $n_{i_1, i_2, \dots, i_m} = n_c$ for the m -subset $\{i_1, i_2, \dots, i_m\}$. Then the entries of $\Lambda(\boldsymbol{\pi})$ are given by

$$\Lambda(\boldsymbol{\pi})_{i_1, i_1} = \pi_{i_1} \sum_{i_2 < i_3 < \dots < i_m} \frac{\lambda_{i_1, i_2, \dots, i_m} \sum_{j=2}^m \pi_{i_j}}{\left(\sum_{j=1}^m \pi_{i_j} \right)^2}, \quad (10.1)$$

and

$$\Lambda(\boldsymbol{\pi})_{i_1, i_2} = -\pi_{i_1} \pi_{i_2} \sum_{i_3 < i_4 < \dots < i_m} \frac{\lambda_{i_1, i_2, \dots, i_m}}{\left(\sum_{j=1}^m \pi_{i_j} \right)^2}. \quad (10.2)$$

Under the null hypothesis that all treatment combinations are equally attractive (i.e., $\boldsymbol{\pi} = \mathbf{j}_L = \boldsymbol{\pi}_0$, where \mathbf{j}_L is a vector of L ones)

$$\Lambda(\boldsymbol{\pi}_0)_{i_1, i_1} = \frac{m-1}{m^2} \sum_{i_2 < i_3 < \dots < i_m} \lambda_{i_1, i_2, \dots, i_m}, \quad (10.3)$$

and

$$\Lambda(\boldsymbol{\pi}_0)_{i_1, i_2} = -\frac{1}{m^2} \sum_{i_3 < i_4 < \dots < i_m} \lambda_{i_1, i_2, \dots, i_m}. \quad (10.4)$$

Two alternatives are said to *differ in attribute q* if the level of attribute q is different in the two alternatives. Each choice set has $m(m-1)/2$ pairs of alternatives, and for each attribute, we let D_q be the number of pairs of alternatives in which the levels of attribute q differ. So if attribute q appears at the same level for all options in a choice set, then $D_q = 0$.

10.5.1.1 Estimating Main Effects Only

Let \mathbf{B}_{ℓ_q} be a normalized contrast matrix for main effects for a factor with ℓ_q levels. We let \mathbf{B}_M be the normalized contrast matrix for main effects for a $\ell_1 \times \ell_2 \times \dots \times \ell_k$ factorial, then $\mathbf{C}_M = \mathbf{B}_M \boldsymbol{\Lambda}(\boldsymbol{\pi}_0) \mathbf{B}'_M$ is the information matrix. Explicitly,

$$\mathbf{B}_M = \begin{bmatrix} \mathbf{B}_{\ell_1} \otimes \frac{1}{\sqrt{\ell_2}} \mathbf{j}'_{\ell_2} \otimes \dots \otimes \frac{1}{\sqrt{\ell_k}} \mathbf{j}'_{\ell_k} \\ \frac{1}{\sqrt{\ell_1}} \mathbf{j}'_{\ell_1} \otimes \mathbf{B}_{\ell_2} \otimes \dots \otimes \frac{1}{\sqrt{\ell_k}} \mathbf{j}'_{\ell_k} \\ \vdots \\ \frac{1}{\sqrt{\ell_1}} \mathbf{j}'_{\ell_1} \otimes \frac{1}{\sqrt{\ell_2}} \mathbf{j}'_{\ell_2} \otimes \dots \otimes \mathbf{B}_{\ell_k} \end{bmatrix}. \quad (10.5)$$

The maximum possible value of the determinant of $\mathbf{C}_M = \mathbf{B}_M \boldsymbol{\Lambda}(\boldsymbol{\pi}_0) \mathbf{B}'_M$ is given by

$$\det(\mathbf{C}_{\text{opt},M}) = \prod_{q=1}^k \left[\frac{2\ell_q S_q}{m^2 L(\ell_q - 1)} \right]^{\ell_q - 1}, \quad (10.6)$$

where

$$S_q = \begin{cases} (m^2 - (\ell_q x^2 + 2xy + y))/2 & 2 \leq \ell_q < m, \\ m(m-1)/2 & \ell_q \geq m, \end{cases} \quad (10.7)$$

and positive integers x and y satisfy the equation $m = \ell_q x + y$ for $0 \leq y < \ell_q$. S_q represents the least upper bound for D_q . Thus, the D -optimal design, for estimating main effects only, is one which consists of choice sets in which D_q is equal to S_q for each attribute q (Burgess and Street 2005).

We now consider the construction of optimal or near-optimal designs when estimating main effects only. A resolution III fractional factorial design is required as a starting design, then the choice sets are formed by adding one or more sets of generators. For an optimal design the sets of generators must be chosen so that the sum of the differences between the levels of an attribute across the choice sets is the maximum possible.

We now describe a construction that results in an optimal design. Let F be the complete factorial for k attributes where the q th attribute has ℓ_q levels. Suppose that the choice sets are to be of size m . Then we must choose a set of m generators $G = \{\mathbf{g}_1 = \mathbf{0}, \mathbf{g}_2, \dots, \mathbf{g}_m\}$ such that $\mathbf{g}_i \neq \mathbf{g}_j$ for $i \neq j$. Suppose that $\mathbf{g}_i = (g_{i1}, g_{i2}, \dots, g_{ik})$ for $i = 1, \dots, m$, and suppose that the multiset of differences for attribute q $\{\pm(g_{i1q} - g_{i2q}) \mid 1 \leq i_1, i_2 \leq m, i_1 \neq i_2\}$ contains each nonzero difference modulo ℓ_q equally often. Then the choice sets given by the rows of $F + \mathbf{g}_1, F + \mathbf{g}_2, \dots, F + \mathbf{g}_m$, for one or more sets of generators G are optimal

for the estimation of main effects only, provided that there are as few zero differences as possible in each choice set.

We begin by calculating the values of x and y (where $m = \ell_q x + y$) so we know that we have y -values (between 0 and $\ell_q - 1$) that are repeated $x + 1$ times each and $(\ell_q - y)$ values that are repeated x times each. We then partition the values between 0 and $\ell_q - 1$ into two disjoint sets, one containing y entries and the other containing the remaining $(\ell_q - y)$ entries. There are $\binom{\ell_q}{y}$ ways to do this. For each partition, we calculate the differences that arise from a vector with m entries in which the entries in the set with y entries are each repeated $x + 1$ times and the entries in the other set are each repeated x times. All such vectors have as few 0 differences as possible in the $m(m - 1)$ differences. Next, we partition the vectors into sets based on the number of times each nonzero entry modulo ℓ_q appears as a difference. We then choose the vectors to take from each partition so that we have all nonzero differences appearing equally often over the set of vectors chosen. If there are several attributes, perhaps with different numbers of levels, then we must choose the same number of vectors for each attribute. Once we have the vectors for each attribute, then we can calculate the entries for the sets of generators by choosing one entry from each vector for each generator in such a way that no generator is repeated. We illustrate this construction in the following example.

Example 10.3. Let $k = 3$, $\ell_1 = 2$, $\ell_2 = 3$, and $\ell_3 = 6$. Suppose $m = 3$. Since we want each set of generators to contain 000, we only consider vectors that contain 0 for each attribute. There is one partition of the vectors for the first attribute with entries $(0,0,1)$ and $(0,1,1)$. There is one vector for the second attribute: $(0,1,2)$. There are $\binom{5}{2} = 10$ vectors for the third attribute (since we must include 0), and these are partitioned into three sets. The first set consists of $(0,1,2)$, $(0,1,5)$, and $(0,4,5)$ with nonzero differences 1 and 5 twice each and 2 and 4 once each. The second set has $(0,2,4)$ with nonzero differences 2 and 4 thrice each. The third set has the remaining six vectors, $(0,1,3)$, $(0,1,4)$, $(0,2,3)$, $(0,2,5)$, $(0,3,4)$, and $(0,3,5)$, and each of these has nonzero differences 1, 2, 4, and 5 once each and 3 twice. Suppose that we have x_1 vectors from the first partition for attribute 3, x_2 from the second, and x_3 from the third. Because we want to have each nonzero difference appearing equally often, we get $2x_1 + x_3 = x_1 + 3x_2 + x_3 = 2x_3$. Suppose $x_2 = 0$. Then $x_1 = x_3 = 0$, which is a contradiction. Instead, try $x_2 = 1$. Then $2x_1 + x_3 = x_1 + x_3 + 3$; so $x_1 = 3$, and hence $x_3 = 6$. Thus, for an optimal design, we need to have 10 sets of generators and 360 choice sets. We could use $(0,1,1)$ in all 10 sets of generators for attribute 1, $(0,1,2)$ in all 10 sets of generators for attribute 2, and $(0,1,2)$, $(0,1,3)$, $(0,1,4)$, $(0,1,5)$, $(0,2,3)$, $(0,2,4)$, $(0,2,5)$, $(0,3,4)$, $(0,3,5)$, and $(0,4,5)$ for attribute 3. This gives the 10 sets of generators $G_1 = (000, 111, 122)$, $G_2 = (000, 111, 123)$, $G_3 = (000, 111, 124)$, $G_4 = (000, 111, 125)$, $G_5 = (000, 112, 123)$, $G_6 = (000, 112, 124)$, $G_7 = (000, 112, 125)$, $G_8 = (000, 113, 124)$, $G_9 = (000, 113, 125)$, and $G_{10} = (000, 114, 125)$. For example, by adding G_1 to treatment combination 105,

we get the choice set (105, 010, 021), where the addition of the levels of the three attributes is performed in modulo 2, 3, and 6, respectively.

10.5.1.2 Estimating Main Effects Plus Two-Factor Interactions

We now consider designs when main effects plus two-factor interactions are to be estimated. Let \mathbf{B}_T be the normalized rows of \mathbf{B} that correspond to the two-factor interactions. The contrast matrix associated with main effects and two-factor interactions is denoted by \mathbf{B}_{MT} and is the concatenation of \mathbf{B}_M and \mathbf{B}_T . Then

$$\mathbf{B}_{MT} = \begin{bmatrix} \mathbf{B}_M \\ \mathbf{B}_T \end{bmatrix}, \quad (10.8)$$

where \mathbf{B}_M is defined in Equation (10.5) and

$$\mathbf{B}_T = \begin{bmatrix} \mathbf{B}_{\ell_1} \otimes \mathbf{B}_{\ell_2} \otimes \frac{1}{\sqrt{\ell_3}} \mathbf{j}'_{\ell_3} \otimes \cdots \otimes \frac{1}{\sqrt{\ell_k}} \mathbf{j}'_{\ell_k} \\ \mathbf{B}_{\ell_1} \otimes \frac{1}{\sqrt{\ell_2}} \mathbf{j}'_{\ell_2} \otimes \mathbf{B}_{\ell_3} \otimes \cdots \otimes \frac{1}{\sqrt{\ell_k}} \mathbf{j}'_{\ell_k} \\ \vdots \\ \frac{1}{\sqrt{\ell_1}} \mathbf{j}'_{\ell_1} \otimes \cdots \otimes \frac{1}{\sqrt{\ell_{k-2}}} \mathbf{j}'_{\ell_{k-2}} \otimes \mathbf{B}_{\ell_{k-1}} \otimes \mathbf{B}_{\ell_k} \end{bmatrix}. \quad (10.9)$$

The $p \times p$ information matrix is $\mathbf{C}_{MT} = \mathbf{B}_{MT} \boldsymbol{\Lambda}(\boldsymbol{\pi}_0) \mathbf{B}'_{MT}$, where

$$p = \sum_{q=1}^k (\ell_q - 1) + \sum_{q_1=1}^{k-1} \sum_{q_2=q_1+1}^k (\ell_{q_1} - 1)(\ell_{q_2} - 1).$$

There are no general results on optimal designs for any choice set size with attributes that can have any number of levels, although Street and Burgess (2007) give tables of optimal designs for various values of m , k , and ℓ_q for $q = 1, \dots, k$. Burgess and Street (2003) show that if all the attributes have two levels, then the optimal design consists of all choice sets in which the number of attributes that differ between any pair of treatment combinations in the choice set is $D_q = (k + 1)/2$ if k is odd, or $D_q = k/2$ or $D_q = k/2 + 1$, if k is even. The maximum determinant is

$$\det(\mathbf{C}_{opt,MT}) = \begin{cases} \left(\frac{(m-1)(k+2)}{m(k+1)2^k} \right)^{k+k(k-1)/2} & k \text{ even,} \\ \left(\frac{(m-1)(k+1)}{mk2^k} \right)^{k+k(k-1)/2} & k \text{ odd.} \end{cases} \quad (10.10)$$

Example 10.4. Let $m = 3$ and $k = 4$. The optimal design will have $k/2 = 2$ or $k/2 + 1 = 3$ attribute level differences between each pair of alternatives in the choice sets, and the maximum obtainable determinant is $\det(\mathbf{C}_{\text{opt},MT}) = (1/20)^{10}$. There are 160 choice sets that have two or three attribute level differences. However, we can use a resolution V fractional factorial design as the starting design and add sets of generators to obtain the remaining alternatives in the choice sets to get a near-optimal design in fewer choice sets. Since there are no fractions of the 2^4 factorial that are resolution V, we must use the 16 treatment combinations from the complete factorial. These treatment combinations are given in the first column of Table 10.3. We choose $G_1 = (\mathbf{0}, \mathbf{g}_{21}, \mathbf{g}_{31}) = (0000, 1100, 0110)$ and form 16 choice sets by adding the elements of G_1 to F . From these, we can estimate all the main effects and two-factor interactions except the main effect of the fourth attribute. So we need another set of generators, G_2 , to add to F to get 16 more choice sets. We choose $G_2 = (\mathbf{0}, \mathbf{g}_{22}, \mathbf{g}_{32}) = (0000, 1100, 0111)$. If we use G_2 only, then the two-factor interaction between attributes 3 and 4 cannot be estimated. With both G_1 and G_2 , all main effects and two-factor interactions can now be estimated; so we have no need to generate any more choice sets. The 32 choice sets shown in Table 10.3 form a design with $\det(\mathbf{C}_{MT}) = (1/18)^8(1/36)^2$. Thus, this design is 96.73% efficient.

Grasshoff et al. (2004a) derive optimal designs when estimating main effects plus two-factor interactions when all the attributes have the same number of levels and $m = 2$. They provide a table that gives the requisite number of attributes to have different between pairs of treatment combinations in a choice

Table 10.3 Near-Optimal Choice Sets for Estimating Main Effects and Two-Factor Interactions for $m = 3$ and $k = 4$

F	$F + \mathbf{g}_{21}$	$F + \mathbf{g}_{31}$	F	$F + \mathbf{g}_{22}$	$F + \mathbf{g}_{32}$
0 0 0 0	1 1 0 0	0 1 1 0	0 0 0 0	1 1 0 0	0 1 1 1
0 0 0 1	1 1 0 1	0 1 1 1	0 0 0 1	1 1 0 1	0 1 1 0
0 0 1 0	1 1 1 0	0 1 0 0	0 0 1 0	1 1 1 0	0 1 0 1
0 0 1 1	1 1 1 1	0 1 0 1	0 0 1 1	1 1 1 1	0 1 0 0
0 1 0 0	1 0 0 0	0 0 1 0	0 1 0 0	1 0 0 0	0 0 1 1
0 1 0 1	1 0 0 1	0 0 1 1	0 1 0 1	1 0 0 1	0 0 1 0
0 1 1 0	1 0 1 0	0 0 0 0	0 1 1 0	1 0 1 0	0 0 0 1
0 1 1 1	1 0 1 1	0 0 0 1	0 1 1 1	1 0 1 1	0 0 0 0
1 0 0 0	0 1 0 0	1 1 1 0	1 0 0 0	0 1 0 0	1 1 1 1
1 0 0 1	0 1 0 1	1 1 1 1	1 0 0 1	0 1 0 1	1 1 1 0
1 0 1 0	0 1 1 0	1 1 0 0	1 0 1 0	0 1 1 0	1 1 0 1
1 0 1 1	0 1 1 1	1 1 0 1	1 0 1 1	0 1 1 1	1 1 0 0
1 1 0 0	0 0 0 0	1 0 1 0	1 1 0 0	0 0 0 0	1 0 1 1
1 1 0 1	0 0 0 1	1 0 1 1	1 1 0 1	0 0 0 1	1 0 1 0
1 1 1 0	0 0 1 0	1 0 0 0	1 1 1 0	0 0 1 0	1 0 0 1
1 1 1 1	0 0 1 1	1 0 0 1	1 1 1 1	0 0 1 1	1 0 0 0

set. However, this does not always result in the optimal design. For example, when there are $k = 8$ attributes, each with 6 levels, the table in Grasshoff et al. (2004a) says that the optimal design will have choice sets with the levels of six and seven attributes different, whereas the optimal design for this situation has seven differences between the levels of the attributes in each pair (see Street and Burgess 2007, chapter 6, appendix A.7).

For the more general case of any choice size and attributes with any number of levels, there is no general result on the optimal designs when estimating main effects plus some or all two-factor interactions. The maximum $\det(\mathbf{C}_{MT})$ for some small examples are given in Street and Burgess (2007), chapter 6, appendix A.7. However, we do have an expression for $\det(\mathbf{C}_{MT})$, which allows us to compare different choice experiments. In this situation, we have found it is best to construct some large designs to try to find the best $\det(\mathbf{C}_{MT})$, then use this as a basis for determining the efficiency of smaller designs that can be used in practice. This can easily be done using freely available software (Burgess 2007). The software can also calculate the appropriate information matrix if only a subset of the two-factor interactions are to be estimated.

We construct designs by using a resolution V fractional factorial design as a starting design and adding sets of generators to get the rest of the alternatives in the choice sets. At the moment, there are not many resolution V fractional factorial designs available with at least one attribute with more than two levels; see Sloane (2006) and constructions in Street and Burgess (2007), Dey (1985) and Hedayat, Sloane, and Stufken (1999).

The sets of generators need to satisfy two conditions:

1. For each attribute, there must be at least one generator with a nonzero value in the corresponding position (to estimate main effects);
2. For two attributes for which the two-factor interaction is to be estimated, there must be at least one generator in which the corresponding positions have a 0 and a nonzero value.

Example 10.5. Suppose that $k = 5$, $\ell_1 = \ell_2 = \ell_3 = \ell_4 = 2$, $\ell_5 = 4$, and $m = 4$. We can use a resolution V fractional factorial design in 32 runs (see Street and Burgess 2007, example 8.2.2). By trying various sets of generators that satisfy the aforementioned conditions, we find the most efficient design (Table 10.4) in 32 choice sets is obtained by using the set of generators $G = (00000, 00111, 10102, 01110)$. This design is 96.29% efficient, relative to the largest $\det(\mathbf{C}_{MT})$ we could find.

10.5.1.3 Practical Techniques for Design Construction

The formal construction of designs for estimating main effects only in Section 10.5.1.1 is complicated to use and often results in a large number of choice sets. In practice, it is usual to use a fractional factorial design, F , of a resolution of at least III as the starting design, then add one set of generators G to get near-optimal choice sets. The generators are chosen so that the choice

Table 10.4 Choice Sets for $k = 5$, $\ell_1 = \ell_2 = \ell_3 = \ell_4 = 2$, $\ell_5 = 4$ when $m = 4$ for Main Effects and All Two-Factor Interactions

F	$F + 00111$	$F + 10102$	$F + 01110$	F	$F + 00111$	$F + 10102$	$F + 01110$
00000	00111	10102	01110	00012	00103	10110	01102
00110	00001	10012	01000	00102	00013	10000	01012
01010	01101	11112	00100	01002	01113	11100	00112
01100	01011	11002	00010	01112	01003	11010	00002
10010	10101	00112	11100	10002	10113	00100	11112
10100	10011	00002	11010	10112	10003	00010	11002
11000	11111	01102	10110	11012	11103	01110	10102
11110	11001	01012	10000	11102	11013	01000	10012
00001	00112	10103	01111	00013	00100	10111	01103
00111	00002	10013	01001	00103	00010	10001	01013
01011	01102	11113	00101	01003	01110	11101	00113
01101	01012	11003	00011	01113	01000	11011	00003
10011	10102	00113	11101	10003	10110	00101	11113
10101	10012	00003	11011	10113	10000	00011	11003
11001	11112	01103	10111	11013	11100	01111	10103
11111	11002	01013	10001	11103	11010	01001	10013

sets have the maximum number of differences in the levels of each of the attributes.

In Example 10.3, using only G_1 gives a design with 36 choice sets that is 97.80% efficient, and using only G_2 gives 36 choice sets that are 99.39% efficient. On the other hand, using only G_6 with elements (0,2,4) for the third attribute results in $\det(\mathbf{C}_M) = 0$, which is not surprising, since only two of the differences are represented. In some cases, one set of generators will give rise to an optimal design if there is a difference set for the appropriate values of ℓ_q and m . For example, there is an optimal set of 56 choice sets when $k = 3$, $\ell_1 = 2$, $\ell_2 = 4$, $\ell_3 = 7$, and $m = 4$ obtained by using the set of generators $G = (000, 011, 122, 134)$. This is because the differences from the set $\{0, 1, 2, 4\}$ contain each nonzero difference modulo 7 exactly once.

We now consider another example.

Example 10.6. Suppose $k = 13$, $\ell_q = 2$, for $q = 1, \dots, 12$, $\ell_{13} = 4$, and $m = 3$. Starting with the $4^5/16$ FFD of resolution III, which is given in the first column in Table 10.5, we replace the four levels of the first attribute by the four treatment combinations of the $2^3//4$. That is, we replace the first four-level attribute by three two-level attributes by replacing 0 in the first attribute with 000, 1 with 011, 2 with 101, and 3 with 110. We repeat this procedure for the second, third, and fourth four-level attributes to get the $2^{12} \times 4//16$ design shown in the second column in Table 10.5. If we add the set of generators $G = (00000000000000, 11111111111111, 1010101010102)$ to this starting design, then we get the choice

Table 10.5 $2^{12} \times 4/16$ Obtained from $4^5/16$ by Expansive Replacement, and Choice Sets for $m = 3$

$4^5/16$	$F = 2^{12} \times 4/16$	$F + 111111111111$	$F + 1010101010102$
00000	00000000000000	111111111111	1010101010102
01111	0000110110111	1111001001002	1010011100013
02222	0001011011012	1110100100103	101110001110
03333	0001101101103	1110010010010	1011000111001
10123	0110000111013	1001111000100	1100101101111
11032	0110110001102	1001001110013	1100011011000
12301	0111011100001	1000100011112	1101110110103
13210	0111101010110	1000010101001	1101000000012
20231	1010001011101	0101110100012	0000100001003
21320	1010111101010	0101000010101	0000010111112
22013	1011010000113	0100101111000	0001111010011
23102	1011100110002	0100011001113	0001001100100
30312	1100001100112	0011110011003	0110100110010
31203	1100111010003	0011000101110	0110010000101
32130	1101010111100	0010101000011	011111101002
33021	1101100001011	0010011110102	0111001011113

sets that are also given in Table 10.5. This design is 100% efficient, and the information matrix is $\mathbf{C}_M = 1/18432 \mathbf{I}_{15}$.

When estimating main effects and two-factor interactions, a larger number of choice sets are required. In practice, it is usual to use a fractional factorial design, F , of a resolution of at least V as the starting design, then create the choice sets by adding as many sets of generators as are required in order to be able to estimate all of the effects of interest. As we noted previously, for each attribute, there must be at least one generator with a nonzero value in the corresponding position to estimate the main effects, and for any two attributes, there should be at least one generator in which the corresponding positions have a zero and a nonzero value. For example, for $k = 5$, $\ell_1 = \ell_2 = \ell_3 = \ell_4 = 2$, and $\ell_5 = 4$, suppose that $m = 3$. By starting with a resolution V design in 32 treatment combinations, we can construct a design in 64 choice sets by adding two sets of generators. One near-optimal design is given by $(F, F + 00111, F + 10102)$ and $(F, F + 01103, F + 11010)$. This design is 96.44% efficient relative to the best design we could find. More examples are given in section 8.2 of Street and Burgess (2007).

When the levels of all of the attributes are ordered in some way, from best to worst or vice versa, then some treatment combinations may dominate other treatment combinations. In this situation, we need to avoid starting designs that contain the treatment combination 000 ... 0 or the treatment combination with all attributes at the high level. This can be achieved by choosing a

different fraction of the factorial design. In this case, the best way of choosing sets of generators is to have a mix of levels so that in some attributes, a low value is added, in some attributes, middle values are added, and in the remaining attributes, high values are added. Thus, no choice set will contain treatment combinations that dominate other treatment combinations. Trial and error may be needed in the selection of F and G to achieve optimality, or to get choice sets that are appropriate for the particular situation.

Example 10.7. In Table 10.5, the first choice set contains the treatment combination 00000000000000. If the attribute levels are ordered from least preferred at level 0 to most preferred at level $\ell_q - 1$, then the treatment combination of all zeros will be dominated by any other treatment combination. We can use a different fraction for the starting design to avoid this problem. First, we choose a treatment combination that is not in F . For example, the treatment combination 1111110000000 is not in F , and if we add it to F , using the modular arithmetic appropriate to each attribute, we will obtain a different fraction that is still an OMEP. This OMEP is displayed in the first column of Table 10.6. Neither the treatment combinations with all low values, 0000000000000, nor the treatment combination with all high values, 111111111113, appears in this OMEP. Then by adding the same set of generators as we used in the previous example, we get the choice sets given in Table 10.6. These choice sets have the same efficiency and C_M matrix as before. Note that in no choice set does one treatment combination dominate another.

Table 10.6 A Different $2^{12} \times 4/16$ and Choice Sets for $m = 3$

$F = 2^{12} \times 4/16 + 1111110000000$	$F + 1111111111111$	$F + 1010101010102$
1111110000000	0000001111111	0101011010102
1111000110111	0000111001002	0101101100013
1110101011012	0001010100103	0100000001110
1110011101103	0001100010010	0100110111001
1001110111013	0110001000100	0011011101111
1001000001102	011011110013	0011101011000
1000101100001	0111010011112	0010000110103
1000011010110	0111100101001	0010110000012
0101111011101	1010000100012	1111010001003
0101001101010	1010110010101	1111100111112
0100100000113	1011011111000	1110001010011
0100010110002	1011101001113	1110111100100
0011111100112	1100000011003	1001010110010
0011001010003	1100110101110	1001100000101
0010100111100	1101011000011	1000001101002
0010010001011	1101101110102	1000111011113

Similarly, if there are some treatment combinations that are unrealistic, it is often possible to avoid them by either using a different starting design or choosing the set(s) of generators so that no unrealistic treatment combinations appear in the choice sets.

The number of choice sets that each respondent can complete will depend on the complexity of the choice sets, including the number of attributes and the number of options in the choice set. Usually, we have each of the respondents complete all of the choice sets, but if the design has more choice sets than a respondent can complete, then the choice sets can be split into blocks (or versions), either randomly or using a spare attribute, if there is one available. In Example 10.7, if 16 choice sets are considered to be too many for the respondents to complete, then it is possible to construct the starting design $F = 2^{12} \times 4//16$ from a $4^5//16$ FFD while retaining one of the four-level attributes to be used as the blocking factor. Then the 16 choice sets could be split into either two blocks of eight choice sets or four blocks of four choice sets.

10.5.2 Extensions

10.5.2.1 Choice Experiments with a None Option

In Section 10.5.1, we have assumed that the respondents must choose one alternative in each choice set. However, there are many situations in which it is more realistic to include a “none of these” alternative in each choice set. King et al. (2007) investigate patient preferences for asthma medications with each choice set consisting of two medications plus a “no medication” option. The third option was included because in practice respondents may choose not to take a medication.

In this section, we consider the results in Street and Burgess (2007) on what happens to the optimality of the designs when we adjoin a none option to each choice set. It turns out that there is a simple relationship between the matrices for a forced choice stated preference experiment and those from the same choice sets with a none option adjoined to each choice set.

We let \mathbf{B}_f be the contrast matrix for the forced choice experiment and it may contain contrasts for main effects and perhaps interaction effects. Let $\Lambda(\boldsymbol{\pi}_0)_f$ be the $\Lambda(\boldsymbol{\pi}_0)$ matrix for the forced choice experiment as defined in Equations (10.3) and (10.4), and let $\mathbf{C}_f = \mathbf{B}_f \Lambda(\boldsymbol{\pi}_0)_f \mathbf{B}'_f$ be the information matrix for the forced choice experiment. We assume that \mathbf{B}_f is $p \times L$ where there are p contrasts of interest, and note that $\Lambda(\boldsymbol{\pi}_0)_f$ will contain rows and columns of 0s if not all treatment combinations appear in the choice experiment. We will use \mathbf{B}_n , $\Lambda(\boldsymbol{\pi}_0)_n$, and $\mathbf{C}_n = \mathbf{B}_n \Lambda(\boldsymbol{\pi}_0)_n \mathbf{B}'_n$ for the corresponding matrices when a none option has been included in each choice set. We assume that each choice set in the forced choice experiment has m options in it and that there are N such choice sets.

Let $d^2 = L(L + 1)$. Then

$$\mathbf{B}_n = \begin{bmatrix} \mathbf{B}_f & \mathbf{0}_p \\ \frac{1}{d} \mathbf{j}'_L & \frac{-L}{d} \end{bmatrix},$$

where $\mathbf{0}_p$ is a $p \times 1$ vector of zeroes and none is the final treatment combination. Thus $\mathbf{B}_n \mathbf{B}'_n = \mathbf{I}_{p+1}$. Let r_i be the number of times the i th treatment combination appears in the stated preference experiment, and let $\mathbf{r} = (r_1, \dots, r_L)'$. Let \mathbf{D} be a matrix with these replication numbers on the diagonal. Then

$$\mathbf{\Lambda}(\boldsymbol{\pi}_0)_n = \frac{1}{(m+1)^2 N} \begin{bmatrix} m^2 N \mathbf{\Lambda}(\boldsymbol{\pi}_0)_f + \mathbf{D} & -\mathbf{r} \\ -\mathbf{r}' & mN \end{bmatrix},$$

and

$$\mathbf{C}_n = \frac{1}{(m+1)^2 N} \begin{bmatrix} m^2 N \mathbf{C}_f + \mathbf{B}_f \mathbf{D} \mathbf{B}'_f & \frac{L+1}{d} \mathbf{B}_f \mathbf{r} \\ \frac{L+1}{d} \mathbf{r}' \mathbf{B}'_f & \frac{mN(L+1)}{L} \end{bmatrix}.$$

This result holds for any stated preference choice experiment in which a none option has been adjoined to each choice set. Obviously, the expression for \mathbf{C}_n is easier to work with if $\mathbf{B}_f \mathbf{r} = \mathbf{0}_p$, and this is true if we assume that all the treatment combinations in the complete factorial appear r times in the choice experiment. Then \mathbf{C}_n can be simplified to

$$\mathbf{C}_n = \frac{1}{(m+1)^2} \begin{bmatrix} m^2 \mathbf{C}_f + \frac{m}{L} \mathbf{I}_p & \mathbf{0}_p \\ \mathbf{0}'_p & \frac{m(L+1)}{L} \end{bmatrix}.$$

Hence, the maximum determinant of the information matrix for estimating main effects when a none option is included in each choice set is given by

$$\det(\mathbf{C}_{\text{opt},M,n}) = \frac{m(L+1)}{L(m+1)^2} \times \prod_{q=1}^k \left[\frac{2\ell_q S_q + m(\ell_q - 1)}{(m+1)^2 (\ell_q - 1)L} \right]^{\ell_q - 1},$$

where S_q is the least upper bound for the sum of the differences for a particular attribute q as defined in Equation (10.7).

Hence, the same designs that are optimal for the estimation of main effects in the forced choice setting are optimal for the estimation of main effects when a none option has been adjoined to each choice set. We can also calculate the effect of estimating just the main effects (i.e., \mathbf{C}_n without the row and column for the none option) relative to the maximum determinant of \mathbf{C}_M for the forced choice setting. We find that the efficiency of the design for the estimation of main effects only is reduced from 100% to

$$\frac{m^2}{(m+1)^2} \left[\prod_{q=1}^k \left(\frac{2\ell_q S_q + m(\ell_q - 1)}{2\ell_q S_q} \right)^{(\ell_q - 1)} \right]^{1/p} \times 100\%.$$

But adjoining a none option does change the properties of the design. If there is no none option, then the optimal design for main effects cannot be used to give any information about two-factor interactions. The inclusion of a none option may make it possible to estimate two-factor interactions as well, although the efficiency may not be very high. The component of the information matrix corresponding to main effects is given by

$$\frac{1}{(m+1)^2 N} \left(m^2 \mathbf{C}_f + \frac{m}{L} \mathbf{I}_p \right),$$

and so correlated effects will remain correlated after the introduction of a none option.

Example 10.8. Suppose that we have $k = 3$ attributes with $\ell_1 = \ell_2 = 2$ and $\ell_3 = 3$. If $m = 2$, the optimal design for estimating main effects in a forced choice setting is obtained by taking the complete factorial and adding the generator (1,1,1) to get the following 12 choice sets:

(000, 111),	(001, 112),	(002, 110),	(010, 101),	(011, 102),	(012, 100),
(100, 011),	(101, 012),	(102, 010),	(110, 001),	(111, 002),	(112, 000).

If we adjoin a none option to each of these choice sets then these 12 choice sets are still optimal for the estimation of main effects, but the component of the information matrix corresponding to main effects has changed. The normalized contrast matrix for the stated preference experiment when the none option is included is given by

$$\mathbf{B}_n = \frac{1}{\sqrt{156}} \begin{bmatrix} \sqrt{156} \mathbf{B}_M & \mathbf{0}'_4 \\ \mathbf{j}'_{12} & -12 \end{bmatrix},$$

where \mathbf{j} is a column vector of ones. For main effects only for the forced choice design, the $\mathbf{C}_{M,f}$ is a diagonal matrix with entries $1/48[4, 4, 3, 3]$. For main effects only, for the “12 plus none” design, $\mathbf{C}_{M,n}$ is a diagonal matrix with entries $1/108[6, 6, 5, 5, 26]$. Thus we see that $\det(\mathbf{C}_{\text{opt},M}) = (1/12)^2 \times (1/16)^2$, and that $\det(\mathbf{C}_{\text{opt},M,n}) = (1/18)^2 \times (5/108)^2 \times (13/54)$. The determinant for the main effects only if the none option is included is $(1/18)^2 \times (5/108)^2$. When we compare the efficiencies of the designs with none adjoined to forced choice designs, we use the determinant for the main effects only (or for main effects plus two-factor interactions only) so that we can see what we gain (or lose) by adjoining the none. In this case, we see that if we adjoin a none option to the 12 choice sets that are optimal for estimating main effects, then the design is 67.2% efficient for estimating main effects plus two-factor interactions. The same set of 12 choice sets is now $((1/18)^2(5/108)^2 / ((1/12)^2(1/16)^2))^{1/4} = 70.3\%$ efficient for estimating main effects only, however.

If $m = 2$, the optimal design for estimating main effects plus two-factor interactions is obtained by taking the complete factorial and adding each of the generators $(0,1,1)$, $(1,0,1)$, and $(1,1,0)$ in turn and has the following 30 choice sets:

(000, 011),	(001, 012),	(002, 010),	(010, 001),	(011, 002),	(012, 000),
(100, 111),	(101, 112),	(102, 110),	(110, 101),	(111, 102),	(112, 100),
(000, 101),	(001, 102),	(002, 100),	(010, 111),	(011, 112),	(012, 110),
(100, 001),	(101, 002),	(102, 000),	(110, 011),	(111, 012),	(112, 010),
(000, 110),	(001, 111),	(002, 112),	(010, 100),	(011, 101),	(012, 102).

In this example, all effects are independently estimated in all of the designs discussed.

Table 10.7 gives the efficiencies of all the designs for estimating both main effects only and main effects plus two-factor interactions relative to the best forced choice experiment and considering only the effects of interest.

In the next example, adjoining a none option does not make the two-factor interaction effects estimable.

Example 10.9. Suppose that we have $k = 3$ attributes with $\ell_1 = \ell_2 = 2$ and $\ell_3 = 4$ levels. Suppose that we use the choice sets given in Table 10.8 to estimate main

Table 10.7 Efficiencies of the Four Designs Discussed in Example 10.8

Design	Estimating	
	Main Effects	ME plus 2fi
12 forced choice	100%	0%
12 plus none	70.3%	67.2%
30 forced choice	69.3%	100%
30 plus none	56.5%	80.4%

Table 10.8 Choice Sets with $k = 3$ Attributes, $\ell_1 = \ell_2 = 2$ and $\ell_3 = 4$

Option 1	Option 2
0 0 0	1 1 1
1 1 0	0 0 1
1 1 1	0 0 2
0 0 1	1 1 2
0 1 2	1 0 3
1 0 2	0 1 3
1 0 3	0 1 0
0 1 3	1 0 0

effects plus two-factor interactions. Then we get a \mathbf{C}_{MT} matrix with $\det(\mathbf{C}_{MT}) = 0$ and so not all effects can be estimated. If we now assume that each choice set has a none option adjoined, then a row and column is adjoined to \mathbf{C}_{MT} for the none option, and we get $\mathbf{C}_{MT,n}$, which also has a determinant of zero. Thus we see that adjoining the none option has not improved the properties of the design.

It should be noted that a binary response experiment is in fact the simplest experiment, which includes a none option in each choice set. In a binary response design, respondents are shown a set of options, in turn, and asked for each option whether or not they would choose it. The options themselves might be actual products for sale, or they might be treatments in a medical setting, for instance. Hall et al. (2002) used a binary response experiment to investigate chickenpox vaccination.

Street and Burgess (2007) have shown that when estimating the main effects only, a resolution III fraction results in a diagonal C matrix with the largest possible determinant among designs with a diagonal C matrix.

Since each treatment combination is shown individually to each respondent the number of choice sets is N , and we can write

$$\Lambda(\boldsymbol{\pi}_0)_n = \frac{1}{4N} \begin{bmatrix} \mathbf{I}_N & \mathbf{0}_{N,L-N} & -\mathbf{j}_N \\ \mathbf{0}_{L-N,N} & \mathbf{0}_{L-N,L-N} & \mathbf{0}_{L-N} \\ -\mathbf{j}'_N & \mathbf{0}'_{L-N} & N \end{bmatrix}.$$

For the information matrix, $\mathbf{C}_n = \mathbf{B}\Lambda_n\mathbf{B}'$, when estimating main effects only, a resolution III, equireplicate fraction gives rise to a diagonal information matrix given by

$$\mathbf{C}_{M,n} = \frac{1}{4N} \begin{bmatrix} \mathbf{I}_p & \mathbf{0}_p \\ \mathbf{0}'_p & \frac{(L+1)N}{L} \end{bmatrix}.$$

with a maximum determinant of

$$\det(\mathbf{C}_{\text{opt},M,n}) = \left(\frac{1}{4N} \right)^p \frac{(L+1)N}{L}.$$

10.5.2.2 DCEs with a Common Base Option

If a particular combination of attribute levels appears in all choice sets, then this treatment combination is called the *common base option*. It may represent the current situation or the current treatment for a particular health condition (examples in Ryan and Hughes 1997, Ryan and Farrar 2000, and Longworth, Ratcliffe, and Boulton 2001), or the common base may be randomly chosen from the main effects plan and have all the other scenarios from the plan compared to it pairwise (examples in Ryan 1999, Ryan et al. 2000, and Scott 2002).

In this section, we discuss the results in Street and Burgess (2007), which give the optimal design when there are two alternatives in each choice set, one of which is the common base option. To ensure that the matrix of contrasts for main effects is unambiguously defined, we assume that the R treatment combinations that appear in the choice experiment form a fractional factorial design of resolution III, which we denote by F . The treatment combinations in F are ordered so that the first treatment combination is the common base, the next $R - 1$ treatment combinations are the other treatment combinations in F , and, as usual, let \mathbf{B}_M be the normalized main effects contrast matrix. Then

$$\Lambda(\boldsymbol{\pi}_0)_c = \frac{1}{4(R-1)} \begin{bmatrix} (R-1) & -1 & -1 & \dots & -1 \\ -1 & 1 & 0 & \dots & 0 \\ -1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ -1 & 0 & 0 & \dots & 1 \\ \mathbf{0}_{L-R,R} & & & & \mathbf{0}_{L-R,L-R} \end{bmatrix}.$$

Then the information matrix, $\mathbf{C}_{M,c} = \mathbf{B}_M \Lambda(\boldsymbol{\pi}_0)_c \mathbf{B}'_M$, is given by

$$\mathbf{C}_{M,c} = \frac{1}{4(R-1)} \left(\frac{R}{L} \mathbf{J}_p + \frac{R}{L} \mathbf{I}_p \right),$$

where $p = \sum_{q=1}^k (\ell_q - 1)$. The optimal design arises from the smallest resolution III design with

$$\det(\mathbf{C}_{\text{opt},M,c}) = \frac{1+p}{(4(R_{\min}-1))^p} \left(\frac{R_{\min}}{L} \right)^p.$$

The efficiency of the design, relative to the optimal forced choice stated preference experiment with choice sets of size 2, is given by $(\det(\mathbf{C}_{M,c}) / \det(\mathbf{C}_{\text{opt},M,m=2}))^{1/p}$.

We can see that from the point of view of statistical efficiency, all resolution III designs with the same number of treatment combinations are equally good for a particular number of attributes with given number of levels. It is also immaterial which of the treatment combinations is used as the common base.

Example 10.10. Consider an experiment where there are four attributes describing each option. Three of the attributes have two levels, and one has four levels. The smallest fractional factorial design of resolution III has eight treatment combinations. Then, using any $2 \times 2 \times 2 \times 4 / 16$ and using any treatment combination as a common base gives a design that is 93.3% efficient relative to a design with eight treatment combinations and 45.17% efficient

relative to a forced choice stated preference design. Using the complete factorial and using any treatment combination as a common base gives a design that is 90.3% efficient relative to a design with eight treatment combinations and 43.71% efficient relative to a forced choice stated preference design. Using a $2 \times 2 \times 2 \times 4 / 8$ gives an efficiency of 48.40% relative to a forced choice stated preference design.

In Scott (2002), a similar experiment was carried out with four attributes (two with two levels, one with three levels, and one with four levels) and using 16 choice sets. In this case, the efficiency depends on the treatment combination chosen to be the common base. If the common base has the level of the three-level attribute, which appears eight times, then the efficiency is 44.49%; and if the common base has either of the other levels of the three-level attribute, then the efficiency is 46.12%.

We can use this approach to estimate main effects plus two-factor interactions by starting with a resolution V fractional factorial design. Unfortunately, in this setting, we can only know that all effects are estimable and compare particular designs. We cannot calculate the efficiency relative to the optimal design, since there is no general expression for the maximum $\det(C)$; see Section 10.5.1.2.

10.5.2.3 DCEs with a Common Base and a None Option

Street and Burgess (2007) have combined the previous results to give an expression for the information matrix for choice sets when there is both a common base and a none option in each choice set. Although in theory the choice sets could be of any size, we will only consider the situation when the choice sets are of size $m = 3$. In this case, each choice set contains the common base, an option unique to that choice set and the “neither of these” option.

In this case, the correct contrast matrix is \mathbf{B}_n , defined in Section 10.5.2.1. By reordering the treatment combinations if necessary, and assuming there are R treatment combinations in the fraction, we can write the $\Lambda(\boldsymbol{\pi}_0)$ matrix, $\Lambda(\boldsymbol{\pi}_0)_{cn}$, as

$$\Lambda(\boldsymbol{\pi}_0)_{cn} = \frac{1}{9(R-1)} \begin{bmatrix} 2(R-1) & -1 & -1 & \dots & -1 & 0 & \dots & 0 & -(R-1) \\ -1 & 2 & 0 & \dots & 0 & 0 & \dots & 0 & -1 \\ -1 & 0 & 2 & \dots & 0 & 0 & \dots & 0 & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & 0 & 0 & \dots & 2 & 0 & \dots & 0 & -1 \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 \\ -(R-1) & -1 & -1 & \dots & -1 & 0 & \dots & 0 & 2(R-1) \end{bmatrix}$$

If we assume that the treatment combinations involved in the experiment form a fractional factorial design, then the information matrix when there is a common base and a none option is \mathbf{C}_{cn} , given by

$$\mathbf{C}_{cn} = \frac{1}{9(R-1)} \begin{bmatrix} \frac{2R}{L} I_p + 2(R-1)\mathbf{b}_1\mathbf{b}'_1 & \frac{(L+1)(R-2)}{d}\mathbf{b}_1 \\ \frac{(L+1)(R-2)}{d}\mathbf{b}'_1 & \frac{2(L+1)(R-1)}{L} \end{bmatrix},$$

where \mathbf{b}_1 is the first column of \mathbf{B}_M .

10.5.2.4 DCEs that Allow Ties

In the previous sections, we have considered choice sets in which the respondents were required to choose one option as their preferred option in each of the choice sets. We now consider the case where respondents are permitted to indicate that two or more options are equally attractive. Davidson (1970) extended the Bradley–Terry model to incorporate ties in paired comparisons, and Bush, Burgess, and Street (2010) have extended this to choice sets of any size.

In a choice set with m alternatives, denoted by $\mathcal{T} = (T_{i_1}, T_{i_2}, \dots, T_{i_m})$, respondents may choose any number of alternatives in the choice set as the most preferred options. Let v be the ties parameter that we wish to estimate in addition to the π_{ij} parameters. We now let the merit of the set of s items T_{i_1}, \dots, T_{i_s} be $v^{\sqrt[s]{\pi_{i_1}\pi_{i_2}\dots\pi_{i_s}}}$, where s can be any value from 2 to m . We let \mathcal{T}_{mer} denote the sum of the merits for each possible decision such that

$$\mathcal{T}_{\text{mer}} = \sum_{a=1}^m \pi_{i_a} + \sum_{x=2}^m \sum_{\{T_{j_1}, \dots, T_{j_x}\} \subseteq \mathcal{T}} v^{\sqrt[x]{\pi_{j_1}\dots\pi_{j_x}}}.$$

Then the probabilities for each decision are given by

$$\begin{aligned} P(T_{i_1} | \mathcal{T}) &= \frac{\pi_{i_1}}{\mathcal{T}_{\text{mer}}}, \\ P(\{T_{i_1}, T_{i_2}\} | \mathcal{T}) &= \frac{v^{\sqrt{2}\sqrt{\pi_{i_1}\pi_{i_2}}}}{\mathcal{T}_{\text{mer}}}, \\ &\vdots \\ P(\{T_{i_1}, T_{i_2}, \dots, T_{i_m}\} | \mathcal{T}) &= \frac{v^{\sqrt[m]{\pi_{i_1}\pi_{i_2}\dots\pi_{i_m}}}}{\mathcal{T}_{\text{mer}}} \end{aligned}$$

The Λ matrix is defined to be

$$\Lambda(\boldsymbol{\pi}, v) = \begin{bmatrix} \Lambda_{\pi\pi}(\boldsymbol{\pi}, v) & \mathbf{0} \\ \mathbf{0} & \Lambda_{vv}(m, v) \end{bmatrix}.$$

Under the null hypothesis of equal merits for each of the items and leaving v unspecified, and recalling that $\Lambda(\pi_0)$ matrix is defined in Equations (10.3) and (10.4), we have

$$\begin{aligned}\Lambda_{\gamma\gamma}(\pi_0, v) &= \left[\frac{m+v \sum_{x=2}^m \left(\frac{m-x}{x(m-1)} \binom{m}{x} \right)}{m+v \sum_{x=2}^m \binom{m}{x}} \right] \Lambda(\pi_0), \\ \Lambda_{vv}(m, v) &= \frac{m \sum_{x=2}^m \binom{m}{x}}{v \left(m+v \sum_{x=2}^m \binom{m}{x} \right)^2}.\end{aligned}$$

Then the information matrix for estimating v and the effects of interest for the attributes is given by

$$\mathbf{C}_{\text{Ties}} = \begin{bmatrix} \mathbf{B}_\gamma \Lambda_{\gamma\gamma}(\pi_0, v) \mathbf{B}'_\gamma & \mathbf{0} \\ \mathbf{0} & \Lambda_{vv}(m, v) \end{bmatrix},$$

where \mathbf{B}_γ represents the contrasts of interest for the attributes, such as \mathbf{B}_M or \mathbf{B}_{MT} , which are defined in Equations (10.5) and (10.8), respectively. Bush, Burgess, and Street (2010) have shown that the design that is D -optimal when ties are not allowed is also D -optimal when estimating the attribute parameters, as well as the ties parameter. The maximum determinant when the ties parameter is included in the model is

$$\det(\mathbf{C}_{\text{opt,Ties}}) = \left(\frac{m+v \sum_{x=2}^m \left(\frac{m-x}{x(m-1)} \binom{m}{x} \right)}{m+v \sum_{x=2}^m \binom{m}{x}} \right)^p \times \Lambda_{vv}(m, v) \times \det(\mathbf{C}_{\text{opt}}), \quad (10.11)$$

where $\det(\mathbf{C}_{\text{opt}})$ is the maximum determinant when there are no ties allowed and will depend on the effects to be estimated for the attributes.

Example 10.11. Consider an experiment with two two-level attributes, and with choice sets of size 3, with ties between alternatives in the choice sets permitted. In each choice set, the respondents can choose just one alternative, two alternatives, or all three alternatives. We have $\ell_1 = \ell_2 = 2$ and $m = 3$, and suppose we wish to estimate the main effects only of the two attributes plus the ties parameter. Since $S_1 = S_2 = (m^2 - 1)/4 = 2$, $\det(\mathbf{C}_{\text{opt},M}) = (2/9)^2$, $\Lambda_{vv}(m, v) = 12/v(3 + 4v)^2$, and

Table 10.9 A DCE with Two Two-Level Attributes

Option 1	Option 2	Option 3
0 0	0 1	1 0
0 1	0 0	1 1
1 0	1 1	0 0
1 1	1 0	0 1

$$\left(\frac{m+v \sum_{x=2}^m \left(\frac{m-x}{x(m-1)} \binom{m}{x} \right)^p}{m+v \sum_{x=2}^m \binom{m}{x}} \right) = \left(\frac{3(4+v)}{4(3+4v)} \right)^2.$$

Then the maximum possible value for the determinant of the information matrix \mathbf{C}_{Ties} for the estimation of main effects of the attributes and ties parameter is $\det(\mathbf{C}_{\text{opt},M,\text{Ties}}) = (4+v)^2/(3v(3+4v)^4)$. The design in Table 10.9 is optimal for estimating main effects and the ties parameter, and this can be checked by calculating $\mathbf{C}_{M,\text{Ties}}$ for this design.

The maximum possible value for the determinant of the information matrix \mathbf{C}_{Ties} for the estimation of main effects of the two attributes plus their two-factor interaction, and ties parameter, using Equation (10.10) for $\det(\mathbf{C}_{\text{opt},MT})$, is $\det(\mathbf{C}_{\text{opt},M,\text{Ties}}) = (4+v)^3/(18v(3+4v)^5)$. By calculating $\mathbf{C}_{MT,\text{Ties}}$, we can easily check that the design in Table 10.9 is also optimal when the main effects plus two-factor interaction of the attributes and the ties parameter are to be estimated.

10.5.2.5 DCEs that Include Position Effects Option

In Section 10.5.1, we assumed that the order of the treatment combinations in the choice experiment is arbitrary, but there are situations in which the order effects are of interest to the researcher. Davidson and Beaver (1977) proposed a modification of the Bradley–Terry model that incorporates the order of presentation of the options in choice sets of size 2. Bush, Street, and Burgess (2010) have extended these results to find optimal designs when position effects are included in the model for any choice set size. Davidson and Beaver (1977) discuss two experiments, one involving judging weights and the other involving taste testing. There was a significant position effect in both experiments. We now assume that choice sets, such as $(T_{i_1}, T_{i_2}, T_{i_3})$ and $(T_{i_2}, T_{i_1}, T_{i_3})$, are distinct choice sets.

The Davidson–Beaver position effects model incorporates position effects by multiplying the merit of item T_i, π_i , by a parameter ψ_a to reflect the effect of the item being presented in position a of the ordered choice set. Thus the probability of choosing the item presented in position a of the ordered choice set $X = (T_{i_1}, T_{i_2}, \dots, T_{i_m})$ is

$$P(T_{i_a} | X) = \frac{\psi_a \pi_{i_a}}{\sum_{b=1}^m \psi_b \pi_{i_b}}.$$

The $\Lambda(\boldsymbol{\pi}, \boldsymbol{\psi})$ matrix, containing the expectations of the second derivatives of the log-likelihood function, is defined to be

$$\Lambda(\boldsymbol{\pi}, \boldsymbol{\psi}) = \begin{bmatrix} \Lambda_{\gamma\gamma}(\boldsymbol{\pi}, \boldsymbol{\psi}) & \Lambda_{\gamma\psi}(\boldsymbol{\pi}, \boldsymbol{\psi}) \\ \Lambda_{\psi\gamma}(\boldsymbol{\pi}, \boldsymbol{\psi}) & \Lambda_{\psi\psi}(\boldsymbol{\pi}, \boldsymbol{\psi}) \end{bmatrix},$$

where under the usual null hypothesis of $\boldsymbol{\pi} = \mathbf{j}_L = \boldsymbol{\pi}_0$, with the entries in $\boldsymbol{\psi}$ remaining unspecified, we have

$$\begin{aligned} \Lambda_{\gamma\gamma}(\boldsymbol{\pi}_0, \boldsymbol{\psi})_{ij} &= -\frac{1}{\Psi_1} \sum_{a=1}^m \sum_{b \neq a} \psi_a \psi_b \lambda_{T_i \text{ in pos } a, T_j \text{ in pos } b}, \\ \Lambda_{\gamma\gamma}(\boldsymbol{\pi}_0, \boldsymbol{\psi})_{ii} &= -\frac{1}{\Psi_1} \sum_{a=1}^m \psi_a \left(\sum_{b=1}^m \psi_b - \psi_a \right) \lambda_{T_i \text{ in pos } a}, \\ \Lambda_{\gamma\psi}(\boldsymbol{\pi}_0, \boldsymbol{\psi})_{ia} &= -\frac{1}{\Psi_1} \sum_{b \neq a} \psi_b (\lambda_{T_i \text{ in pos } a} - \lambda_{T_i \text{ in pos } b}), \\ \Lambda_{\psi\psi}(\boldsymbol{\pi}_0, \boldsymbol{\psi})_{a_1 a_2} &= -\frac{1}{\Psi_1}, \text{ and} \\ \Lambda_{\psi\psi}(\boldsymbol{\pi}_0, \boldsymbol{\psi})_{aa} &= \frac{\left(\sum_{b=1}^m \psi_b \right) - \psi_a}{\psi_a \Psi_1}, \end{aligned}$$

with $\Psi_1 = (\sum_{b=1}^m \psi_b)^2$, $\lambda_{T_i \text{ in pos } a} = \delta_{T_i \text{ in pos } a} \times n_C / N$, and $\lambda_{T_i \text{ in pos } a, T_j \text{ in pos } b} = \lambda_{T_i \text{ in pos } a} \times \delta_{T_j \text{ in pos } b}$. $\Lambda_{\psi\psi}(\boldsymbol{\pi}_0, \boldsymbol{\psi})$ is independent of the design for a given choice set size. The contrast matrix is given by

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_\gamma & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_\psi \end{bmatrix},$$

where \mathbf{B}_γ contains the contrasts for the effects of interest for the attributes, such as \mathbf{B}_M or \mathbf{B}_{MT} , which are defined in Equations (10.5) and (10.8), respectively, and \mathbf{B}_ψ is an $(m-1) \times m$ matrix of contrasts for the position effects.

The D -optimal design for the estimation of main effects of the attributes and the position effects is given by the set of choice sets where for each attribute q , the sum of the differences is equal to S_q (defined in Eq. 10.7). The maximum possible value for the determinant of the information matrix $\mathbf{C}_{\text{opt}, M, P} = \mathbf{B}_{M, P} \Lambda(\boldsymbol{\pi}_0, \boldsymbol{\psi}) \mathbf{B}'_{M, P}$ is

$$\det(\mathbf{C}_{\text{opt}, M, P}) = \prod_{q=1}^k \left[\frac{2S_q \ell_q \Psi_2}{Lm(m-1)\Psi_1(\ell_q - 1)} \right]^{\ell_q - 1} \times \det(\mathbf{C}_{\psi\psi}(\boldsymbol{\pi}_0, \boldsymbol{\psi})),$$

where $(\mathbf{C}_{\psi\psi}(\boldsymbol{\pi}_0, \boldsymbol{\psi})) = \det(\mathbf{B}_{\psi\psi}\boldsymbol{\Lambda}_{\psi\psi}(\boldsymbol{\pi}_0, \boldsymbol{\psi})\mathbf{B}'_{\psi})$, which is a function of the ψ values only, and Ψ_2 as defined in Example 10.12.

Bush, Street, and Burgess (2010) have also shown that when the main effects plus two-factor interactions of the attributes and the position effects are to be estimated, if all the attributes have two levels, then the optimal design consists of all choice sets in which the number of attributes that differ between any pair of treatment combinations in the choice set is $D_q = (k+1)/2$ if k is odd, or $D_q = k/2$ or $D_q = k/2 + 1$, if k is even. The information matrix is $\mathbf{C}_{MT,P} = \mathbf{B}_{MT,P}\boldsymbol{\Lambda}(\boldsymbol{\pi}, \boldsymbol{\psi})\mathbf{B}'_{MT,P}$ and the maximum determinant is

$$\det(\mathbf{C}_{opt,MT,P}) = \begin{cases} \left(\frac{\Psi_2(k+2)}{\Psi_1(k+1)2^k} \right)^{k+k(k-1)/2} \times \det(\mathbf{C}_{\psi\psi}(\boldsymbol{\pi}_0, \boldsymbol{\psi})) & k \text{ even,} \\ \left(\frac{\Psi_2(k+1)}{\Psi_1 k 2^k} \right)^{k+k(k-1)/2} \times \det(\mathbf{C}_{\psi\psi}(\boldsymbol{\pi}_0, \boldsymbol{\psi})) & k \text{ odd.} \end{cases}$$

Example 10.12. Consider the experiment in Example 10.11, with $\ell_1 = \ell_2 = 2$, $m = 3$, and $S_1 = S_2 = (m^2 - 1)/4 = 2$. For $m = 3$, $\Psi_1 = (\sum_{b=1}^m \psi_b)^2 = (\psi_1 + \psi_2 + \psi_3)^2$, $\Psi_2 = \sum_{a=1}^m \sum_{b \neq a} \psi_a \psi_b = 2(\psi_1 \psi_2 + \psi_1 \psi_3 + \psi_2 \psi_3)$, and $\det(\mathbf{C}_{\psi\psi}(\boldsymbol{\pi}_0, \boldsymbol{\psi})) = 1/(3\psi_1 \psi_2 \psi_3 (\psi_1 + \psi_2 + \psi_3))$. The maximum possible value for the determinant of the information matrix $\mathbf{C}_{M,P}$ for the estimation of main effects of the attributes and contrasts of the position effects is

$$\det(\mathbf{C}_{opt,M,P}) = \left(\frac{\Psi_2}{3\Psi_1} \right)^2 \times \det(\mathbf{C}_{\psi\psi}(\boldsymbol{\pi}_0, \boldsymbol{\psi})) = \frac{4(\psi_1 \psi_2 + \psi_1 \psi_3 + \psi_2 \psi_3)^2}{27\psi_1 \psi_2 \psi_3 (\psi_1 + \psi_2 + \psi_3)^5}.$$

The design in Table 10.9 is optimal for estimating main effects and the position effects. This can be checked by calculating $\mathbf{C}_{M,P}$ for this design.

The maximum possible value for the determinant of the information matrix $\mathbf{C}_{M,P}$ for the estimation of main effects of the two attributes plus their two-factor interaction, and contrasts of the position effects is

$$\det(\mathbf{C}_{opt,MT,P}) = \left(\frac{\Psi_2}{3\Psi_1} \right)^3 \times \det(\mathbf{C}_{\psi\psi}(\boldsymbol{\pi}_0, \boldsymbol{\psi})) = \frac{8(\psi_1 \psi_2 + \psi_1 \psi_3 + \psi_2 \psi_3)^3}{81\psi_1 \psi_2 \psi_3 (\psi_1 + \psi_2 + \psi_3)^7}.$$

By calculating $\mathbf{C}_{MT,P}$, we can easily check that the design in Table 10.9 is also optimal when the main effects plus two-factor interaction of the attributes and the position effects are to be estimated.

10.5.3 Alternative-Specific Attributes

Previously, we have been considering choice sets in which the attributes are generic, with all alternatives in each choice set being described by the same set of attributes. In some experiments, alternatives are described not only by

attribute levels but also by a brand name or label. The attributes are nested within each brand, so there can be different numbers of attributes as well as different attributes within each brand, complete with different numbers of levels; see, for example, Louviere, Hensher, and Swait (2000) and Viney, Savage, and Louviere (2005). The following example is based on one in Louviere, Hensher, and Swait (2000) and features two modes of transport (or brands) for a commuter trip to work. Suppose that we are interested in the effect of the attributes on how people travel to work by either train or bus. Each choice set will have a train alternative and a bus alternative, and each alternative is described by the levels of three attributes. Suppose that the three attributes in the train alternative are *cost* (with levels \$2.00 and \$3.00), *frequency* (with levels 5 and 15), and *time* (with levels 10 and 20). Similarly, the three attributes in the bus alternative are *cost* (with levels \$1.20 and \$2.20), *frequency* (with levels 15 and 30), and *time* (with levels 15 and 30). The levels of these two-level attributes are coded as 0 and 1. The choice sets for this example are given in Table 10.10.

Burgess (2010) has established the form of the *D*-optimal designs under the usual null hypothesis when the attributes are alternative specific. A FFD of resolution III or more is optimal when estimating main effects only, and a FFD of resolution V or more is optimal when estimating main effects plus two-factor interactions within each brand. In the example discussed at the beginning of this section, the design in Table 10.10 is optimal for estimating main effects only since the design is a resolution III FFD. If main effects plus two-factor interactions for the attributes within the brands are to be estimated, then we cannot use this design since we cannot estimate any interactions between the attributes within the brands. Designs that are optimal for estimating main effects plus interactions between the attributes within the brands are the resolution V fractional factorial design in 32 runs, and the complete factorial in 64 runs.

Table 10.10 Example of a DCE with Alternative Specific Attributes

Choice Set	Train			Bus		
	Cost	Freq	Time	Cost	Freq	Time
1	0	0	0	0	0	0
2	0	0	1	0	1	1
3	0	1	0	1	1	1
4	0	1	1	1	0	0
5	1	0	0	1	1	0
6	1	0	1	1	0	1
7	1	1	0	0	0	1
8	1	1	1	0	1	0

10.6 USING COMBINATORIAL DESIGNS TO CONSTRUCT DCEs

Many combinatorial designs had their genesis as the solution of practical design problems, usually in the ordinary least squares setting. Possibly the most famous example of this is the introduction of balanced incomplete block designs by Frank Yates to provide designs in which every pair of treatments can be compared with equal precision; see Yates (1935), Yates (1936), Fisher (1940), HK2 (chapters 2 and 3).

As combinatorial designs have nice structural properties, they are a natural tool to use in an attempt to develop DCEs with useful properties. Some of these constructions are summarized in this section.

10.6.1 OAs and BIBDs

Green (1974) suggested that when the goal of the choice experiment is to estimate the main effects of the attributes, the set of items to be used for constructing a BIBD should be the treatment combinations that appear in an orthogonal array of strength 2. The choice sets are then the blocks of the BIBD. (BIBDs are defined in HK2.) One advantage of this approach over the construction outlined above is that the total number of items that are described is typically much smaller, and this can be very important when prototypes have to be constructed.

Example 10.13. Suppose that there are $k = 5$ attributes with $\ell_1 = \ell_2 = \ell_3 = \ell_4 = 2$ and $\ell_5 = 4$, and that we are interested in estimating main effects. We use the blocks of an (8,14,7,4,3) BIBD as the choice sets where the items are the treatment combinations in the OA[8;2,2,2,2,4;2]. The choice sets that result from this construction are given in Table 10.11. These 14 choice sets have $\det(\mathbf{C}) = (-\frac{3}{224})^7$, and hence have an efficiency of 85.7% under the null hypothesis. An optimal design in four choice sets is given in Table 10.12, but it contains 16 distinct treatment combinations, which may be unacceptable if there is a cost associated with making a prototype of each, for example.

Table 10.11 The 14 Choice Sets from a (8,14,7,4,3) BIBD Using Items from an OA[8;2,2,2,2,4;2]

00000	00112	01011	11110	00000	00112	01011	10013
00000	01103	10013	11110	00112	01011	01103	10101
00000	10101	11002	11110	01011	01103	10013	11002
00112	01103	10101	11110	01103	10013	10101	00000
00112	10013	11002	11110	10013	10101	11002	00112
01011	01103	11002	11110	10101	11002	00000	01011
01011	10013	10101	11110	11002	00000	00112	01103

Table 10.12 Four Optimal Choice Sets

00000	11111	00112	11003
01011	10102	01103	10010
10013	01100	10101	01012
11002	00113	11110	00001

10.6.2 A Recursive Construction using DCEs and BIBDs

A natural extension of the construction described in the previous subsection is to construct a DCE and then to use the items in each of the choice sets, in turn, as the items for a BIBD, the blocks of which become the choice sets in another DCE with smaller choice sets. Thus each of the original choice sets gives rise to b smaller choice sets.

This approach can give rise to optimal DCEs (Street and Donovan 2009). For instance, using Kuhfeld (2006), we can find optimal DCEs for $k = p^s$ attributes each with $\ell_q = p^s$ levels, with choice sets of size $m = p^s$, where p is a prime. There is a BIBD with $m = p^s$ treatments, $b = p^{s-t}\phi(s, t, p)$ blocks, replication number $r = \phi(s, t, p)$, block size $u = p^t$, and index $\ell = \phi(s-1, t-1, p)$, where

$$\phi(s, t, p) = \begin{cases} \frac{(p^s - 1)(p^s - p)\dots(p^s - p^{t-1})}{(p^t - 1)(p^t - p)\dots(p^t - p^{t-1})}, & 1 \leq t \leq s \\ 1, & t = 0; \end{cases}$$

see, for example, Street and Street (1987). Taking the DCE from Kuhfeld (2006), and using the items in each choice set in turn to construct the BIBD, we obtain an optimal DCE with up to $k = p^s$ attributes each, with $\ell_q = p^s$ levels with choice sets of size p^t for any prime p and any $1 \leq t \leq s$.

For example, if $p = 3$ then Kuhfeld (2006) gives choice sets of size 9 for nine attributes each with nine levels. Using the items in each choice set to construct a (9,12,4,3,1) BIBD, we end up with an optimal DCE with 108 choice sets of size 3 for the nine attributes.

10.6.3 Using the OA Symbols as Ordered Pairs

Grasshoff et al. (2004b) use the symbols of each column in an orthogonal array to represent a pair of levels for an attribute in a DCE in which the choice sets have size $m = 2$. Since the symbols in the OA represent pairs of levels, we are interested in OAs with the number of levels equal to 3 ($= \binom{3}{2}$), 6 ($= \binom{4}{2}$), 10 ($= \binom{5}{2}$) and so on.

Example 10.14. Suppose that $k = 7$, $\ell_1 = \dots = \ell_6 = 3$ and $\ell_7 = 4$, and that we want to construct a DCE with $m = 2$ and with $N = 18$ choice sets. The original OA[18;3⁶,6;2], the corresponding choice sets, and the 18 pairs constructed

Table 10.13 (a) An OA[18;3⁶;6;2], (b) the 18 Pairs Obtained by Symbol Replacement and (c) 18 Pairs from Section 10.5.1

The OA	Option 1	Option 2	Option 1	Option 2
0 0 0 0 0 0 0	0 0 0 0 0 0 0	1 1 1 1 1 1 1	0 0 0 0 0 0 0	1 1 1 1 1 1 1
0 0 1 1 2 2 1	0 0 0 0 1 1 0	1 1 2 2 2 2 2	0 0 1 1 2 2 1	1 1 2 2 0 0 2
0 1 0 2 2 1 2	0 0 0 1 1 0 0	1 2 1 2 2 2 3	0 1 0 2 2 1 2	1 2 1 0 0 2 3
0 1 2 0 1 2 3	0 0 1 0 0 1 1	1 2 2 1 2 2 2	0 1 2 0 1 2 3	1 2 0 1 2 0 0
0 2 1 2 1 0 4	0 1 0 1 0 0 1	1 2 2 2 2 1 3	0 2 1 2 1 0 0	1 0 2 0 2 1 1
0 2 2 1 0 1 5	0 1 1 0 0 0 2	1 2 2 2 1 2 3	0 2 2 1 0 1 1	1 0 0 2 1 2 2
1 0 0 2 1 2 5	0 0 0 1 0 1 2	2 1 1 2 2 2 3	1 0 0 2 1 2 1	2 1 1 0 2 0 2
1 0 2 0 2 1 4	0 0 1 0 1 0 1	2 1 2 1 2 2 3	1 0 2 0 2 1 0	2 1 0 1 0 2 1
1 1 1 1 1 1 0	0 0 0 0 0 0 0	2 2 2 2 2 2 1	1 1 1 1 1 1 0	2 2 2 2 2 2 1
1 1 2 2 0 0 1	0 0 1 1 0 0 0	2 2 2 2 1 1 2	1 1 2 2 0 0 1	2 2 0 0 1 1 2
1 2 0 1 2 0 3	0 1 0 0 1 0 1	2 2 1 2 2 1 2	1 2 0 1 2 0 3	2 0 1 2 0 1 0
1 2 1 0 0 2 2	0 1 0 0 0 1 0	2 2 2 1 1 2 3	1 2 1 0 0 2 2	2 0 2 1 1 0 3
2 0 1 2 0 1 3	1 0 0 1 0 0 1	2 1 2 2 1 2 2	2 0 1 2 0 1 3	0 1 2 0 1 2 0
2 0 2 1 1 0 2	1 0 1 0 0 0 0	2 1 2 2 2 1 3	2 0 2 1 1 0 2	0 1 0 2 2 1 3
2 1 0 1 0 2 4	1 0 0 0 0 1 1	2 2 1 2 1 2 3	2 1 0 1 0 2 0	0 2 1 2 1 0 1
2 1 1 0 2 0 5	1 0 0 0 1 0 2	2 2 2 1 2 1 3	2 1 1 0 2 0 1	0 2 2 1 0 1 2
2 2 0 0 1 1 1	1 1 0 0 0 0 0	2 2 1 1 2 2 2	2 2 0 0 1 1 1	0 0 1 1 2 2 2
2 2 2 2 2 2 0	1 1 1 1 1 1 0	2 2 2 2 2 2 1	2 2 2 2 2 2 0	0 0 0 0 0 0 1
(a)	(b)		(c)	

using the results in Section 10.5.1 are given in Table 10.13. The first set of pairs are 48.075 % efficient, and the second set of pairs are 98% efficient, in both cases assuming $\pi = j_L$, or, equivalently, that $\beta = 0$.

10.6.4 Using Hadamard Matrices to Construct DCEs

Hadamard matrices are also used in Grasshoff et al. (2004b) to construct paired comparison designs. Let \mathbf{A}_ℓ have as its rows the pairs in an optimal design with $k = 1$ and $m = 2$. Let $\mathbf{H}_{n,k}$ be k columns of a Hadamard matrix of order n , where $k \leq n$. Then the pairs are obtained from the matrix $\mathbf{H}_{n,k} \otimes \mathbf{A}_\ell$. Negating a pair corresponds to reversing the order of the levels within that pair.

Example 10.15. If $\ell = 2 = k$ then $\mathbf{A}_2 = [(0,1)]$ and $\mathbf{H}_{2,2} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$. Thus we get $\mathbf{H}_{2,2} \otimes \mathbf{A}_2 = \begin{bmatrix} (0,1) & (0,1) \\ (0,1) & -(0,1) \end{bmatrix}$. We get the choice set (00, 11) from the first row and the choice set (01, 10) from the second row. These pairs are optimal for the estimation of main effects and give no information about the two-factor interaction effect.

The 12 choice sets of size 2 that you get from this approach when $\ell = 3$ and $k = 4$ are optimal for the estimation of main effects and again give no information about the two-factor interaction effect. The construction of Section 10.5.1 gives nine choice sets that are also optimal for the estimation of main effects.

10.6.5 Partial Profiles

Items that are described by many attributes can be too cognitively demanding for respondents (see Miller 1956 and Maddala, Phillips, and Johnson 2002, for example). The most common way of dealing with this in DCEs is to only allow a subset of attributes to vary between the options in a choice set. There are two strategies routinely employed for the attributes that are not varying between the options in the choice set. The first is to not display those attributes at all. This idea can be found in Green (1974). The other approach is to show the other attributes but fix the levels of those attributes for all the options in each choice set. In its most complete form, the levels of the fixed attributes are given by each of the rows of an OA in turn for each of the choice sets in a DCE. The next example illustrates this idea.

Example 10.16. Suppose that there are $k = 7$ three-level attributes, and only three may vary in each choice set. In Table 10.14, we give a $(7,7,3,3,1)$ BIBD, an optimal set of choice sets of size $m = 2$ for three 3-level attributes, an OA $[9;3^4;2]$, and the nine choice sets that arise from using block (1,2,7) from the BIBD (so attributes 1, 2, and 7 can vary in the DCE), and using run 8 from the OA to define the levels of attributes 3, 4, 5, and 6.

Table 10.14 Choice Sets with Four Unvarying Attributes Constructed from a BIBD, an Optimal DCE, and an OA

(7,3,1) BIBD	Choice Sets		OA	Choice Sets from Block {127} and Run 8	
127	000	111	0000	0021020	1121021
134	011	122	0111	0121021	1221022
156	022	100	0222	0221022	1021020
235	101	212	1012	1021021	2121022
246	112	220	1120	1121022	2221020
367	120	201	1201	1221020	2021021
457	202	010	2021	2021022	0121020
210	201	021	2102	2121020	0221021
221	002	220	2210	2221021	0021022

10.7 BAYESIAN WORK

We have seen that in the MNL model, the variance–covariance matrix for the parameter estimates depends on the unknown parameters. Thus the determination of the optimal design is (probably) dependent on the unknown parameter values (since it is possible for the same design to be optimal across all parameter values, although to date no such situation has been found).

So far in this chapter, we have found the optimal design when $\beta = \mathbf{0}$, and exactly the same ideas could be used to determine the optimal design for any fixed value of β . In the absence of prior information, Ferrini and Scarpa (2007), for example, recommend the use of designs that are optimal when $\beta = \mathbf{0}$.

In a Bayesian setting, we deal with the uncertainty about the parameter values by assuming a prior distribution for the parameter values and evaluating the performance of the design by integrating over that distribution. Assuming a specific parameter value is the same as assuming that the prior distribution has a point mass of 1 at the assumed value.

Thus, there are a number of questions that we now have to address. How do the optimality criteria introduced above change in a Bayesian setting? What prior distributions might be used? How easy is it to do the integrations needed?

The links between the usual optimality criteria and a Bayesian approach are spelt out in Chaloner and Verdinelli (1995). There are similar remarks, but specific to DCEs, in Kessels, Goos, and Vandebroek (2006).

Prior information about parameter values can be available from a pilot study or from managers' expectations, and the prior distribution is typically assumed to be normal (see Sandor and Wedel 2001, for example) or uniform (see Kessels, Goos, and Vandebroek 2004, for example). Kessels et al. (2008) show that it is very easy to construct prior distributions that are internally inconsistent, and they give a three-step "sanity check" for any prior distribution.

The integrations that are required to evaluate the Bayesian optimality criteria do not have a closed form. The algorithmic difficulties and suggested strategies for efficient computation have been investigated by various authors. Bliemer, Rose, and Hess (2008) and Kessels et al. (2009) provide an introduction to the issues.

10.8 BEST-WORST EXPERIMENTS

One generalization of choice experiments asks respondents to indicate both the best item and the worst item from each choice set. Thus, in each choice set, each respondent indicates the pair of items that are most different and which of the pair is better. The idea of choosing from three objects the pair that is most different dates back to Richardson (1938), and was extended by

Torgerson (1958) by asking respondents to select one object from two that they perceived to be most similar to a third object. Modeling for these tasks was developed by Ennis, Mullen, and Fritters (1988) and Ennis et al. (1989). Asking explicitly for best and worst from a set of items, each described by the levels of one attribute, seems to have first been done by Finn and Louviere (1992), and an MNL approach to modeling such data is developed in Marley and Louviere (2005).

Two variants of the best and worst task have been considered. In the first, the items being compared have been extended from the levels of one attribute to correspond to the multiattribute items usually used in DCEs. In the second, the items being compared are the features of one multiattribute item. We discuss the design aspects of each below.

10.8.1 Multiattribute Best–Worst Experiments

These designs have been used recently; see Mueller, Francis, and Lockshin (2009) for an application to the testing of wines, and Dejaeger et al. (2008) for a comparison between a best–worst task and a conventional rating task in comparing food products. The only design results that appear to be available are those in Vermeulen, Goos, and Vandebroek (2010), where a semi-Bayesian design constructed to be optimal for multiattribute best–worst is compared with designs that are optimal for the usual DCE. For the case that they considered ($N = 9$ choice sets, each of size $m = 4$, with $\ell_1 = \ell_2 = \ell_3 = 3$ and $\ell_4 = \ell_5 = 2$), the purpose-specific design did better than the usual DCEs in terms of D -optimality. They do note that the prediction accuracy from using a D -optimal DCE or a D -optimal best-worst design are comparable however.

10.8.2 Attribute-Level Best–Worst Experiments

Designs in which respondents see one item at a time and are asked for the best feature and the worst feature of that item have been called *attribute-level best–worst* choice experiments (Marley, Flynn, and Louviere 2008) or *best–worst scaling experiments* (Flynn et al. 2007). Street and Knox (2010) establish the general form of the information matrix for the estimation of main effects for an attribute-level best–worst choice experiment. They use this general expression, and a prior assumption of no differences between the attribute levels, to show that for the estimation of main effects only, resolution III fractional factorial designs perform as well as the complete factorial, and that for the estimation of main effects plus two-factor interactions, resolution V fractional factorial designs perform as well as the complete factorial. They conduct simulation studies that show that the designs they propose recover a range of non-null prior parameter values well.

10.9 MISCELLANEOUS TOPICS

10.9.1 Other Models

We have focused on the MNL model as it is the model used most often to analyze results from a DCE. There are other models that have been proposed to address various deficiencies of the MNL model. These include the nested logit, the mixed logit, the scale heterogeneity model, and, most recently, the G-MNL model (Fiebig et al. 2010), which nests the mixed logit and scale heterogeneity models. Train (2003) says of the mixed logit model, “It obviates the three limitations of the standard logit by allowing random taste variation, unrestricted substitution patterns, and correlation in unobserved factors over time.” Both the mixed logit and scaled heterogeneity models are discussed more extensively in his book. Some design comparisons are available; see Sandor and Wedel (2002), Carlsson and Martinsson (2003) and Ferrini and Scarpa (2007) for example.

If you are not sure whether or not interaction effects should be included in a model, then of course it makes no sense to design only for a main effects model since usually a design that is optimal for the estimation of main effects does not allow the estimation of interaction effects. A small simulation study that illustrates this point may be found in Yu, Goos, and Vandebroek (2008).

10.9.2 Complete Determination of Optimal Designs

Grasshoff and Schwabe (2008) determine the optimal set of pairs for $k = 2$ binary attributes for all possible values of β , and remark that it does not appear to be possible to find such a characterization for models that involve more than two parameters.

10.9.3 Analyzing Results from a DCE

Many software packages can be used to analyze discrete choice experiments, often by utilizing the link between the DCEs and Cox’s proportional hazards model. Kuhfeld (2005) gives a very complete description, including many worked examples, together with appropriate macros, on how to use SAS to do the analysis. There is a description of using S-PLUS to fit a Cox’s proportional hazards model in Venables and Ripley (2003). Multinomial models and advice about how to fit them may be found in Agresti (2002) and Thompson (2005). Thompson (2005) gives worked examples in both S-PLUS and R. A detailed description of using GLIM to analyze paired and triple comparisons may be found in Critchlow and Fligner (1991). Long and Freese (2006) have an extensive discussion on how to use STATA to analyze MNL models and related models, such as the conditional logit and multinomial probit. Train (2003) discusses discrete choice methods with a focus on the use of estimation by simulation.

Table 10.15 DCE for the Light Bulb Survey

Option 1	Option 2	Option 3	Option 1	Option 2	Option 3
0 1 0 1 1 3	2 3 0 1 1 1	1 2 1 0 0 0	0 2 0 1 1 3	1 3 1 0 0 0	3 1 1 0 0 2
2 1 1 0 0 3	3 2 0 1 1 0	1 0 0 1 1 2	3 1 1 0 1 2	2 0 0 1 0 1	1 3 1 0 1 0
2 1 1 1 1 0	0 3 1 1 1 2	3 2 0 0 0 1	3 0 1 1 0 1	2 3 0 0 1 0	1 2 1 1 0 3
3 3 0 0 1 3	0 0 1 1 0 0	1 1 0 0 1 1	0 1 0 0 0 2	1 2 1 1 1 3	3 0 1 1 1 1
2 2 0 0 0 2	1 1 1 1 1 1	0 0 0 0 0 0	1 0 0 0 1 3	2 1 1 1 0 0	0 3 1 1 0 2
1 3 0 1 0 2	0 2 1 0 1 1	2 0 1 0 1 3	2 0 1 0 0 3	3 1 0 1 1 0	0 2 1 0 0 1
3 3 0 0 0 3	2 2 1 1 1 2	1 1 0 0 0 1	3 0 1 0 1 2	0 1 0 1 0 3	2 3 0 1 0 1
3 2 0 1 0 0	1 0 0 1 0 2	0 3 1 0 1 1	3 3 1 1 0 3	0 0 0 0 1 0	2 2 0 0 1 2

Example 10.17. In this example, we analyze results for 100 respondents for the generic forced choice DCE in Example 10.1 with $\ell_1 = \ell_2 = \ell_6 = 4$, $\ell_3 = \ell_4 = \ell_5 = 2$, and $m = 3$ for estimating main effects only. The aim of the experiment was to investigate the effect of various attributes of energy efficient light bulbs on the decision to purchase such a bulb. The experiment was motivated by recent legislation that will ban the sale of incandescent bulbs in Australia. The design, which is optimal, is shown in Table 10.15. The SAS code and output are given below, but due to limited space, only the data for the first five choice sets for the first respondent are shown. For each option in each choice set for each subject, in the input, we require the levels of each of the six attributes and whether or not the option was chosen ($c = 1$ if the option is chosen, $c = 2$ if the option is not chosen). Therefore, there will be a total of $16 \times 3 \times 100 = 4800$ observations. The levels of the attributes were effects coded, as described in Example 10.2, by PROC transreg. See Kuhfeld (2005) for more details on the SAS code.

The parameter estimates in the SAS output are the estimates of the entries in $\hat{\beta}$, as specified in Example 10.2, with

$$\hat{\beta} = \begin{pmatrix} 0.10328, -0.11032, -0.09791, -0.33903, -0.06407, -0.08512, \\ -0.05473, -0.10545, 0.10535, 0.75065, 0.34754, -0.33322 \end{pmatrix}'.$$

We can use $\hat{\beta}$ to estimate the proportion of respondents who would choose each of the options in each of the choice sets in the DCE. If we consider the sample choice set in Table 10.2 and using the effects coding specified in Example 10.2, we can calculate the probabilities that each of the options in that choice set will be chosen with

$$P(T_{i_1} > T_{i_2}, T_{i_3}) = \frac{\hat{\pi}_{i_1}}{\hat{\pi}_{i_1} + \hat{\pi}_{i_2} + \hat{\pi}_{i_3}}.$$

Hence, the probability that option 1 (010113) is chosen is

$$\frac{\text{Exp}[\hat{\beta}\mathbf{x}_{(0,1,0,1,1,3)}]}{\text{Exp}[\hat{\beta}\mathbf{x}_{(0,1,0,1,1,3)}] + \text{Exp}[\hat{\beta}\mathbf{x}_{(2,3,0,1,1,1)}] + \text{Exp}[\hat{\beta}\mathbf{x}_{(1,2,1,0,0,0)}]} = 0.106.$$

Similarly, the probability that option 2 (230111) is chosen is 0.388, and the probability that option 3 is chosen is 0.506.

We can also estimate probabilities for hypothetical choice sets. For instance, if we had a choice set with four light bulbs that only differed on the cost attribute, then the estimated proportion who would choose the cheapest light bulb is given by

$$\frac{\text{Exp}[0.75065]}{\text{Exp}[0.75065] + \text{Exp}[0.34754] + \text{Exp}[-0.33322] + \text{Exp}[-0.75065 - 0.34754 + 0.33322]} = 0.449,$$

since the terms corresponding to the other levels would cancel out. The estimated proportion who would choose the most expensive light bulb is

$$\frac{\text{Exp}[-0.75065 - 0.34754 + 0.33322]}{\text{Exp}[0.75065] + \text{Exp}[0.34754] + \text{Exp}[-0.33322] + \text{Exp}[-0.75065 - 0.34754 + 0.33322]} = 0.099.$$

Cost is the attribute with the largest difference in preferences, as seen by the magnitude of the parameter estimates. For the three binary attributes, the estimated proportions are all about 0.45 for the less attractive level and all about 0.55 for the more attractive level. Quality of light is also fairly evenly split among its four levels. In a choice set with four light bulbs that only differ on life length, light bulbs with the longest lifetime are preferred by about a third of the sample.

```
/* The full data set can be found at ftp://ftp.wiley.com/
public/sci_tech_med/special_designs */

data lightbulbdata1;
input Subj Set Quality Lifetime Recycling Dimmable
Brightness Cost c @@;
datalines;
1 1 0 1 0 1 1 3 2   1 1 2 3 0 1 1 1 2   1 1 1 2 1 0 0 0 1
1 2 2 1 1 0 0 3 2   1 2 3 2 0 1 1 0 1   1 2 1 0 0 1 1 2 2
1 3 2 1 1 1 1 0 1   1 3 0 3 1 1 1 2 2   1 3 3 2 0 0 0 1 2
1 4 3 3 0 0 1 3 2   1 4 0 0 1 1 0 0 1   1 4 1 1 0 0 1 1 2
1 5 2 2 0 0 0 2 2   1 5 1 1 1 1 1 1 2   1 5 0 0 0 0 0 0 1
      :           :           :
;
```

```

PROC transreg design data=lightbulbdata1 nozeroconstant
norestoremissing;
model class(Quality Lifetime Recycling Dimmable Brightness
Cost/ zero=last effects) /lprefix=0;
output out=coded(drop=_type_ _name_ intercept);
id Subj Set c; run;

PROC phreg data=coded outest=betas;
title2 'Light Bulb DCE';
model c*c(2) = &_trgind / ties=breslow;
strata Subj Set;
run;

```

Light Bulb DCE

The PHREG Procedure

Model Information

Data Set	WORK.CODED
Dependent Variable	c
Censoring Variable	c
Censoring Value(s)	2
Ties Handling	BRESLOW
Number of Observations Read	4800
Number of Observations Used	4800

Summary of Subjects, Sets, and Chosen and Unchosen Alternatives

Pattern	Number of Choices	Number of Alternatives	Chosen Alternatives	Not Chosen
1	1600	3	1	2

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	3515.559	3038.326
AIC	3515.559	3062.326
SBC	3515.559	3126.859

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	477.2333	12	<.0001
Score	444.5916	12	<.0001
Wald	383.6305	12	<.0001

The PHREG Procedure

Multinomial Logit Parameter Estimates					
	DF	Parameter	Standard		
		Estimate	Error	Chi-Square	Pr > ChiSq
0	1	0.10328	0.05150	4.0216	0.0449
1	1	-0.11032	0.05339	4.2692	0.0388
2	1	-0.09791	0.05249	3.4791	0.0621
0	1	-0.33903	0.05764	34.5935	<.0001
1	1	-0.06407	0.05211	1.5117	0.2189
2	1	0.08512	0.05037	2.8562	0.0910
0	1	-0.05473	0.02923	3.5065	0.0611
0	1	-0.10545	0.02985	12.4791	0.0004
0	1	0.10535	0.02922	12.9938	0.0003
0	1	0.75065	0.04606	265.5507	<.0001
1	1	0.34754	0.04678	55.1829	<.0001
2	1	-0.33322	0.05438	37.5483	<.0001

REFERENCES

- Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley.
- Atkinson, A.C., A.N. Donev, and R.D. Tobias (2007). *Optimum Experimental Designs, with SAS*. Oxford: Oxford University Press.
- Bliemer, M.C., J.M. Rose, and S. Hess (2008). Approximation of bayesian efficiency in experimental choice designs. *Journal of Choice Modelling*, **1**, 98–127.
- Bunch, D.S., J.J. Louviere, and D. Anderson (1996). A comparison of experimental design strategies for multinomial logit models: The case of generic attributes. Technical report, University of California, Davis. Available at <http://faculty.gsm.ucdavis.edu/~bunch/>.
- Burgess, L. (2007). Discrete choice experiments [computer software]. Technical report, Department of Mathematical Sciences, University of Technology, Sydney, Available at <http://crsu.science.uts.edu.au/choice/>.
- Burgess, L. (2010). Optimal designs for discrete choice experiments when attributes are alternative-specific. Technical report. University of Technology: Sydney.
- Burgess, L. and D.J. Street (2003). Optimal designs for 2^k choice experiments. *Communications in Statistics—Theory and Methods*, **32**, 2185–2206.
- Burgess, L. and D.J. Street (2005). Optimal designs for asymmetric choice experiments. *Journal of Statistical Inference*, **134**, 288–301.
- Bush, S., L. Burgess, and D.J. Street (2010). Optimal designs for stated choice experiments that incorporate ties. *Journal of Statistical Planning and Inference*, **140**, 1712–1718.
- Bush, S., D.J. Street, and L. Burgess (2010). Optimal designs for stated choice experiments that incorporate position effects. *Communications in Statistics—Theory and Methods*, **140**, 1712–1718.
- Carlsson, F. and P. Martinsson (2003). Design techniques for stated preference methods in health economics. *Health Economics*, **12**, 281–294.

- Chakraborty, G., R. Ettenson, and G. Gaeth (1994). How consumers choose health insurance. *Journal of Health Care Marketing*, **14**, 21–33.
- Chaloner, K. and I. Verdinelli (1995). Bayesian experimental design: A review. *Statistical Science*, **10**, 273–304.
- Critchlow, D.E. and M.A. Fligner (1991). Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on GLIM. *Psychometrika*, **56**, 517–533.
- Davidson, R.R. (1970). On extending the Bradley–Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, **65**, 317–328.
- Davidson, R. and R. Beaver (1977). On extending the Bradley-Terry model to incorporate within-pair order effects. *Biometrics*, **33**, 693–702.
- Dejaeger, S., A. Jorgensen, M. Aaslyung, and W. Bredie (2008). Best-worst scaling: An introduction and initial comparison with monadic rating for preference elicitation with food products. *Food Quality and Preference*, **19**, 579–588.
- Dey, A. (1985). *Orthogonal Fractional Factorial Designs*. New York: Wiley.
- Ennis, D.M., K. Mullen, and J.E.R. Fritters (1988). Variants of the method of triads: Unidimensional Thurstonian models. *British Journal of Mathematical and Statistical Psychology*, **41**, 25–36.
- Ennis, D.M., K. Mullen, J.E.R. Fritters, and J. Tindall (1989). Decision conflicts: Within-trial resampling in Richardson's method of triads. *British Journal of Mathematical and Statistical Psychology Society*, **42**, 265–269.
- Ferrini, S. and R. Scarpa (2007). Designs with a priori information for nonmarket valuation with choice experiments: A Monte Carlo study. *Journal of Environmental Economics and Management*, **53**, 342–336.
- Fiebig, D.G., M.P. Keane, J. Louviere, and N. Wasi (2010). The generalized multinomial logit model: Accounting for scale and coefficient heterogeneity. *Marketing Science*, **29**, 393–421.
- Finn, A. and J.J. Louviere (1992). Determining the appropriate response to evidence of public concern: The case of food safety. *Journal of Public Policy and Marketing*, **11**, 12–25.
- Fisher, R.A. (1940). An examination of the different possible solutions of a problem in incomplete blocks. *Annals of Eugenics*, **10**, 52–75.
- Flynn, T.N., J.J. Louviere, T.J. Peters, and J. Coast (2007). Best-worst scaling: What it can do for health care research and how to do it. *Journal of Health Economics*, **26**, 171–189.
- Grasshoff, U., H. Grossmann, H. Holling, and R. Schwabe (2004a). Optimal paired comparison designs for first order interactions. *Statistics*, **37**, 373–386.
- Grasshoff, U., H. Grossmann, H. Holling, and R. Schwabe (2004b). Optimal designs for main effects in linear paired comparison models. *Journal of Statistical Planning and Inference*, **126**, 361–376.
- Grasshoff, U. and R. Schwabe (2008). Optimal design for the Bradley-Terry paired comparison model. *Statistical Methods and Applications*, **17**, 275–289.
- Green, P. (1974). On the design of choice experiments involving multifactor alternatives. *Journal of Consumer Research*, **1**, 61–68.

- Guttmann, R., R. Castle, and D.G. Fiebig (2009). Use of discrete choice experiments in health economics: An update of the literature. Technical report, University of New South Wales.
- Hall, J., P. Kenny, M. King, J. Louviere, R. Viney, and A. Yeoh (2002). Using stated preference discrete choice modelling to evaluate the introduction of varicella vaccination. *Health Economics*, **11**, 457–465.
- Hanley, N., S. Mourato, and R.E. Wright (2001). Choice modeling approaches: A superior alternative for environmental valuation? *Journal of Economic Surveys*, **15**, 435–462.
- Hedayat, A., N.J.A. Sloane, and J. Stufken (1999). *Orthogonal Arrays: Theory and Applications*. New York: Springer.
- Huber, J. and K. Zwerina (1996). The importance of utility balance in efficient choice designs. *Journal of Marketing Research*, **33**, 307–317.
- Kessels, R., P. Goos, and M. Vandebroek (2004). Comparing algorithms and criteria for designing Bayesian conjoint choice experiments. Technical report, Department of Applied Economics, Katholieke Universiteit Leuven, Belgium.
- Kessels, R., P. Goos, and M. Vandebroek (2006). A comparison of criteria to design efficient choice experiments. *Journal of Marketing Research*, **43**, 409–419.
- Kessels, R., B. Jones, P. Goos, and M. Vandebroek (2008). Recommendations on the use of Bayesian optimal designs for choice experiments. *Quality and Reliability Engineering International*, **24**, 737–744.
- Kessels, R., B. Jones, P. Goos, and M. Vandebroek (2009). An efficient algorithm for constructing Bayesian optimal choice designs. *Journal of Business and Economic Statistics*, **27**, 279–291.
- King, M., J. Hall, E. Lancsar, D. Fiebig, I. Hossain, J. Louviere, H.K. Reddel, and C. Jenkins (2007). Patient preferences for managing asthma: Results from a discrete choice experiment. *Health Economics*, **16**, 703–717.
- Kuhfeld, W.F. (2005). Marketing research methods in SAS. Technical report, SAS Institute. Available at <http://support.sas.com/techsup/technote/ts722.pdf>.
- Kuhfeld, W.F. (2006). Orthogonal arrays. Technical report, SAS Institute. Available at <http://support.sas.com/techsup/technote/ts723.html>.
- Long, J.S. and J. Freese (2006). *Regression Models for Categorical Dependent Variables Using Stata*. College Station, TX: Stata Press.
- Longworth, L., J. Ratcliffe, and M. Boulton (2001). Investigating women's preferences for intrapartum care: Home versus hospital births. *Health and Social Care in the Community*, **9**, 404–413.
- Louviere, J., D. Hensher, and J. Swait (2000). *Stated Choice Methods: Analysis and Application*. Cambridge: Cambridge University Press.
- Maddala, T., K. Phillips, and F. Johnson (2002). Measuring preferences for health care interventions using conjoint analysis: An application to HIV testing. *Health Service Research*, **37**, 1681–1705.
- Marley, A. and J. Louviere (2005). Some probabilistic models of best, worst, and best-worst choices. *Journal of Mathematical Psychology*, **49**, 464–448.
- Marley, A.A.J., T.N. Flynn, and J.J. Louviere (2008). Probabilistic models of set-dependent and attribute-level best-worst choice. *Journal of Mathematical Psychology*, **52**, 281–296.

- Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, **63**, 81–97.
- Mueller, S., I. Francis, and L. Lockshin (2009). Comparison of best-worst and hedonic scaling for the measurement of consumer wine preferences. *Australian Journal of Grape and Wine Research*, **15**, 205–215.
- Pukelsheim, F. (1993). *Optimal Design of Experiments*. New York: Wiley.
- Richardson, M. (1938). Multidimensional psychophysics. *Psychological Bulletin*, **35**, 659–660.
- Ryan, M. (1999). Using conjoint analysis to take account of patient preferences and go beyond health outcomes: An application to in vitro fertilisation. *Social Science and Medicine*, **48**, 535–546.
- Ryan, M. and S. Farrar (2000). Using conjoint analysis to elicit preferences for health care. *British Medical Journal*, **320**, 1530–1533.
- Ryan, M. and K. Gerard (2003). Using discrete choice experiments to value health care programmes: Current practice and future research reflections. *Applied Health Economics and Health Policy*, **2**, 55–64.
- Ryan, M., K. Gerard, and M. Amaya-Amaya (2008). *Using Discrete Choice Experiments to Value Health and Health Care*. Dordrecht: Springer-Verlag.
- Ryan, M. and J. Hughes (1997). Using conjoint analysis to assess women's preferences for miscarriage management. *Health Economics*, **6**, 261–273.
- Ryan, M., E. McIntosh, T. Dean, and P. Old (2000). Trade-offs between location and waiting times in the provision of health care: The case of elective surgery on the Isle of Wight. *Journal of Public Health Medicine*, **22**, 202–210.
- Sandor, Z. and M. Wedel (2001). Designing conjoint choice experiments using managers' prior beliefs. *Journal of Marketing Research*, **38**, 430–444.
- Sandor, Z. and M. Wedel (2002). Profile construction in experimental choice designs for mixed logit models. *Marketing Science*, **21**, 455–475.
- Scott, A. (2002). Identifying and analysing dominant preferences in discrete choice experiments: An application in health care. *Journal of Economic Psychology*, **23**, 383–398.
- Sloane, N.J.A. (2006). A library of orthogonal arrays. Technical report, AT&T Shannon Lab. Available at <http://www.research.att.com/~njas/oadir/>.
- Street, A.P. and D.J. Street (1987). *Combinatorics of Experimental Design*. Oxford: Clarendon Press.
- Street, D.J. and L. Burgess (2007). *The Construction of Optimal Stated Choice Experiments: Theory and Methods*. Hoboken, NJ: Wiley.
- Street, D.J. and D.M. Donovan (2009). Constructing discrete choice experiments using BIBDs. Technical report, University of Technology, Sydney.
- Street, D.J. and S. Knox (2010). Designing for attribute-level best-worst choice experiments. Technical report, University of Technology, Sydney.
- Thompson, L. (2005). An S manual to accompany Agresti's categorical data analysis. Technical report, University of Houston-Clear Lake. Available at <https://home.comcast.net/~lthompson221/#CDA>.
- Torgerson, W. (1958). *Theory and Methods of Scaling*. New York: Wiley.
- Train, K.E. (2003). *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.

- Venables, W.N. and B.D. Ripley (2003). *Modern Applied Statistics with S*. New York: Springer.
- Vermeulen, B., P. Goos, and M. Vandebroek (2010). Obtaining more information from conjoint experiments by best-worst choices. *Computational Statistics and Data Analysis*, **54**, 1426–1433.
- Viney, R., E. Savage, and J. Louviere (2005). Empirical investigation of experimental design properties of discrete choice experiments in health care. *Health Economics*, **14**, 349–362.
- Yates, F. (1935). Complex experiments (with discussion). *Journal of the Royal Statistical Society*, **2**(Suppl), 181–247.
- Yates, F. (1936). Incomplete randomized blocks. *Annals of Eugenics*, **7**, 121–140.
- Yu, J., P. Goos, and M. Vandebroek (2008). Model-robust design of conjoint choice experiments. *Communications in Statistics—Simulation and Computation*, **37**, 1603–1621.

C H A P T E R 11

Computer Experiments

Max D. Morris

11.1 INTRODUCTION

One classical dichotomy of science divides activity into *theoretical* and *experimental* investigations, with interaction between the communities when experimentalists test theories and theoreticians generalize experimental observations. More recently, *computational science* has come to be recognized as a connection between theory and experiment in that it allows more complex descriptions of system theory than were previously possible, and generates numerical values that are often called “data” even though they are produced through entirely artificial means. In fact, the computer models used to produce these data, not only in science but in many other fields, are often so complex that they can only be understood, or used for prediction, through empirical *computer experiments*—designed and controlled studies in which a computer model is the object (or one of the objects) of interest. In recent decades, statisticians and others have considered what it means to regard the output of a complex computer model as data, and have developed appropriate ideas and methods for the design and analysis of computer experiments. This chapter is a brief summary of some topics that have been addressed in this effort.

11.1.1 Models

A *mathematical model* is a representation of a physical system or process. A simple example of a mathematical model is Kepler’s third law, expanded by

Design and Analysis of Experiments: Special Designs and Applications, First Edition. Edited by Klaus Hinkelmann.

© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

Newton, describing the time required for a planet to make one orbit around the sun:

$$P = 2\pi \frac{a^{3/2}}{\sqrt{G(M+m)}}, \quad (11.1)$$

where P is the time per orbit, a is the semi-major axis of the elliptical orbit, m is the mass of the planet, M is the mass of the sun, and G is the *gravitational constant*. Use of Equation (11.1) is straightforward—a matter of evaluating a simple algebraic expression. A mathematical model that requires somewhat more evaluation effort is the following generic expression of the amount or concentration of a substance in a *compartment* (e.g., a drug in the circulatory system of a human) when it is injected over time and is depleted at a rate proportional to the current concentration:

$$\frac{\partial q}{\partial t} = u(t) - kq(t), \quad (11.2)$$

where $q(t)$ is the concentration of the substance at time t , $u(t)$ is the injection rate, and k is a depletion rate constant. This model is not completely specified until an *initial condition*, $q(0) = q_0$, is specified. Unless simplifications are made, evaluation of $q(t)$ for given q_0 , $u(t)$, and k requires iterative numerical methods.

Physical systems of greater complexity generally require mathematical models that are correspondingly more complex for accurate representation. Examples of complex systems of scientific interest include biological species population growth and diversity as a function of resources, material fatigue as a function of stress, and the global climate as a function of greenhouse gas emission.

Mathematical models can be described by their structure; examples include:

- *ordinary differential equations* defined in terms of derivatives with respect to time only, for example, the compartmental model above, and *partial differential equations*, also including derivatives with respect to other quantities, often spatial dimensions;
- *agent-based* models that explicitly represent the actions and interactions of autonomous units in a network, for example, of vehicular traffic;
- *discrete event simulations* based on a chronological sequence of events, for example, of the behavior of complicated queuing systems; and
- *deterministic* models in which all quantities can be expressed with complete certainty, and *stochastic* models where unpredictable random quantities are part of the model.

When a mathematical model is analytically complex, cannot be written in closed form, contains a large number of variables, or contains random elements, the relationships among variables may only be practically understood

by numerical evaluation. For our purposes, we will define a *computer model* as a program written to numerically evaluate one or more of the variables in a mathematical model as a function of the others.

11.1.2 Some Notation

As suggested above, we regard a computer model as a numerical implementation of a function relating some variables called *inputs* to others called *outputs*. Symbolically, say:

$$\mathbf{y} = f(\mathbf{x}, \mathbf{t}, \boldsymbol{\epsilon}), \quad \mathbf{x} = (x_1, x_2, x_3, \dots, x_k)' \in \Delta, \quad (11.3)$$

where \mathbf{y} is a vector of outputs, f is the computer model, \mathbf{x} is a k -dimensional vector of inputs, \mathbf{t} is a vector of model parameters, and $\boldsymbol{\epsilon}$ is a vector of random perturbations that enter into the determination of \mathbf{y} . In some cases, the distinction between model inputs and model parameters can be blurred; the values of both vectors must be specified before the computer model can be executed. Model parameters usually specify the values of entities that are regarded as constant, whether known or not, e.g., the speed of light, chemical kinetics coefficients, or the gravitational constant G in Equation (11.1). In contrast, model inputs specify quantities that might be controlled or subject to uncontrolled variation, e.g., the feed rate of a chemical reactor, or the rate of precipitation over a watershed. While all variables associated with a computer model have meaningful ranges, we emphasize the set of possible input vectors Δ because, in many computer experiments, the collection of model executions is specified by a set of input vectors. Where \mathbf{t} is held constant, it can be regarded as part of the definition of f , and reference to \mathbf{t} can be suppressed. Where the model is deterministic, we can further simplify our notation to focus on input and output vectors:

$$\mathbf{y} = f(\mathbf{x}), \quad \mathbf{x} \in \Delta. \quad (11.4)$$

11.1.3 Computer Experiments

A computer model is generally written because the corresponding mathematical model is so complex that it can only be understood through numerical studies. When these studies involve a series of N model executions, most often with different input vectors \mathbf{x} and sometimes with different parameter vectors \mathbf{t} , we call this a *computer experiment*. Throughout most of this chapter, we focus on computer experiments in which attention is limited to a scalar output y , \mathbf{t} does not change, and for which the computer model is deterministic. (The case of vector-valued outputs \mathbf{y} is discussed in Section 11.6, and model calibration centered on estimation of appropriate model parameter values is described in Section 11.7.2.) For each of a selected set of input vectors, or *experimental design* $D = \{\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3, \dots, \mathbf{x}^N\}$, the corresponding outputs are computed, and inferences about f are made based on these outputs.

Some common differences distinguish many computer experiments from physical experiments. In computer experiments, there is generally no uncontrolled variation in the response vector (except in the case of stochastic models), so there is less or no concern about blocking and replication. Also, the dimension of \mathbf{x} and/or \mathbf{y} are often much larger than their counterparts in physical experiments.

Some particular settings for computer experiments include sensitivity/uncertainty analysis to determine which inputs are the most influential in determining the value of outputs, output optimization, inverse problems where the challenge is to find the value or values of \mathbf{x} that lead to a specified value of \mathbf{y} , and approximation of the computer model.

11.2 SENSITIVITY/UNCERTAINTY ANALYSIS

Sensitivity or uncertainty analysis is generally undertaken to discover which components of \mathbf{x} are most important in determining the value of y . An input can be “important” in a number of different ways.

11.2.1 Descriptive Methods for Local Analysis

Use of terms has never been completely standardized in this field, but the phrase *local sensitivity analysis* has often been used to describe the examination of derivatives of outputs with respect to inputs. Where computer models are too complex to allow local sensitivity analysis based on analytical derivatives, a computer experiment may be undertaken to numerically approximate them. If f is assumed to be approximately linear over a small subset of Δ , first partial derivatives of f with respect to each element of \mathbf{x} can be approximated from data collected using a resolution III or IV two-level fractional factorial design (see HK1, chapter 11). With additional assumptions, such as effect sparsity or known signs of potential effects, sequential plans can be more economical. Examples of these include group screening approaches (Watson 1961) and fully sequential Bayesian strategies (Mitchell and Scott 1987). For deterministic f , any formal basis for uncertainty characterization of the slopes estimated from such an experiment and taken as approximations of output derivatives must be based on Bayesian arguments, since the residuals have no interpretation as random quantities.

Alternatively, if f is assumed to be at least approximately monotonic in each component of \mathbf{x} , correlations between inputs and outputs are sometimes used to characterize the relative importance of each input. These may be estimated by sampling values of \mathbf{x} from Δ , computing the related outputs by executing f , and computing the resulting sample input–output correlations. In order to lessen the effects of any nonlinear aspects of f , correlations are sometimes calculated after rank transformation of outputs, for example, Iman and Conover (1980). Accidental/spurious sample correlation between elements of

\mathbf{x} can introduce ambiguity into this kind of analysis unless N is large relative to k .

Note that correlations between inputs and outputs cannot really be defined unless \mathbf{x} is regarded as a random vector with some distribution over Δ . When such a distribution can be specified, this provides a basis for defining input “importance” in statistical terms. In the following section, we describe an approach to sensitivity/uncertainty analysis that does not (unlike input–output correlation methods) assume any particular structure, such as approximate linearity or monotonicity, between inputs and outputs.

11.2.2 Methods Based on Input Sampling and Conditional Variance

Suppose a distribution has been specified for \mathbf{x} , $\pi(\mathbf{x})$. A scalar-valued output y is then a function of random variables, and so is also random with variance $\text{Var}(y)$. Variance-based sensitivity/uncertainty analysis focuses on determining the elements of \mathbf{x} that are responsible for most of the variability in y . Let $\mathbf{x}_{(i)}$ represent a vector of all inputs except x_i . Using this notation, two commonly used variance-based sensitivity indices are defined as:

- $T_i = E_{(i)} \text{Var}_i(y|\mathbf{x}_{(i)})$, total variation associated with x_i ; and
- $S_i = \text{Var}_i E_{(i)}(y|x_i)$, first-order variation associated with x_i .

Conceptually, T_i is the expected variability of y if $\mathbf{x}_{(i)}$ is fixed at a randomly selected value, and S_i is the expected *reduction* in variability of y if x_i is fixed at a randomly selected value. In many cases, each index is normalized to the unconditional variance of the output, that is, $T_i = T_i/\text{Var}(y)$ and $S_i = S_i/\text{Var}(y)$.

Sensitivity indices are commonly estimated by output values computed from randomly selected input vectors. T_i can be estimated via the following algorithm:

1. Do n times: Randomly pick $\mathbf{x}_{(i)}$.
 - (a) Do m times: Randomly pick x_i from $\pi(x_i|\mathbf{x}_{(i)})$, compute y for each \mathbf{x} , compute the sample variance, an estimate of $\text{Var}_i(y|\mathbf{x}_{(i)})$.
2. Compute the average of n sample variances, an estimate of $E_{(i)} \text{Var}_i(y) = T_i$.

S_i can be estimated via the following algorithm:

1. Obtain an estimate of $\text{Var}(y)$, e.g., based on output values computed from a random sample from $\pi(\mathbf{x})$.
2. Do n times: Randomly pick x_i .
 - (a) Do m times: Randomly pick $x_{(i)}$ from $\pi(x_{(i)}|x_i)$, compute y for each x , compute the sample variance, an estimate of $\text{Var}_{(i)}(y|x_i)$.
3. Compute the average of n sample variances, an estimate of $E_i \text{Var}_{(i)}(y)$.
4. Estimate S_i as $\widehat{\text{Var}}(y)$ minus the expected variance estimated in step 3.

In each of the above algorithms, much of the computational effort is the result of the generality of $\pi(\mathbf{x})$, specifically that $\pi(x_i|\mathbf{x}_{(i)})$ may be different for every $\mathbf{x}_{(i)}$, and that $\pi(\mathbf{x}_{(i)}|x_i)$ may be different for every x_i . The computational burden is substantially lessened when the elements of \mathbf{x} are independent, so that $\pi(\mathbf{x}) = \prod_{i=1}^k \pi_i(x_i)$.

Sobol (1993) and Saltelli (2002) describe how both T_i and S_i can be estimated for all inputs when the inputs are independent. Let \mathbf{x}_i and \mathbf{x}_i^* be two n -element column vectors containing independent realizations of $\pi_i(x_i)$. Define $k+2$ arrays of size $n \times k$ as:

$$\begin{aligned}\mathbf{A}_T &= (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_k), \mathbf{A}_S = (\mathbf{x}_1^*, \mathbf{x}_2^*, \mathbf{x}_3^*, \dots, \mathbf{x}_k^*), \\ \mathbf{A}_i &= (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_{i-1}, \mathbf{x}_i^*, \mathbf{x}_{i+1}, \dots, \mathbf{x}_k), i = 1, 2, 3, \dots, k\end{aligned}\quad (11.5)$$

Interpreting each row of each array as an input vector, compute the corresponding $n(k+2)$ output values y . T_i is estimated as the average (estimating the expectation) of half the squared difference (estimating the variance) of outputs corresponding to the rows of \mathbf{A}_T and \mathbf{A}_i . Total variation for all inputs except x_i taken as a group, $T_{(i)}$, is estimated as the average of half the squared difference of outputs corresponding to the rows of \mathbf{A}_S and \mathbf{A}_i . The unconditional variance $\text{Var}(y)$ can be estimated from the within-array variability of outputs, and S_i is estimated as the difference between estimates of $\text{Var}(y)$ and $T_{(i)}$. Individually, estimates of T_i and S_i are equivalent to those computed using the general algorithm shown above for $m = 2$. Each set of indices (S_i or T_i for $i = 1, 2, 3, \dots, k$) would require $n(k+1)$ executions of the model using the general algorithms. But independence allows estimation of both sets of indices based on $n(k+2)$ runs, because the runs specified by $\mathbf{A}_1 - \mathbf{A}_k$ can be used for each.

McKay (1995) described a sampling method that is even more efficient for estimating first-order sensitivities when total sensitivities are not needed. Again, let \mathbf{x}_i be an n -element column vector containing independent realizations of $\pi_i(x_i)$. Let $p(\mathbf{x})$ denote a vector with elements that are a random permutation of the elements of \mathbf{x} . Construct m arrays of size $n \times k$ as:

$$\mathbf{A}_1 = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_k), \mathbf{A}_i = (\mathbf{x}_1, p_2^i(\mathbf{x}_2), p_3^i(\mathbf{x}_3), \dots, p_k^i(\mathbf{x}_k)), i = 2, 3, 4, \dots, m \quad (11.6)$$

where each random permutation $p_j^i(-)$ is generated independently of the others. $T_{(i)}$ is estimated as the average of sample variances of outputs across corresponding rows of \mathbf{A}_i . $\text{Var}(y)$ can be estimated from the within-array variability of outputs, and S_i estimated as the difference between estimates of $\text{Var}(y)$ and $T_{(i)}$. This sampling plan does not support estimation of T_i , but for main-effect sensitivity, *all* model evaluations are used to estimate *each* index. The idea is that within each group of input vectors containing any fixed value of x_1 , the m values of x_2 form a random sample from $\pi_2(x_2)$, and similarly for

any other pair of inputs. This is not actually true, as can be seen by noting that even when π_2 is absolutely continuous, the probability of a repeated x_2 value is not really zero. When relatively few inputs are important and n is fairly large, the resulting bias in estimating S_i is small. Morris, Moore, and McKay (2008) described a modification of this sampling plan that eliminates estimation bias.

11.2.3 Fourier Amplitude Sensitivity Test

Cukier et al. (1973) suggested a fundamentally different way to estimate main-effect variance sensitivity indices when inputs are independent random variables. Moments of the distribution of y (both conditional and unconditional) require high-dimensional (k) integration. The *Fourier amplitude sensitivity test* (FAST) is based on a reexpression of these moments as one-dimensional integrals using Wyle's (1938) theorem:

$$\int_{\Delta} y^p(\mathbf{x}) \pi(\mathbf{x}) d\mathbf{x} = \lim_{T \rightarrow \infty} \int_{-T}^T y^p(x_1(s), x_2(s), x_3(s), \dots, x_k(s)) ds, \quad (11.7)$$

for appropriately chosen functions:

$$x_i(s) = G_i(\sin(\omega_i s)), \quad (11.8)$$

where G_i is determined by the marginal distribution π_i , and the integer parameters ω_i are chosen to be *incommensurate* to avoid aliasing at low Fourier frequencies. When ω_i are selected to be positive integers, $T \rightarrow \infty$ is not needed, and an experimental design is comprised of the input vectors generated by a uniform grid of s -values between $-\pi$ and $+\pi$. The *fast Fourier transform* of the vector of computed output values is calculated, and the sum of squared Fourier coefficients indexed by frequencies associated with x_i is taken as the estimate of S_i .

The fundamental difference between FAST and the random sampling techniques discussed above is the implicit assumption in FAST that f is a smooth function of \mathbf{x} , and that the required Fourier expression is an adequate approximation of the model. When these assumptions are justified, FAST can be expected to work well because it explicitly takes both \mathbf{x} and y values into account. In contrast, the sampling approaches of both Sobol and McKay do not depend on the (randomly drawn) values of \mathbf{x} , resulting in a loss of efficiency. However, the sampling approaches are valid when the assumptions upon which FAST is based do not hold.

11.3 GAUSSIAN STOCHASTIC PROCESS MODELS

Traditionally, many of the models used in statistical analysis are *data smoothers* of relatively simple form. For example, polynomial regression is often used with the understanding that the polynomial model may not reflect the actual form of the function linking expected response to controlled variables, but

with the hope that this inaccuracy in the mean function is minor compared with the random noise in the data, and with an emphasis on understanding how this noise affects uncertainty in the modeling effort. In analyzing the output of deterministic computer models, the absence of noise (in the sense of variation in output values corresponding to repeated executions of f with the same input vector) makes this logic questionable. In recent years, methods based on “spatial” *Gaussian stochastic process* (GaSP) models have become popular because they simultaneously provide a means of interpolating noiseless data and a coherent approach for expressing uncertainty of the output at input vectors where the computer model has not been executed.

11.3.1 Model Structure

In this context, the output of a computer model indexed by any input vector \mathbf{x} is regarded as a Gaussian (normal) random variable. The GaSP model specifies the *a priori* characteristics of this (generally infinite) collection of random variables through:

$$\begin{aligned} \text{a mean function : } & E(y(\mathbf{x})) = \mu(\mathbf{x}) \\ \text{a variance function : } & \text{Var}(y(\mathbf{x})) = \sigma^2(\mathbf{x}) \\ \text{a correlation function : } & R(\mathbf{x}, \mathbf{x}^*) = \text{Corr}(y(\mathbf{x}), y(\mathbf{x}^*)), \end{aligned} \quad (11.9)$$

for any \mathbf{x} and \mathbf{x}^* taken from Δ . For any GaSP, certain restrictions on these functions are necessary. For example, $\sigma^2(\mathbf{x})$ must be positive, and $R(\mathbf{x}, \mathbf{x}^*) = R(\mathbf{x}^*, \mathbf{x})$ must be such as to generate a positive semi-definite correlation matrix for any finite set of points taken from Δ . Any or all of these functions can be specified up to some unknown set of parameters, and the analysis is then developed to accommodate the associated uncertainty.

Suppose we have data (output values) from a set of N executions of the model defined by the vectors specified in an experimental design:

$$D = \{\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3, \dots, \mathbf{x}^N\}, \quad (11.10)$$

and denote the N -vector of scalar outputs by \mathbf{y}_D . Let $\boldsymbol{\mu}_D$ be the vector of *a priori* means, and $\boldsymbol{\Sigma}_{DD}$ be the *a priori* covariance matrix of \mathbf{y}_D as determined by functions μ , σ^2 , and R . For any other input vector \mathbf{x}^0 (not in D), let $\boldsymbol{\sigma}_{0D}$ be the N -vector of *a priori* covariances between responses at \mathbf{x}^0 and at each point in D . Then for a fully characterized GaSP, the conditional distribution of $y(\mathbf{x}^0)$ given \mathbf{y}_D is Gaussian, and is characterized by:

$$\begin{aligned} E(y(\mathbf{x}^0) | \mathbf{y}_D) &= \mu(\mathbf{x}^0) - \boldsymbol{\sigma}'_{0D} \boldsymbol{\Sigma}_{DD}^{-1} (\mathbf{y}_D - \boldsymbol{\mu}_D) \\ \text{Var}(y(\mathbf{x}^0) | \mathbf{y}_D) &= \sigma^2(\mathbf{x}^0) - \boldsymbol{\sigma}'_{0D} \boldsymbol{\Sigma}_{DD}^{-1} \boldsymbol{\sigma}_{0D}. \end{aligned} \quad (11.11)$$

In practice, the model is often simplified to a *stationary* form in which the functions that define process moments are invariant with respect to location

in Δ . Abusing notation slightly, this can be expressed by requiring $\mu(\mathbf{x}) = \mu$, $\sigma^2(\mathbf{x}) = \sigma^2$, and $R(\mathbf{x}, \mathbf{x}^*) = R(\mathbf{x} - \mathbf{x}^*)$; that is, the mean and variance are the same at each location, and the correlation between responses at any two locations is a function only of the difference between the corresponding input vectors. While not required, the functional forms for correlation functions most often used in this context decrease as the difference between input vectors grows, but remain non-negative for any pair of vectors. For example, $R(\mathbf{x} - \mathbf{x}^*) = e^{-\sum_{i=1}^k (\mathbf{x}_i - \mathbf{x}_i^*)^2}$, while positive for any arguments, decreases as the Euclidean distance between points \mathbf{x} and \mathbf{x}^* increases. For specified parameter values, $E(y(\mathbf{x}^0)|\mathbf{y}_D) = \hat{y}(\mathbf{x}^0)$ is the best (minimum variance) linear unbiased predictor (BLUP) of the unobserved output at \mathbf{x}^0 . From a Bayesian perspective, it is the minimum expected squared-error predictor.

In computer experiments, the dimension of Δ is often much larger than 2 or 3, with the result that selection of a correlation function R cannot be as firmly based on data examination as is common in the application of spatial processes to, for example, environmental prediction problems. Some knowledge of the analytical behavior of y as a function of \mathbf{x} can be helpful in selecting an appropriate correlation function, for example:

- Stationary processes that are (mean square) continuous, i.e., those that are most appropriate when it is assumed that $f(\mathbf{x})$ is a continuous function, require $R(\boldsymbol{\delta}) \rightarrow 1$ as $\boldsymbol{\delta} \rightarrow \mathbf{0}$. (Here, $\boldsymbol{\delta}$ denotes a difference vector, $\mathbf{x} - \mathbf{x}^*$.)
- Stationary processes that are d -times (mean square) differentiable at all \mathbf{x} require that R be $2d$ -times differential at $\mathbf{0}$.
- Because $\hat{y}(\mathbf{x})$ is a linear combination of $R(\mathbf{x} - \mathbf{x}^i)$, $i = 1, 2, 3, \dots, N$ (apart from σ^2 , these are the elements of $\boldsymbol{\sigma}_{0D}$ in Eq. 11.11), \hat{y} has the same number of derivatives with respect to \mathbf{x} as does R .

While a number of spatial correlation functions are available for modeling data in one or a few dimensions, many of these do not generalize easily to higher dimension. As a result, the *product correlation form*:

$$R(\boldsymbol{\delta}) = \prod_{i=1}^k R_i(\delta_i),$$

where δ_i is the i th element of $\boldsymbol{\delta}$, is often used in computer experiments. So long as the individual R_i are valid correlation functions of one-dimensional arguments, the product is positive-semidefinite regardless of the value of k .

Example 11.1. As a simple example of how a GaSP can be used to model data, consider a small data set of $N = 5$ observed values, each indexed by a $k =$ two-dimensional vector x taken from $\Delta = (0, 1)^2$:

(x_1, x_2)	(0.2, 0.2)	(0.2, 0.8)	(0.8, 0.2)	(0.8, 0.8)	(0.5, 0.5)
y	9.0	9.0	9.0	12.0	10.0

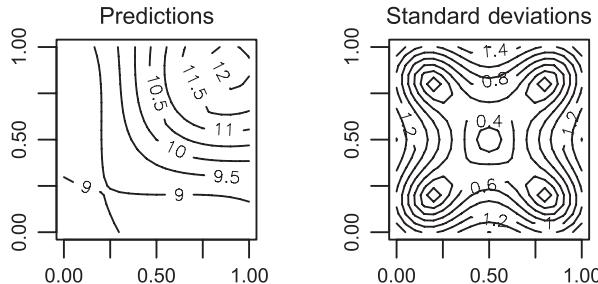


Figure 11.1 Conditional mean and standard deviation for Example 4.1.

Suppose we take as parameter values $\mu = 10$ and $\sigma = 3$, and take

$$R(\delta) = e^{-3(\delta_1^2 + \delta_2^2)}. \quad (11.12)$$

(The more practical case of inferences with unknown process parameters is discussed in Section 11.4.) Then, conditioned on the five data values given, the point prediction of output $\hat{y}(\mathbf{x})$ and the predictive standard deviation $\sqrt{\text{Var}(y(\mathbf{x}) | \mathbf{y}_D)}$ for $\mathbf{x} \in \Delta$ (Eq. 11.11) are displayed in Figure 11.1.

Note that \hat{y} is equal to the output at the five values of \mathbf{x} in D , i.e., the predictions interpolate the observed values of y , and the predictive standard deviations at these five points are zero.

11.3.2 Accommodating Random Noise

Some computer models simulate both the theoretically known or hypothesized deterministic structure of the modeled system and the random deviations from that structure associated with incompletely known processes and influences that are external to the system of interest. When stochastic simulation is used in a computer model, the output is not deterministic; multiple runs of the model using one input vector result in output values that differ due to the random quantities involved in the calculation. In this case, the practical interest is usually in modeling or predicting what the noiseless output would be, that is, the average output from many runs at the same value of \mathbf{x} or the result of a single simulation run long enough so that the noise component of the output is negligible. As in more traditional applications of statistics to data that include sampling or measurement error, producing model predictions that

interpolate \mathbf{y}_D is not the best approach, but retaining the flexibility of the GaSP model to mimic the potentially complicated behavior of the output is still desirable.

Such random noise can be accommodated in the GaSP model through use of what is often called a covariance *nugget*. Suppose the data we observe, \mathbf{y}_D , are actually “noisy” versions of the values we wish to predict, $\tilde{\mathbf{y}}_D$:

$$\mathbf{y}_D = \tilde{\mathbf{y}}_D + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma_{MC}^2 \mathbf{I}),$$

where $\tilde{\mathbf{y}}_D$ and $\boldsymbol{\epsilon}$ are independent, MC stands for *Monte Carlo*, and we wish to predict the noiseless $\tilde{y}(\mathbf{x}^0)$ for any $\mathbf{x}^0 \in \Delta$. Then prediction of \tilde{y} based on the conditional/posterior process takes the same form as in Equation (11.11) with appropriate modification of $\text{Var}(\mathbf{y}_D)$:

$$E(\tilde{y}(\mathbf{x}^0) | \mathbf{y}_D) = \hat{y}(\mathbf{x}^0) = \mu + \boldsymbol{\sigma}'_{0D} [\boldsymbol{\Sigma}_{DD} + \sigma_{MC}^2 \mathbf{I}]^{-1} (\mathbf{y}_D - \mu \mathbf{1}).$$

$$\text{Var}(\tilde{y}(\mathbf{x}^0) | \mathbf{y}_D) = \sigma^2 - \boldsymbol{\sigma}'_{0D} [\boldsymbol{\Sigma}_{DD} + \sigma_{MC}^2 \mathbf{I}]^{-1} \boldsymbol{\sigma}_{0D}$$

When $\sigma_{MC}^2 > 0$, the result is that $\hat{y}(\mathbf{x})$ smooths, rather than interpolates, the observed output values, and conditional variances at design points are relatively small, but are not zero as they are when a nuggetless model is used.

Stein (1999) presents a more detailed discussion of the properties of GaSP models and their use in spatial applications.

11.4 INFERENCE

Section 11.3 outlines the application of a fully characterized stationary GaSP for function prediction. More commonly, a functional form may be assumed for R only up to the values of unknown parameters θ , and μ and σ^2 are also regarded as unknown. In this case, inference about these parameters must be derived from \mathbf{y}_D . Common approaches to this are substitution of maximum likelihood estimates for the unknown values, and a more complete Bayesian approach.

11.4.1 Maximum Likelihood Parameter Estimation

One common practice, sometimes described as an *empirical Bayesian* method, is to estimate the process parameters using data \mathbf{y}_D , usually via the method of maximum likelihood. The log-likelihood function following from the Gaussian model is:

$$L = -\frac{1}{2} (N \ln(2\pi) + N \ln |\mathbf{R}_{DD}| + (\mathbf{y}_D - \mu \mathbf{1})' \mathbf{R}_{DD}^{-1} (\mathbf{y}_D - \mu \mathbf{1}) / \sigma^2), \quad (11.13)$$

where \mathbf{R}_{DD} is the $N \times N$ *a priori* correlation matrix for responses at points in D , with (i,j) element $R(\mathbf{x}^i - \mathbf{x}^j)$. The correlation function R is parameterized by

a positive vector $\boldsymbol{\theta}$; for simplicity, we limit attention to the product correlation form that associates one element of $\boldsymbol{\theta}$ with each dimension in Δ , $R(\boldsymbol{\delta}) = \prod_{i=1}^k R_i(\delta_i; \theta_i)$, and assume that each $R_i(\delta_i; \theta_i)$ is written so that for fixed nonzero δ_i , R_i decreases to or asymptotically toward zero as θ_i increases. Correlation parameters $\boldsymbol{\theta}$ influence L only through \mathbf{R}_{DD} . Given a particular value of $\boldsymbol{\theta}$, maximum likelihood estimates (MLEs) of μ and σ^2 can be derived in closed form:

$$\hat{\mu} | \boldsymbol{\theta} = \mathbf{1}' \mathbf{R}_{DD}^{-1} \mathbf{y}_D / \mathbf{1}' \mathbf{R}_{DD}^{-1} \mathbf{1}, \quad \hat{\sigma}^2 | \boldsymbol{\theta} + (\mathbf{y}_D - \hat{\mu} \mathbf{1})' \mathbf{R}_{DD}^{-1} (\mathbf{y}_D - \hat{\mu} \mathbf{1}) / N. \quad (11.14)$$

So the aspect of maximum likelihood estimation that is numerically most demanding is the search over values of $\boldsymbol{\theta}$ for the maximizer of $L, \hat{\boldsymbol{\theta}}$.

In this application, MLEs do not have the asymptotic properties found with independent data because of the *infill* phenomenon: As more design points are added to D , the distances between pairs of design points become smaller, and *a priori* correlation between output values at these points increases. In fact, MLEs are not even *consistent* in many cases. Regardless of this, for smooth f and suitably chosen R , $\hat{y}(\mathbf{x}) \rightarrow y(\mathbf{x})$ for large samples distributed evenly throughout Δ .

11.4.2 Numerical Issues

The correlation structure of a GaSP model can lead to likelihood surfaces that present numerical difficulties in inference. These surfaces can be multi-modal and/or near-flat, with the result that parameter MLEs can be challenging to compute. In addition, numerical difficulties can be encountered when some pairs of points in D are such that each element of $\mathbf{x}_i - \mathbf{x}_j$ is relatively close to zero, because this can cause the matrix \mathbf{R}_{DD} to be ill-conditioned (large ratio of maximum to minimum eigenvalues), leading to increased numerical error in calculated values of \hat{y} and its predictive standard error, and in extreme cases can produce a *numerically singular* correlation matrix. Finally, even when \mathbf{R}_{DD} is not ill-conditioned, the computational burden of calculating its determinant and inverse is of order N^3 , making full likelihood maximization prohibitive for large data sets. In the following sections, we briefly review one approach that has been proposed for dealing with each of these numerical problems.

11.4.2.1 Penalized Likelihood

Li and Sudjianto (2005) suggested adding a penalty term to the log-likelihood function in order to reduce convergence (and other) problems associated with flat likelihoods. They substitute:

$$Q = -\frac{1}{2} (N \ln(2\pi) + N \ln |\mathbf{R}_{DD}| + (\mathbf{y}_D - \mu \mathbf{1})' \mathbf{R}_{DD}^{-1} (\mathbf{y}_D - \mu \mathbf{1}) / \sigma^2) - N \sum_{i=1}^k p_\lambda(\theta_i) \quad (11.15)$$

for L (Eq. 11.13), where p_λ is a nonnegative, nondecreasing function added to penalize large values of θ_i . Note that because the added term does not involve μ or σ^2 , for given θ , the values of μ and σ^2 that maximize L are the same as those that maximize Q . A particular penalty function advocated by Li and Sudjianto is the *smoothly clipped absolute deviation* penalty of Fan and Li (2001):

$$p_\lambda(\theta) = \begin{cases} \lambda\theta & \theta \leq \lambda \quad (\text{linear with positive slope}) \\ \frac{1}{a-1} \left[a\lambda\theta - \frac{1}{2}\theta^2 \right] & \lambda < \theta \leq a\lambda \quad (\text{concave quadratic}) \\ \frac{1}{2} \frac{a^2 \lambda^2}{a-1} & a\lambda < \theta \quad (\text{constant}). \end{cases}$$

where a is selected, and a value of λ is determined by a cross-validation process. Generally, estimates of θ that maximize Q are smaller than their corresponding MLEs, and estimates of σ^2 are often larger. Together, these two effects often have relatively little impact (i.e., they “cancel out”) on output predictions and their predictive standard errors.

11.4.2.2 Sum of Processes

In order to avoid numerical difficulties associated with ill-conditioned correlation matrices, Booker (2000) suggested modeling output using the sum of two GaSP models with substantially different correlation length scales. The context of his work was function optimization, with the goal of finding one or more input vectors \mathbf{x} that lead to the largest output or function of multiple outputs. This is done through a sequential procedure in which the value of \mathbf{x} that leads to the maximum *predicted* output at one stage is evaluated with the computer model and the results are added to the data set at the next stage. With such a scheme, data points tend to “pile up” in regions where model output is large, leading to near-singular \mathbf{R}_{DD} matrices.

Booker’s solution to this problem is to model y with a single GaSP as described above in the early stages of the investigation; for clarity, in this context, label the parameters of this process as μ_1 , σ_1^2 , and θ_1 . When point-clustering results in ill-conditioned \mathbf{R}_{DD} , the model for y is changed to a sum of the first model, and another mean-zero stationary GaSP with parameters σ_2^2 , and θ_2 for which the elements of θ_2 are large relative to their counterparts in θ_1 . Estimates of the first model are “frozen” at this point, while those of the second model continue to be selected by maximum likelihood with enforcement of the difference in length-scales.

As a practical matter, this helps because, for relatively large δ , correlations are modeled essentially as they would be under the first GaSP alone. For smaller δ , a linear combination of the two correlation functions decreases with distance more quickly than does the first-stage correlation, weakening near-collinearities among the columns/rows of \mathbf{R}_{DD} . Although this may at first

appear to be much like application of a *nugget* as described in Section 11.3.2, this is not really the case—predicted values interpolate observed data, but with some local roughness in the predictive surface due to the short correlation scale of the second GaSP. In some cases, this can also be thought of as an approach for modeling outputs subject to local numerical error accumulated in the execution of the computer model.

11.4.2.3 Tapered Covariance Functions

Most correlation functions used in modeling computer codes decrease with increasing distance between points, but remain positive for all distances. This results in a *dense* correlation matrix \mathbf{R}_{DD} , that is, one that has no elements that are *structurally zero*. Because the computational complexity of eigenanalysis of dense \mathbf{R}_{DD} grows much more quickly than N , GaSP modeling for large data sets can be prohibitive. In contrast, special numerical algorithms have been developed that substantially accelerate matrix computations for *sparse* matrices—those with a large number of elements equal to zero—relative to requirements for dense matrices of the same size. A tempting approach taking advantage of these algorithms in GaSP modeling might be to simply redefine the correlation function R to be zero for vector arguments (differences in input vectors) greater than some critical length chosen so that many pairs of points in D are separated by more than this critical distance. This may seem a very minor alteration, since the correlation functions most often used in computer experiments assign relatively small correlations to pairs of input vectors separated by relatively large distances in Δ . However, the result of this modification is generally *not* a covariance function—specifically, it is not positive semidefinite—and so leads to specification of a structure that is not really a stochastic process. This is not simply a theoretical concern; negative “variances” and other nonsensical results often follow when the predictive formulae are used after such seemingly minor modifications.

Furrer, Genton, and Nychka (2006) describe how sparse \mathbf{R}_{DD} matrices *can* be correctly produced through the use of *tapered* covariance functions. Suppose R is the correlation function that would, apart from numerical concerns, be used. Let T be another correlation function that is zero for large δ . Then the tapered correlation function is $R_T(\delta) = R(\delta)T(\delta)$. While this is a simple adjustment, it is a fundamental change in the GaSP being used. Intuition suggests that if tapering can be done within a radius that includes the data that are most useful for making any given prediction, but excludes enough of the $N(N - 1)/2$ design point pairs to yield a correlation matrix \mathbf{R}_{DD} with a relatively large number of zero elements, the speed of calculation may be dramatically improved at only a minor cost in terms of quality of prediction.

11.4.3 Bayesian Approach

A more complete Bayesian approach to predictive analysis based on a GaSP model requires specification of prior distributions for process parameters and

derivation of the associated posterior after knowing \mathbf{y}_D . Development for known $\boldsymbol{\theta}$ is treated, for example, by Zellner (1971). If μ is given the noninformative prior on the real line, the prior density of σ^2 is inversely proportional to its value, and the two are independent, the posterior distribution of $y(\mathbf{x}^0)$ is a scaled-and-shifted t with $N - 1$ degrees of freedom, for which the mean is the same as the BLUP where $\hat{\mu}$ is substituted for μ , and the variance is:

$$\text{Var}(y(\mathbf{x}^0) | \mathbf{y}_D) = \hat{\sigma}^2 \frac{N-1}{N-3} \left(1 - \mathbf{r}'_{0D} \mathbf{R}_{DD}^{-1} \mathbf{r}_{0D} + \frac{(1 - \mathbf{1}' \mathbf{R}_D D^{-1} \mathbf{r}_{0D})^2}{\mathbf{1}' \mathbf{R}_{DD}^{-1} \mathbf{1}} \right), \quad (11.16)$$

where $\mathbf{r}_{0D} = \boldsymbol{\sigma}_{0D}/\sigma^2$, $\hat{\sigma}^2$ is the MLE of σ^2 (Eq. 11.14), and the last term represents uncertainty associated with μ . Handcock and Stein (1993) developed a full Bayesian approach by adding an independent (of μ and σ^2) prior for $\boldsymbol{\theta}$, leading to:

$$p(\boldsymbol{\theta} | \mathbf{y}_D) \propto \frac{p(\boldsymbol{\theta})}{\sqrt{|\mathbf{R}_{DD}| (\mathbf{1}' \mathbf{R}_{DD}^{-1} \mathbf{1}) \hat{\sigma}^{2(N-1)}}}. \quad (11.17)$$

Standard numerical integration over the domain of $\boldsymbol{\theta}$ is tractable for small k , but becomes less practical in higher dimension. A popular alternative is the use of the Metropolis–Hastings sampler (e.g., Marin and Robert 2007), requiring a starting value $\boldsymbol{\theta}_0$, the posterior $p(\boldsymbol{\theta} | \mathbf{y}_D)$, and a *jumping distribution* $q(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1})$.

One interesting extension of the basic methodology described here is the Bayesian treatment of *treed Gaussian process models* by Gramacy and Lee (2008). The focus of this work is to address inadequacies of the (usually) assumed stationary form of the GaSP, which can be problematic when the computer model output exhibits substantially different behavior in different regions of Δ . The methodology of Gramacy and Lee overcomes this difficulty by sequential bifurcation or division of Δ along hyperplanes, each defined by a value of one input, and a separate GaSP used for inference within each element of the resulting partition. Their implementation is accomplished through *reversible jump* Markov chain Monte Carlo sampling (Green 1995), combining inference associated with a tree representation of the partitioning of Δ , and conditionally on this, the Gaussian process structure within each partition.

The following numerical example provides an illustration of the concepts discussed in this section.

Example 11.2. Heebner and Toran (2000) used MODFLOW, a finite-difference three-dimensional groundwater flow model (McDonald and

Harbaugh 1988) to simulate a spray irrigation sewage treatment operation proposed for a site in Pennsylvania. Their goal was to perform a sensitivity analysis to predict which physical parameters would have the greatest influence on certain groundwater flow characteristics at the site. In one exercise, they used 40 runs of the computer model to investigate the effects of four inputs on outputs of interest. Inputs varied were the hydraulic conductivity of the top layer of soil (K_1 , m/day), the irrigation rate (IR , m/day), the ratio of the hydraulic conductivity parameters of two soil layers (K_{rat} , unitless), and a factor that determines the amount of leakage between the two layers (V , unitless). The experimental design they employed is a Latin hypercube (see Section 11.5.3), where values for each of the input variables were evenly spaced over an interval reflecting knowledge of the site. The output we model here is the resulting steady-state outflow of water ($DRAINS$, m³/day). Values selected for the four inputs and the resulting $DRAINS$ values are listed in Table 11.1.

A matrix of plots depicting the pairwise relationships among these five variables is displayed in Figure 11.2. Because the collection of 40 $DRAINS$ values is severely skewed, this variable is represented on a log scale, both in this plot and in the modeling to follow. The figure suggests, for example, that K_1 and IR are perhaps the most influential variables in the determination of the output.

JMP (SAS Institute, Inc.) is one widely available commercial software package that supports fitting a stationary GaSP model via maximum likelihood (Section 11.4.1). (In version 8.0.2, this is easily accomplished through selecting “Analyze > Modeling > Gaussian Process,” and specification of input and output variables, and correlation function form in the resulting window.) Figure 11.3 displays part of the JMP output describing the GaSP fit to the data in Table 11.1, after transforming $DRAINS$ to the log scale, and using the *Gaussian product correlation* function:

$$R(\boldsymbol{\delta}) = e^{-\sum_{i=1}^k \theta_i \delta_i^2}. \quad (11.18)$$

(The correlation can be modified with a nugget term, described in Section 11.3.2, for cases in which the output contains random noise.)

The top panel of the output is a display of the observed response values plotted against predictions made from the remainder of the data set. Deviations from the 45° line are not true cross-validation errors because the model parameters are not re-estimated for each subsample of size $N - 1$, but the plot does give a visual sense of how reliable the fitted model may be. Note that in this case, one output value, the smallest of the 40 by a considerable margin, even on the log scale, is poorly predicted by the remaining 39, and may indicate that further experimentation should be carried out in this region of the input space. The column labeled “Theta” contains MLEs for each of the

Table 11.1 Data from the Groundwater Flow Study of McDonald and Harbaugh (1988)

<i>K₁</i> (m/day)	<i>IR</i> (m/day)	<i>K_{rat}</i> —	<i>V</i> —	<i>DRAINS</i> (m ³ /day)
77.15	0.077	1.42	0.00	763.69
89.85	0.136	0.99	1.18	778.94
61.92	0.095	1.01	0.87	659.15
94.92	0.018	0.94	1.08	514.18
34.00	0.050	0.88	1.59	439.15
59.38	0.032	1.37	1.23	519.50
23.85	0.159	1.32	0.41	1982.3
49.23	0.027	0.70	0.46	395.71
6.08	0.082	1.09	1.49	2755.0
26.38	0.054	0.83	0.36	426.54
74.62	0.009	0.73	1.79	423.74
56.85	0.132	1.29	0.10	802.31
84.77	0.127	0.91	1.54	740.77
97.46	0.104	1.27	1.44	749.39
8.61	0.004	1.40	1.33	124.22
64.46	0.145	0.55	1.85	697.07
18.77	0.154	1.47	1.28	2335.0
79.69	0.064	0.50	0.56	526.04
31.46	0.100	0.65	1.95	706.30
1.00	0.086	1.06	0.26	4671.0
11.15	0.150	0.68	0.51	4846.3
36.54	0.059	1.14	1.64	502.17
28.92	0.091	0.63	0.20	672.49
92.38	0.163	1.17	1.74	846.20
3.54	0.109	1.12	0.31	4797.8
87.31	0.177	1.24	1.13	868.50
82.23	0.045	1.45	1.69	576.83
72.08	0.036	1.35	1.90	561.77
39.08	0.118	0.78	2.00	730.26
100.0	0.000	1.22	1.02	527.44
46.69	0.014	0.86	0.62	365.75
67.00	0.073	1.50	0.92	657.06
16.23	0.141	0.96	0.72	2950.6
69.54	0.041	1.19	0.05	549.14
41.62	0.168	0.76	0.77	1200.2
44.15	0.023	0.60	0.82	351.55
51.77	0.122	0.81	0.67	684.93
54.31	0.172	1.04	0.15	915.63
21.31	0.068	0.58	1.38	600.58
13.69	0.113	0.52	0.97	2861.8

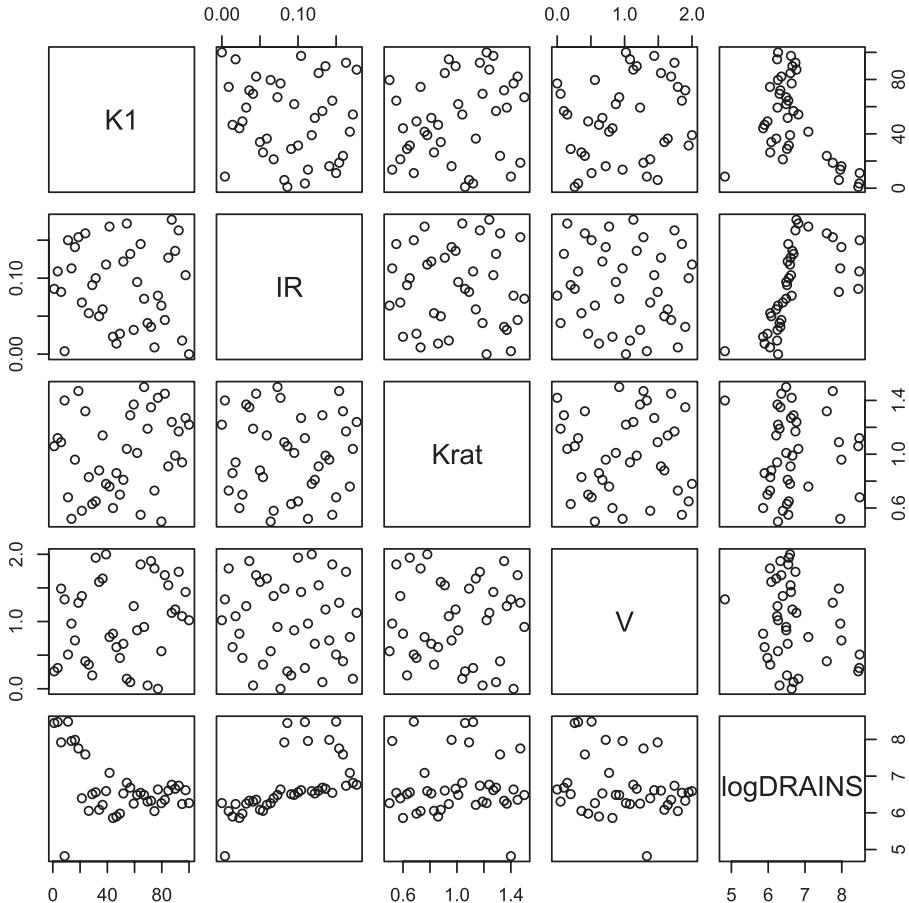


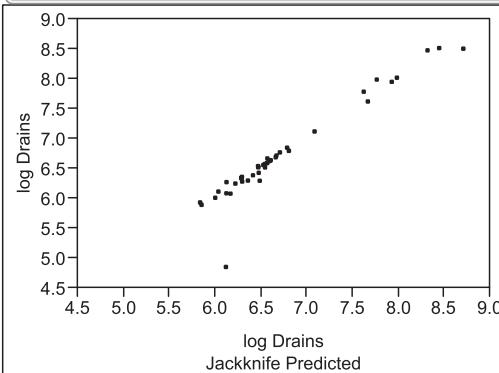
Figure 11.2 Scatter plots of the input and output value for Example 11.2.

four parameters of the correlation function (Eq. 11.18), and MLEs of the process mean and variance appear below the table. As indicated by entries in the “Total Sensitivity” and “Main Effect” columns (related to the indices introduced in Section 11.2.2), K_1 and IR are relatively influential inputs, while K_{rat} and V have little influence on the output.

The bottom panel in this output segment shows a contour plot of \hat{y} , the conditional mean or point predictor of $\log DRAINS$ as a function of K_1 and IR , the most important inputs in this case, with fixed values for K_{rat} and V . This plot shows that $DRAINS$ is an increasing function of IR , but that this relationship is modulated by larger values of K_1 . Again, the corner of this plot in the foreground is heavily influenced by the lowest of the 40 values of $DRAINS$; a

Gaussian Process Model of log Drains

Actual by Predicted Plot



Model Report

Column	Theta	Total	K1	IR	Krat	VC
		Sensitivity	Main Effect	Interaction	Interaction	Interaction
K1	0.0003635	0.5082654	0.2122055		0.2913248	0.0042378
IR	139.30293	0.7730428	0.4791159	0.2913248		0.0017846
Krat	0.1365858	0.0089803	0.0029565	0.0042378	0.0017846	
VC	0.0017332	0.0014343	0.0001184	0.0004972	0.0008174	1.3668e-6

μ σ^2
6.4763284 1.5825966

-2*LogLikelihood
-11.71935

Fit using the Gaussian correlation function.

log Drains

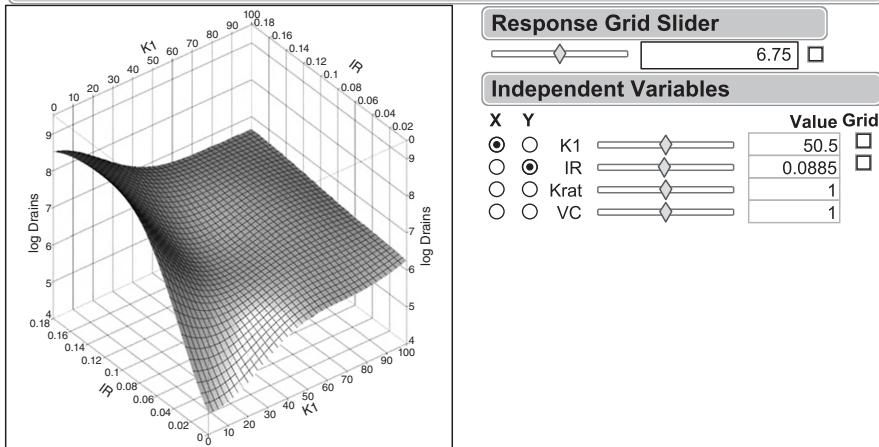


Figure 11.3 JMP output for Example 11.2.

second stage of experimenting focused on relatively small values of both $K1$ and IR would likely improve quality of predictions in this region.

Full Bayesian analysis using GaSP models is more computationally demanding, and at this point in time is often implemented using experimental code written by individual researchers. One concerted effort toward developing software of this type is being undertaken by the MUCM (Management of Uncertainty in Complex Models) research group directed by Professor Tony O'Hagan; the *MUCM Toolkit* is available online at <http://mucm.aston.ac.uk/MUCM/MUCMToolkit>.

11.5 EXPERIMENTAL DESIGNS

A number of approaches to the design of computer experiments have been discussed in the literature. We will briefly describe a few of the most prominent here. Throughout, we assume that the immediate purpose of the experiment is to develop a “good” approximation/prediction of the computer model that can be used throughout Δ .

11.5.1 Model-Based Designs

If a GaSP will be used as the modeling framework for a computer experiment, a statistical design strategy based on an *optimality criterion* can be used to select a design. With standard linear models theory, popular optimality criteria (e.g., D -optimality and E -optimality, see HK2, chapter 1) are not functions of unknown parameters. Here, as in nonlinear regression, the most obvious criteria for optimal designs in the context of GaSP models depend on unknown parameters, particularly θ . As a result, optimal design construction is usually based on a preliminary value of θ , or proceeds sequentially based on updated estimates of θ at each stage (see also Chapter 4).

Sacks, Schiller, and Welch (1989) developed a minimum *integrated mean square error* (IMSE) criterion and design construction algorithm for GaSP models. In essence, integrated mean square predictive error, normalized by σ^2 , can be written as:

$$1 - \text{trace} \left[\begin{pmatrix} 0 & \mathbf{1}' \\ \mathbf{1} & \mathbf{R}_{DD} \end{pmatrix}^{-1} \int_{\Delta} \begin{pmatrix} 1 & \mathbf{r}'_D \\ \mathbf{r}_0 D & \mathbf{r}_0 D \mathbf{r}'_0 D \end{pmatrix} d\mathbf{x}^0 \right], \quad (11.19)$$

where $\mathbf{r}_0 D$ is the vector of correlations σ_{0D}/σ^2 . The integration of matrix elements is simplified when product correlations are used and Δ is a cube, but even so, identifying a design that minimizes this function is computationally demanding. Sacks et al. also concluded through numerical robustness studies that, unless the error in specifying θ is large, there is relatively little loss in

predictive efficiency resulting from constructing an IMSE-optimal design based on incorrect values of the correlation parameters.

Shewry and Wynn (1987) discussed the use of entropy as a design criterion for GaSP models. Entropy is a scalar measure of the uncertainty represented by a uni- or multivariate distribution. For an absolutely continuous random vector \mathbf{z} with probability density function π , entropy is given as

$$\text{Ent} = - \int_{\mathbf{z}} \ln(\pi(\mathbf{z}))\pi(\mathbf{z})d\mathbf{z}.$$

For multivariate normal \mathbf{z} , this reduces to

$$\text{Ent} = c + \ln |\text{Var}(\mathbf{z})|.$$

In our case, this means we want to *minimize* the determinant of the *conditional* (or *posterior*) variance matrix for prediction of output at any collection of points in Δ . But in our setting, this is equivalent to finding the design for which $|\mathbf{R}_{DD}|$ is *maximized*. A major practical implication of this *partitioning of entropy* is that the optimality criterion is a function only of the points in D , not of those points in Δ not included in the design. Hence, while searching over designs still conceptually requires consideration of each collection of N points, no other values of x are required in evaluating the criterion function, as, for example, with the integration over Δ required by the IMSE criterion.

11.5.2 Distance-Based Designs

Because the correlation functions used in GaSP models can be most easily thought of as links between strength of a prior correlation between two responses and *distance between two points in Δ* , it is reasonable to think of experimental designs in geometric terms. An intuitively reasonable criterion for constructing N -point designs that are well spread throughout Δ is the *maximin distance criterion* calling for the design for which

$$\min_{1 \leq i < j \leq N} d(\mathbf{x}^i, \mathbf{x}^j)$$

is maximized for some distance measure d . This criterion actually has more than intuitive appeal. Suppose the correlation function R is rewritten so that its single scalar argument is $\theta \times d(\mathbf{x}, \mathbf{x}^*)$, where correlation decreases as this quantity increases, and where d is interpreted as a measure of distance between \mathbf{x} and \mathbf{x}^* . Johnson, Moore, and Ylvisaker (1990) showed that in the limit as $\theta \rightarrow \infty$ (and the resulting correlation between any pair of points consequently becomes weak), the limiting form of the entropy-optimal design is a maximin distance design (with respect to distance measure d) of *minimum index*, where “index” is the number of pairs of runs included in the design separated by the minimum distance. While this connection to entropy is asymptotic, it does

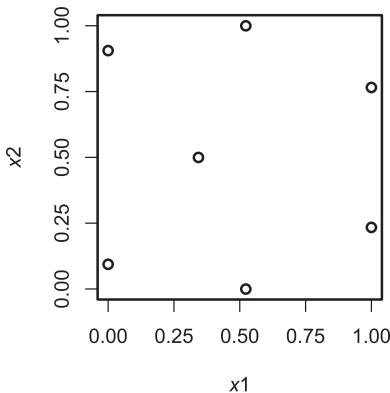


Figure 11.4 Maximin Euclidean distance design in $k = 2$ dimensions, in $N = 7$ runs.

suggest that for situations in which local correlations are expected to be weak (generally when y is expected to be significantly influenced by each element of \mathbf{x}), maximin distance designs can be expected to provide good predictive performance. An example of a maximin Euclidean distance design, in $k = 2$ dimensions and $N = 7$ runs, is shown in Figure 11.4.

11.5.3 Latin Hypercube Designs

Perhaps the designs most commonly used in computer experiments are generated as *Latin hypercube samples*. Latin hypercube samples were proposed by McKay, Beckman, and Conover (1979) as randomized sampling schemes for numerical integration. Suppose, as in sensitivity analysis (Section 11.2), that a probability distribution π has been specified for \mathbf{x} , and that interest centers on estimating the p th noncentral moment of y , thinking of the computer model as the transform function in a change-of-variables problem. A basic Monte Carlo approach to this problem is carried out by selecting a sample of N independent realizations of \mathbf{x} from π , evaluating the computer model for each, and estimating the required moment as the average of the observed output values, each raised to the power p .

McKay et al. described a stratified sampling plan, Latin hypercube sampling, that often improves the efficiency of this estimate when the elements of \mathbf{x} are distributed independently, that is, $\pi(\mathbf{x}) = \prod_{i=1}^k \pi_i(x_i)$. For each input x_i , the range of possible values is divided into a partition of N nonoverlapping, equal-probability intervals according to π_i , and one value is sampled from the conditional distribution associated with each interval. Let these N values be the entries of the N -element column vector \mathbf{x}_i . The array of input vectors to be sampled is:

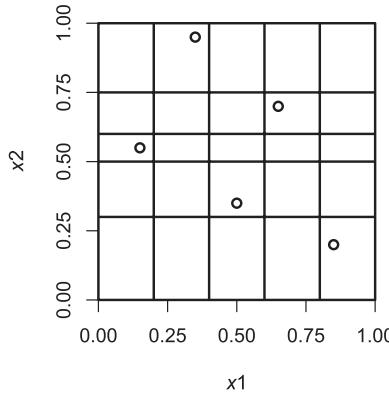


Figure 11.5 Example of a five-run Latin hypercube sample in $k = 2$ dimensions.

$$(p_1(\mathbf{x}_1), p_2(\mathbf{x}_2), p_3(\mathbf{x}_3), \dots, p_k(\mathbf{x}_k)), \quad (11.20)$$

where each p_i is an independent permutation operator. Each row of the array is an input vector at which y is to be evaluated. An example of a five -run Latin hypercube sample in $k = 2$ dimensions for $\Delta = (0, 1)^2$ is displayed in Figure 11.5.

McKay et al. showed that when y is a monotonic function in each element of \mathbf{x} , the variance of the Monte Carlo estimate is smaller when input vectors are selected by Latin hypercube sampling, than when they are selected by simple random sampling. Subsequent research (e.g., Stein 1987) has further elaborated the precision advantages of Latin hypercube sampling when used for estimating integrals.

The *sampling* properties of the Latin hypercube are not so critical in the present context, but aside from the variance reduction in integral estimation, the Latin hypercube *design* (as opposed to *sample*) has the intuitive appeal of being *space filling* in each input dimension (separately). If y is actually a function of only x_1 , say, there are N widely spread (x_1, y) pairs, and no “piling up” of runs to the same x_1 value as would be the case with factorial and fractional factorial plans with a small number of factor levels. As an experimental design, the Latin hypercube structure is sometimes simplified; rather than making draws from interval–conditional distributions, input values are sometimes selected as the conditional mean in each interval, the mid-point of each interval, or the N interior endpoints of $N + 1$ equiprobable intervals. If there is no especially relevant input distribution π , uniform marginal distributions over the respective intervals of interest are often used. In some cases, the Latin hypercube structure is maintained, but values of inputs are combined so as to optimize a design criterion, rather than through random permutation. For example, Morris and Mitchell (1995) constructed Latin hypercube designs via

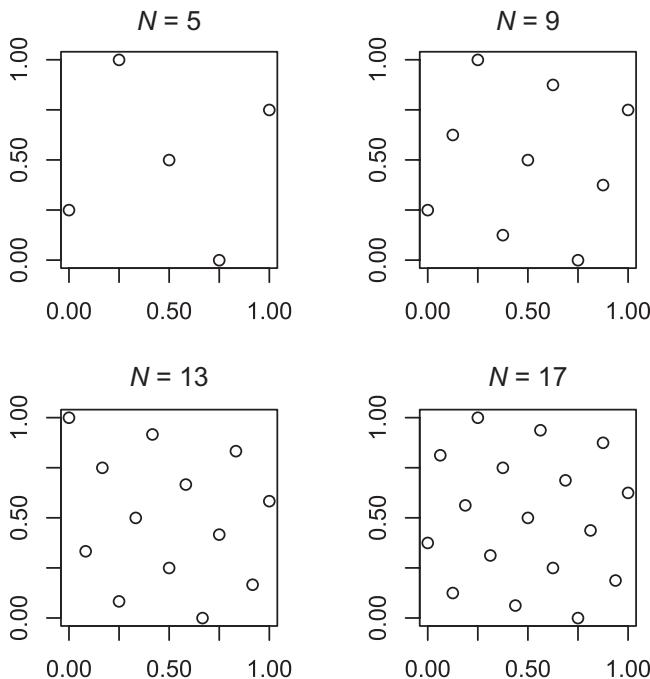


Figure 11.6 Maximin Euclidean distance Latin hypercube designs in $k = 2$ dimensions, in $N = 5$, 9, 13, and 17 runs.

the maximin distance criterion of Johnson, Moore, and Ylvisaker (1990); examples for $k = 2$ with $N = 5, 9, 13$, and 17 based on Euclidean distance are displayed in Figure 11.6.

Another use of Latin hypercube structure in the construction of designs for computer experiments is the orthogonal array-based Latin hypercube of Tang (1993). These are constructed by perturbing an orthogonal array (OA) of strength 2 in s symbols (so that N is necessarily a multiple of s^2) so that the values for each input are distributed uniformly over its range, but in such a way that the OA structure is recovered if the values of each input are rounded back to the nearest of the s original symbol values. Tang discussed variance reduction properties for numerical integration when randomized OA-based Latin hypercubes are used as sampling plans. More recent contributions to the literature on Latin hypercube designs are the method for generating Latin hypercube designs with orthogonal columns by Steinberg and Lin (2006), and the nested Latin hypercube designs of Qian (2009).

11.5.4 Uniform Designs

Uniform designs, or *minimum discrepancy designs* are constructed with the goal of making D as uniformly spread over Δ as possible (e.g., Fang et al. 2000).

Of course, “uniformly spread” can be defined in a number of ways. With uniform designs, this is done through specification of a cumulative distribution function F over Δ ; in practice, this is often the CDF of the k -variate uniform distribution with independent elements. Thinking of \mathbf{x} as a k -dimensional random vector, uniform designs are constructed by selecting D so that the empirical cumulative distribution function of the result is as close as possible to F , for example, so as to minimize a Kolmogorov–Smirnov statistic:

$$\max_{\mathbf{x} \in \Delta} |\hat{F}(\mathbf{x}) - F(\mathbf{x})|, \quad (11.21)$$

where $\hat{F} = 1/N \sum_{j=1}^N I(\mathbf{x}^j \leq \mathbf{x})$. More generally, D can be chosen to minimize one of a family of criteria, or *discrepancy functions*:

$$\phi_p(D) = \left[\int_{\Delta} |\hat{F}(\mathbf{x}) - F(\mathbf{x})|^p d\mathbf{x} \right]^{1/p}, \quad (11.22)$$

for positive p . When $p \rightarrow \infty$, the Kolomogorov–Smirnov index is recovered, and ϕ_∞ is sometimes called the *star discrepancy* function.

As with Latin hypercubes, original justification for uniform designs was in the context of numerical integration, but the resulting uniformity of design points throughout Δ is also effective for spatial modeling. The construction of uniform designs is computationally challenging due to the behavior of ϕ_p as a function of the design. In practice, ϕ_p (as well as other hard-to-optimize criterion functions, such as the minimum interpoint distance) is sometimes applied within the class of Latin hypercubes by randomly generating a large number of designs and applying the criterion to determine which is best.

11.6 MULTIVARIATE OUTPUT

In most applications, computer models actually produce far more than a single scalar-valued output. In some cases, it can be adequate to use scalar-valued modeling techniques for each input separately (as in Mitchell and Morris 1992), or to reduce/transform a vector-valued output to a scalar that reflects the information relevant to a specific set of experimental questions. But this approach does not take advantages of relationships between components of the output, and it is often not adequate when experimental questions are more broad and/or complex. Our notation for the more general case is $\mathbf{y} = f(\mathbf{x})$, where the M -dimensional output vector \mathbf{y} results from a single run of f . While a wide variety of multivariate output structures are possible, some more common forms are time series, spatial maps, and space–time combinations.

11.6.1 Extending the Univariate GaSP Model

Suppose that \mathbf{y} is a time series so that each element is associated with one of M times, $t_1, t_2, t_3, \dots, t_M$. One way to extend the basic GaSP modeling approach

described in Section 11.3 is to regard time as one of the inputs. For the u th output element resulting from input vector \mathbf{x}^i and the v th output element resulting from input vector \mathbf{x}^j , say:

$$\begin{aligned} E(y_u^i) &= E(y_v^j) = \mu, \quad \text{Var}(y_u^i) = \text{Var}(y_v^j) = \sigma^2, \\ \text{Corr}(y_u^i, y_v^j) &= R_t(t_u - t_v) \prod_{l=1}^k R_l(x_l^i - x_l^j). \end{aligned} \quad (11.23)$$

A similar approach can be taken for two-dimensional spatial output arrays (or maps), treating longitude and latitude as two additional inputs.

This approach can be useful when M is small, and the overall behavior of the output is similar over the entire span of simulated time or space. If output is a variable that, for example, always decreases with time, the mean structure can be altered to account for common trends, for example:

$$E(y_u^i) = \mu_u. \quad (11.24)$$

If M is not small, the resulting correlation matrix \mathbf{R}_{DD} can become impractically large (of order $N \times M$). In addition, \mathbf{R}_{DD} can easily become ill-conditioned because each input vector \mathbf{x}^i specified in the design is represented at all M times, and if $R_t(t_u - t_{u+1})$ is near one for some pairs of adjacent time points, this immediately leads to near-one off-diagonal elements of \mathbf{R}_{DD} .

Drignei (2006) suggested a modification to this approach based on the assumption that most of the features of the time series can be reasonably expressed using a small subgrid of $m \ll M$ times. His two-stage modeling process proceeds as described above for the relatively coarse m -point subgrid of time. Modeling in the second stage (for intermediate time values) is based on a *discrete Brownian bridge* process, conditioned on the end points modeled in the first stage. More specifically, suppose t_u and t_{u+T} are two consecutive time points on the coarse grid, so that the $T - 1$ intermediate times $t_{u+1}, t_{u+2}, t_{u+3}, \dots, t_{u+T-1}$ are bracketed by these. Let \mathbf{x}^0 denote a model input vector not included in the design. Response values for input vector \mathbf{x}^0 and times on the coarse grid are predicted directly using all observed data on the coarse grid. The *a priori* statistical model for responses at the intermediate times, for input vector \mathbf{x}^0 , is:

$$y_{v+1}(\mathbf{x}^0) = y_v(\mathbf{x}^0) + \phi \varepsilon_{v+1}, \quad v = u, u+1, u+2, \dots, u+T-1, \quad \varepsilon \text{ i.i.d. } \sim N(0, 1), \quad (11.25)$$

which is modified by conditioning on $y_u(\mathbf{x}^0)$ and $y_{u+T}(\mathbf{x}^0)$. The parameter ϕ is reasonably easy to estimate from the data. The predictive distribution of output indexed by a time not on the coarse grid is derived by integrating $y_u(\mathbf{x}^0)$ and $y_{u+T}(\mathbf{x}^0)$ out of

$$\pi_1(y_v(\mathbf{x}^0) | y_u(\mathbf{x}^0), y_{u+T}(\mathbf{x}^0)) \times \pi_2(y_u(\mathbf{x}^0), y_{u+T}(\mathbf{x}^0)), \quad (11.26)$$

where π_2 is the joint density of $y_u(\mathbf{x}^0)$ and $y_{u+T}(\mathbf{x}^0)$ conditioned on the data indexed by the coarse time grid from all model runs, and π_1 is the conditional (only on the two endpoints) density for the discrete Brownian bridge process. The result is univariate normal, with a mean function that is linear in time, and a variance function that is quadratic in time, between any two coarse-grid time points.

11.6.2 Principal Components

Dimension reduction through the use of principal components is common in many statistical modeling contexts, and has been usefully employed in computer experiments, for example, Higdon et al. (2004). The approach does not depend on a particular data structure, such as a time series, and so is broadly applicable. The underlying assumption is that much of the variation among functional outputs can be described in terms of a relative few characteristic patterns, and that most individual output functions can be accurately approximated as a linear combination of these patterns. After computing N output vectors, a general approach is to construct principle components of these, approximately reexpress each output function as a weighted sum of these principle components, and then model the weight associated with each component as an independent, univariate GaSP across the input domain. Then, output prediction is accomplished by predicting weights at an input vector of interest, and constructing a functional output prediction from these, usually treating the computed principle components as known.

11.6.3 Derivatives

A special case of multivariate output occurs with computer models that compute both a scalar output and one or more derivatives of the output with respect to inputs. Such models are sometimes called *enhanced*, and are of particular interest in applications where the sensitivity of y to small changes in \mathbf{x} is important, for example, Grewank and Walther (2008). Morris, Mitchell, and Ylvisaker (1993) demonstrate how the GaSP modeling approach is especially natural for such applications, because where y is modeled as a GaSP indexed by \mathbf{x} with correlation function R , derivatives of y with respect to the elements of \mathbf{x} , of all orders, are also Gaussian stochastic processes indexed by \mathbf{x} with correlation functions that can be derived from R , and y and its derivatives are also *jointly* Gaussian. The implication is that, for example, for N evaluations of a computer model involving k input variables, the collection of all $N(k + 1)$ values of y and its first derivatives can be jointly modeled, and predictions made for any of them at any point in the Δ . A practical computational difficulty, as with direct extension of the univariate GaSP to modeling time series outputs (Section 11.6.1) is that the order of \mathbf{R}_{DD} is large if either N or k is large.

11.7 MULTIPLE DATA SOURCES

In many applications, data characterizing the behavior of a system are not limited to the output of a single model. Multiple scientifically reasonable models of a physical system may be available that produce different output values for the same set of input values. Also (and sometimes in addition), at least some limited measurement data can be collected from the system of interest that can be expected to differ at least somewhat from output if either the model or the measurement process is not perfect. This section outlines a framework that has been proposed for dealing with multiple sources of data such as these.

11.7.1 Multiple Models

It is often the case that several computer models, or versions of the same computer model, are available for simulating the same physical system. Examples include (1) finite difference or finite element codes that are essentially the same, but that use grids of differing size, (2) codes that are structurally similar, but in which some are simplified by omitting or modifying subprocesses that are believed to be of minor importance, and (3) codes that are fundamentally different, representing the same reality with different mathematical structures. Here, we focus on pairs of computer models that (1) require the same set of inputs, and (2) can be ordered such that *fidelity of prediction* is inversely related to *computational cost*. Note that the second point is likely true in cases 1 and 2 above, but not necessarily in case 3.

Kennedy and O'Hagan (2000) presented a framework in which a collection of ordered (by fidelity, and inversely by cost) models can be jointly analyzed. We focus on the case of two models and refer to such a pair as an *accurate code*, $y_A = f_A(\mathbf{x})$, and a *fast code*, $y_F = f_F(\mathbf{x})$. Beginning with the requirements:

1. y_F can be modeled as a stationary GaSP with parameters μ_F , σ_F^2 , and $\boldsymbol{\theta}_F$; and
2. y_A can also be modeled as a stationary GaSP, but such that for any two input vectors $\mathbf{x} \neq \mathbf{x}^*$, $\text{Cov}(y_A(\mathbf{x}), y_F(\mathbf{x}^*)|y_F(\mathbf{x})) = 0$,

Kennedy and O'Hagan noted that as a result:

$$y_A(\mathbf{x}) = \rho y_F(\mathbf{x}) + \delta(\mathbf{x}), \quad (11.27)$$

where $\delta(\mathbf{x})$ is another stationary GaSP with parameters μ_δ , σ_δ^2 , and $\boldsymbol{\theta}_\delta$, independent of $y_F(\mathbf{x})$. Denote designs and resulting scalar data for the accurate and fast models as, respectively:

$$\begin{aligned} D_A &= \{\mathbf{x}_A^i, i = 1, 2, 3, \dots, N_A\} \quad \mathbf{y}_A \\ D_F &= \{\mathbf{x}_F^i, i = 1, 2, 3, \dots, N_F\} \quad \mathbf{y}_F \end{aligned}$$

Interest centers on making predictive inference about the accurate model, given a relatively few runs of it, but relatively more of the fast model, i.e., $N_A < N_F$. The prior joint model for observed data is multivariate Gaussian with

$$\begin{aligned} E(\mathbf{y}_F) &= \mu_F \mathbf{1} & \text{Var}(\mathbf{y}_F) &= \sigma_F^2 \mathbf{R}_{FF} \\ E(\mathbf{y}_A) &= (\rho\mu_F + \mu_\delta)\mathbf{1} & \text{Var}(\mathbf{y}_A) &= \rho^2\sigma_F^2 \mathbf{R}_{AA} + \sigma_\delta^2 \mathbf{S}_{AA}, \\ & & \text{Cov}(\mathbf{y}_A, \mathbf{y}_F) &= \rho\sigma_F^2 \mathbf{R}_{AF} \end{aligned} \quad (11.28)$$

where \mathbf{R}_{FF} , \mathbf{R}_{AA} , and \mathbf{R}_{AF} are correlation matrices based on the correlation function for y_F (i.e., parameterized by $\boldsymbol{\theta}_F$), containing correlations between pairs of points in D_F , pairs of points in D_A , and pairs of points with one in D_A and one in D_F , respectively; and \mathbf{S}_{AA} is the correlation matrix based on the correlation function for the process δ for pairs of points in D_A . Prediction of y_A at a new point depends on inference about the parameters ρ , μ_F , μ_A , σ_F^2 , σ_δ^2 , $\boldsymbol{\theta}_F$, and $\boldsymbol{\theta}_\delta$. Kennedy and O'Hagan begin with a fully Bayesian framework that leads to elimination of the process means, and then use point estimates of the remaining parameters by maximizing the likelihood:

$$p(\mathbf{y}_F, \mathbf{y}_A | \rho, \sigma_F^2, \boldsymbol{\theta}_F, \sigma_\delta^2, \boldsymbol{\theta}_\delta).$$

In some applications, the model is simplified by setting $\rho = 1$. The function $\delta(\mathbf{x})$ is called the *discrepancy function* and is of interest in its own right since it is useful in determining where (in Δ) the fast code approximates the accurate code well, and where it does not.

11.7.2 Model and Reality

With modifications, the Kennedy and O'Hagan (2000) approach for joint analysis of fast and accurate models can also be used for joint analysis of computer model output y_M (now taking the place of the fast model above) and experimental data y_E (now taking the place of the accurate model above). Here, the discrepancy function can play an important role in at least one version of model *validation*, that is, in assessing the conditions (x) under which the model represents reality well. When all model parameters \mathbf{t} are specified, the primary technical modification that is necessary in most applications is the addition of a nugget to the covariance function for the physical data to accommodate measurement error, or other random variation that would be observed at repeated determinations of physical y at the same \mathbf{x} .

Another important aspect of the joint analysis of model output and physical data is that of model *calibration*—the estimation of model parameters that cannot be directly observed or measured based on the agreement of model output and physical measurements. Express the experimental data as “model” plus “discrepancy” plus “measurement error” as follows:

$$y_E(\mathbf{x}) = \rho y_M(\mathbf{x}, \boldsymbol{\tau}) + \delta(\mathbf{x}) + \varepsilon, \quad (11.29)$$

where model inadequacy is represented by δ , calibration uncertainty is associated with not knowing τ , the *true* value of the parameter vector \mathbf{t} , and measurement error is associated with ε . Borrowing from the last section, we can model y_M as a stationary GaSP with parameters μ_M , σ_M^2 , and $\boldsymbol{\theta}_M$. y_E is then modeled via Equation (11.27) where $\delta(\mathbf{x})$ is an independent GaSP with parameters μ_δ , σ_δ^2 , and $\boldsymbol{\theta}_\delta$, and ε is i.i.d. $N(0, \sigma_\varepsilon^2)$, implying that for fixed parameters, y_E and y_M are jointly Gaussian. Designs for both experimental and computer data and resulting outputs are:

$$\begin{aligned} D_E &= \{\mathbf{x}_E^i, i = 1, 2, 3, \dots, N_E\} & \mathbf{y}_E \\ D_M &= \{(\mathbf{x}_M^i, \mathbf{t}^i), i = 1, 2, 3, \dots, N_M\} & \mathbf{y}_M \end{aligned} \quad (11.30)$$

Interest centers on making inferences both about outputs and the “best fitting” value of \mathbf{t} (which, depending on model inadequacy, may or may not be τ), and prediction of y_E for conditions not observed.

The prior joint model for all data is then:

$$\begin{aligned} E(\mathbf{y}_M) &= \mu_M \mathbf{1} & \text{Var}(\mathbf{y}_M) &= \sigma_M^2 \mathbf{R}_{MM} \\ E(\mathbf{y}_E) &= (\rho \mu_M + \mu_\delta) \mathbf{1} & \text{Var}(\mathbf{y}_E) &= \rho^2 \sigma_M^2 \mathbf{R}_{EE} + \sigma_\delta^2 \mathbf{S}_{EE} + \sigma_\varepsilon^2 \mathbf{I}, \\ \text{Cov}(\mathbf{y}_E, \mathbf{y}_M) &= \rho \sigma_M^2 \mathbf{R}_{EM}(\tau) \end{aligned} \quad (11.31)$$

where \mathbf{R}_{MM} , \mathbf{R}_{EE} , and \mathbf{R}_{EM} are correlation matrices based on the correlation function for the y_M GaSP (i.e., parameterized by $\boldsymbol{\theta}_M$), containing correlations between pairs of points in D_M , pairs of points in D_E , and pairs of points with one in D_E and one in D_M , respectively; and \mathbf{S}_{EE} is the correlation matrix based on the correlation function for δ for the pairs of points in D_E . Note that \mathbf{R}_{EM} is written as a function of the unknown parameter vector τ , since the correlations being expressed here are between the component of y_M in the first term of y_E , and y_M as output from a computer run based on a value of \mathbf{t} specified in the design.

Now let $\mathbf{y} = (\mathbf{y}'_M, \mathbf{y}'_E)'$, and let \mathbf{V} denote $\text{Var}(\mathbf{y})$. Define:

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{N_M} & \mathbf{0}_{N_M} \\ \rho \mathbf{1}_{N_E} & \mathbf{1}_{N_E} \end{pmatrix}.$$

Then for fixed values of the other parameters:

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\mu}_M \\ \hat{\mu}_E \end{pmatrix} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{y}.$$

Let $\tilde{y}_E(\mathbf{x}^0)$ be the “noiseless truth” (i.e., reality without measurement error) at \mathbf{x}^0 :

$$\tilde{y}_E(\mathbf{x}^0) = \rho y_M(\mathbf{x}^0, \tau) + \delta(\mathbf{x}^0).$$

Again, putting a noninformative prior on μ and integrating it out, but conditional on all other parameters, the posterior distribution of $\tilde{y}_E(\mathbf{x}^0)$ is normal with:

$$\begin{aligned} E(\tilde{y}_E(\mathbf{x}^0) | \sigma_M^2, \boldsymbol{\theta}_M, \sigma_\delta^2, \boldsymbol{\theta}_\delta, \rho, \boldsymbol{\tau}) &= \mathbf{m}'\hat{\mu} + \mathbf{v}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mu}) \\ Var(\tilde{y}_E(\mathbf{x}^0) | \sigma_M^2, \boldsymbol{\theta}_M, \sigma_\delta^2, \boldsymbol{\theta}_\delta, \rho, \boldsymbol{\tau}) &= (\rho^2\sigma_M^2 + \sigma_\rho^2) - \mathbf{v}'\mathbf{V}^{-1}\mathbf{v} + \\ &\quad (\mathbf{m}' - \mathbf{v}'\mathbf{V}^{-1}\mathbf{X})(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{m}' - \mathbf{v}'\mathbf{V}^{-1}\mathbf{X})', \end{aligned} \quad (11.32)$$

where $\mathbf{m}' = (\rho, 1)$, and $\mathbf{v}' = \text{Cov}(y_E^0, \mathbf{y}) = (\rho\sigma_M^2\mathbf{R}_{0M}, \rho^2\sigma_M^2\mathbf{R}_{0E} + \sigma_\delta^2\mathbf{S}_{0E})$. The involved vectors and matrix are \mathbf{m} , which depends on ρ ; \mathbf{v} , which depends on $\rho, \sigma_M^2, \sigma_\delta^2, \boldsymbol{\theta}_M, \boldsymbol{\theta}_\delta$, and $\boldsymbol{\tau}$, and \mathbf{V} , which depends on $\rho, \sigma_M^2, \sigma_\delta^2, \boldsymbol{\theta}_M, \boldsymbol{\theta}_\delta, \sigma_e^2$, and $\boldsymbol{\tau}$. Development can follow from MLE's used in place of parameters, via a full Bayesian treatment, or a combination.

11.8 CONCLUSION

Experimental studies focused (at least partly) on computer models are often performed to answer the same kinds of questions as experiments performed in reality. In many cases, computer experiments are undertaken because their physical counterparts are too expensive, too dangerous, or impossible. Of course, when only a model is studied, the results apply only to the model— inference about the system or phenomenon being modeled requires at least some information about the model's validity.

Still, models that are carefully constructed from what is known and what has been observed can be useful tools in understanding phenomena ranging from economic processes to environmental systems to particle physics to national security issues (see Chapter 12). The methods briefly reviewed here have been shown to be useful in understanding computer models used in many fields. The rapid evolution of computer resources suggests that computer models of entirely different scale (and perhaps even fundamental form) may soon be central tools in many application areas. The statistical field of computer experiments is in its infancy, and as computer modeling continues to advance, there will be fascinating challenges in the development of methodology that takes advantage of these models.

REFERENCES

- Booker, A. (2000). Well-conditioned Kriging models for optimization of computer simulations. Boeing Technical Report M&CT-TECH-00-002.

- Cukier, R.I., C.M. Fortuin, K.E. Shuler, A.G. Petschek, and J.H. Schiably (1973). Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients, I. Theory. *Journal of Chemical Physics*, **59**, 3873–3878.
- Drignei, D. (2006). Empirical Bayesian analysis for high-dimensional computer output. *Technometrics*, **48**, 230–240.
- Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.
- Fang, K.T., D.K.J. Lin, P. Winker, and Y. Zhang (2000). Uniform design: Theory and applications. *Technometrics*, **42**, 237–248.
- Furrer, R., M.G. Genton, and D. Nychka (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, **15**, 502–523.
- Gramacy, R.B. and H.K.H. Lee (2008). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, **103**, 1119–1130.
- Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.
- Grewank, A. and A. Walther (2008). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation* (2nd ed.). Philadelphia: SIAM.
- Handcock, M.S. and M.L. Stein (1993). A Bayesian analysis of Kriging. *Technometrics*, **35**, 403–410.
- Heebner, D. and L. Toran (2000). Sensitivity analysis of three-dimensional steady-state and transient spray irrigation models. *Ground Water*, **38**, 20–28.
- Higdon, D., M. Kennedy, J. Cavendish, J. Cafeo, and R.D. Ryne (2004). Combining field observations and simulations for calibration and prediction. *SIAM Journal of Scientific Computing*, **26**, 448–466.
- Iman, R.L. and W.J. Conover (1980). Small sample sensitivity analysis techniques for computer models with an application to risk assessment. *Communications in Statistics A: T&M*, **9**, 1749–1842.
- Johnson, M.E., L.M. Moore, and D. Ylvisaker (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, **26**, 131–148.
- Kennedy, M. and A. O'Hagan (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, **87**, 1–13.
- Li, R. and A. Sudjianto (2005). Analysis of computer experiments using penalized likelihood in Gaussian Kriging models. *Technometrics*, **47**, 111–120.
- Marin, J.-M. and C.P. Robert (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. New York: Springer.
- McDonald, M.G. and A.W. Harbaugh (1988). A modular three-dimensional finite difference ground-water flow model. *U.S. Geological Survey Techniques of Water-Resources Investigations*, Book 6, chapter A1.
- McKay, M.D. (1995). Evaluating prediction uncertainty. Report NUREG/CR-6311, LA-12915-MS, Los Alamos National Laboratory.
- McKay, M., R. Beckman, and J. Conover (1979). A comparison of three methods of selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21**, 239–245.

- Mitchell, T.J. and M.D. Morris (1992). Bayesian design and analysis of computer experiments: Two examples. *Statistica Sinica*, **2**, 359–379.
- Mitchell, T.J. and D.S. Scott (1987). A computer program for the design of group testing experiments. *Communications in Statistics—Theory and Methods*, **16**, 2943–2955.
- Morris, M.D. and T.J. Mitchell (1995). Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, **43**, 381–402.
- Morris, M.D., T.J. Mitchell, and D. Ylvisaker (1993). Bayesian design and analysis of computer experiments: Use of derivatives in surface prediction. *Technometrics*, **35**, 243–255.
- Morris, M.D., L.M. Moore, and M.D. McKay (2008). Using orthogonal arrays in the sensitivity analysis of computer models. *Technometrics*, **50**, 205–215.
- Qian, P.Z.G. (2009). Nested Latin hypercube designs. *Biometrika*, **96**, 957–970.
- Sacks, J., S.B. Schiller, and W.J. Welch (1989). Designs for computer experiments. *Technometrics*, **31**, 41–47.
- Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, **145**, 280–297.
- Shewry, M.C. and H.P. Wynn (1987). Maximum entropy sampling. *Journal of Applied Statistics*, **14**, 165–170.
- Sobol, I.M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, **1**, 407–414.
- Stein, M.L. (1987). Large sample properties of simulations using Latin hypercube sampling. *Technometrics*, **29**, 143–151.
- Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer.
- Steinberg, D.M. and D.K.J. Lin (2006). A construction method for orthogonal Latin hypercube designs. *Biometrika*, **93**, 279–288.
- Tang, B. (1993). Orthogonal array-based Latin hypercubes. *Journal of the American Statistical Association*, **88**, 1392–1397.
- Watson, G.S. (1961). A study of the group screening method. *Technometrics*, **3**, 371–388.
- Wyle, H. (1938). Mean motion. *American Journal of Mathematics*, **60**, 889–896.
- Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

C H A P T E R 12

Designs for Large-Scale Simulation Experiments, with Applications to Defense and Homeland Security

***Susan M. Sanchez, Thomas W. Lucas, Paul J. Sanchez,
Christopher J. Nannini, and Hong Wan***

12.1 INTRODUCTION

Computer experimentation is integral to modern scientific research, national defense, industry and manufacturing, and in public policy debates. Traditional design of experiments (DOE) focuses on small-scale experimentation, whether dealing with experiments involving physical systems or computer models. In contrast, computer models tend to be extremely complex, often with thousands of factors and many sources of uncertainty. Historically, even if experimental designs have been used, they have typically been applied to only a handful of factors—even for computer models having hundreds or thousands of inputs. This suggests that more modelers and analysts need to be aware of the power of experimental design—especially the recent breakthroughs in large-scale experimental designs that enable us to understand the impact of many factors and their intricate interactions on model outcomes.

In this chapter, we begin by considering some philosophical issues about exploring complex, large-scale systems. We review a portfolio of designs we have developed or collected and successfully used to support decision makers in defense and homeland security. These include single-stage designs appropriate for hundreds of factors, as well as sequential approaches that can be used to screen thousands of factors. We illustrate these concepts with a simulation

Design and Analysis of Experiments: Special Designs and Applications, First Edition. Edited by Klaus Hinkelmann.

© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

model developed to assist in the analysis of unmanned aerial vehicle (UAV) requirements for the U.S. Army. This analysis was part of a larger effort by the U.S. Army to utilize advanced technologies to conduct the most comprehensive transformation of the force since World War II. We end with a few parting thoughts about the future of experimental design.

12.2 PHILOSOPHY: EVOLUTION OF COMPUTATIONAL EXPERIMENTS

If anyone, here or later, can tell us how the approach of certainty—traditional mathematics—is going to answer the questions that practical data analysts are going to have to have answered, I will rejoice. But until I am reliably informed of such a utopian prospect, I shall expect the critical practical answers of the next decade or so to come from the approach of simulation—from a statistician’s form of mathematics, in which ever more powerful computing systems will be an essential partner and effective . . .

JOHN TUKEY (1986)

12.2.1 Context

Terminology is used differently in different fields of study, so we want to clarify our usage of terms. For example, those coming from the world of physical experimentation often use the terms *computer model* or *mathematical model* to describe different types of models (see Chapter 11). The Department of Defense (DoD) makes distinctions between *live*, *virtual*, and *constructive* simulations, although specific simulations may involve elements from more than one category. Live simulations involve real people operating real systems, and are extremely limited in terms of replicability; these tend to be heavily-scripted exercises, but are often referred to within DoD as “experiments.” Virtual simulations are often used for training purposes and based around realistic human/machine interactions, such as physical flight simulators. Constructive simulations are computer simulations that do not involve real-time human interaction, thus they are amenable to high-dimensional exploration. In the field of operations research, the term *computer model* is often used to describe numerical evaluations of, for example, partial differential equation or finite element models. In contrast, the term *simulation model* describes a computerized implementation of a mathematical or algorithmic model: The model can be deterministic or stochastic; static or dynamic; and, if dynamic, time can be advanced in a discrete-event or time-stepped fashion.

In this chapter, we emphasize the analysis of constructive simulations, and consider *simulation model* and *computer model* to be interchangeable terms. Model inputs are one potential source of factors, but others are possible. For example, the model might require the number of entities of each of two types as inputs, while the factor of interest might be their proportion. Embedded

components, such as the shape or parameters of random distributions, can also be explored. It is often the case that we have a firm understanding of the structural form of the model, but the parameterization is uncertain. For example, under certain assumptions, we may be quite certain that an arrival process is Markovian without having any certainty about the arrival rate. Thus, it would be suitable to study the impact of the parameterization using a designed experiment. Note that this differs from the use of *parameter* in Chapter 11 to denote values for model entities that are constant.

It is important to realize that there are two levels of modeling involved in simulation experiments. The first level involves modeling the phenomena of interest by building a simulation. The second involves building statistical models of the simulation's output—i.e., a model of a model. Because of the two-tiered nature of this modeling process, the end result is often referred to within the simulation community as a *response surface metamodel* or simply a *metamodel*.

Simulation is regularly used to inform decision makers about a broad range of important problems. For example, DoD regularly makes decisions on acquisition programs, equipment employment options, and recommended tactics, techniques, and procedures. These choices often involve billions of dollars, and, more importantly, can save or cost many lives. Indeed, there are hundreds of simulations in the Modeling and Simulation Coordination Office's online repository (see <http://www.msco.mil>). These models can be almost unimaginably complex and expensive. Models can have literally hundreds of thousands of possible input variables that can be set to different levels (Saeger and Hinch 2001). Many of these input variables are highly uncertain—such as the when, where, and with whom future security and defense events may take place. Furthermore, the performance and reliability of human participants and the systems they use are unknown.

This is a rich environment for applying experimental design. Countless man-hours have been invested in building simulation models, and, in some cases, millions of dollars have been spent collecting data to feed into these models. Given this investment, it is important to glean as much insight as possible from the simulations. Well-designed experiments make this possible.

12.2.2 Why Simulation?

Clearly, there are circumstances where we have little or no choice but to turn to models because of the expense, risk, or feasibility of gathering real-world data. In the past, analytic mathematical models were considered to be the gold standard—and a cliché sometimes bandied about is that simulation should be used as a modeling method of last resort. We disagree. As Hammersley and Handscomb (1964) said, “The idea behind [simulation] . . . is to [replace] theory by experiment whenever the former falters.” In an attempt to achieve mathematical tractability, modelers often commit a type III error, defined by Mitroff and Featheringham (1974) as “the error . . . [of] choosing the wrong

problem representation.” Consider some of the assumptions often made in creating models that can be solved analytically: Systems are often modeled as linear, time or state invariant, memoryless, or deterministic. Even if they are modeled as stochastic, they frequently incorporate mathematically convenient distributions for i.i.d. random components and assume stationary responses.

In contrast, real-world systems are often characterized by nonlinear behaviors, state or time dependent trajectories, many sources of uncertainty, heteroskedasticity, feedback loops with and without phase delays, and transient effects. By incorporating these features, simulation models help us avoid type III errors.

In addition to providing suitably realistic representations of the problem, simulation allows us to address several potential goals. Sacks et al. (1989) state that simulation can be used for prediction, calibration, and optimization. Earlier in this book (see Chapter 11), Morris states “that the immediate purpose of [an] experiment is to develop a ‘good’ approximation/prediction of the computer model that can be used throughout [the set of possible input vectors].” We take the broader view described by Kleijnen (2005) that, in practice, the goals of a simulation experiment often are:

- developing a basic understanding of a particular simulation model or system;
- finding robust decisions or policies; and
- comparing the merits of various decisions or policies.

Seeking a basic understanding differs from “testing hypotheses about factor effects” because we may seek insight into situations where the underlying mechanisms are not well understood, and where real-world data are limited or even nonexistent. It may include identifying the most important factors and interactions, or identifying factor ranges, thresholds, or regions of stability or abrupt transition. Robust solutions are ones that yield acceptable performance across a broad range of circumstances. By contrast, so-called *optimal* solutions often implicitly depend on knowing or controlling a large number of circumstances within the simulation that are unknown or uncontrollable in reality. Comparing the merits of qualitatively different alternatives can be accomplished through statistical ranking and selection techniques, such as those found in Goldsman, Kim, and Nelson (2005).

12.2.3 Why DOE?

Recent advances in high-performance computing have pushed computational capabilities to a petaflop (a thousand trillion operations per second) in a single computing cluster. This breakthrough has been hailed as a way to fundamentally change science and engineering by letting people perform experiments that were previously beyond reach. But for those interested in exploring the I/O behavior of their simulation model, efficient experimental design has a

much higher payoff at a much lower cost. A well-designed experiment allows the analyst to examine many more factors than would otherwise be possible, while providing insights that cannot be gleaned from trial-and-error approaches or by sampling factors one at a time.

In June 2008, a new supercomputer called the “Roadrunner” was unveiled. This petaflop bank of machines was assembled from components originally designed for the video game industry and it cost \$133 million. The *New York Times* coverage included the following description: “*By running programs that find a solution in hours or even less time, compared with as long as three months on older generations of computers, petaflop machines like Roadrunner have the potential to fundamentally alter science and engineering, supercomputer experts say. Researchers can ask questions and receive answers virtually interactively and can perform experiments that would previously have been impractical*” (Markoff 2008).

Yet let us take a closer look at the practicality of a brute-force approach to simulation experiments. Suppose a simulation has 100 factors, each with two levels. In order to fully evaluate all of the factor combinations, 2^{100} (about 10^{30}) runs of the model are necessary. Is this feasible? Former Air Force Major General Jasper Welch succinctly summarized the analyst’s dilemma by the phrase “ 10^{30} is forever” (Hoeber 1981). When he said this roughly three decades ago, using a computer that could evaluate a model run in a nanosecond, an analyst who started making runs at the dawn of the universe would have completed less than one-tenth of 1% of the runs. Even today, with a petaflop computer and a simulation that runs as fast as a single machine hardware operation, running a single replication of this experiment would take over 40 million years.

Efficient design of experiments can break this curse of dimensionality at a tiny fraction of the cost. For example, we can now use a 2_{V}^{100-85} Resolution V fractional factorial (with 32,768 design points) to study 100 factors and all their two-factor interactions. How quickly can we finish a single replication of the design? On a desktop computer with a simulation that takes a full second to run, this experiment takes under 9.5 hours; even if the simulation takes a more reasonable one minute to run, we can finish this experiment on an 8-core desktop over a weekend. For a more complicated simulation that takes, say, one hour per run, this experiment can still be completed over a weekend by using a sixty-node computing cluster. We now have other designs that are even more efficient, and may provide more detailed insights into the simulation model’s behavior.

12.2.4 Which DOE?

Which experiment setup should we choose? Although the question is straightforward, the answer may not always be easy, as different criteria have been developed to compare competing designs . . .

HINKELMANN AND KEMPTHORNE (2008, p. 59)

DOE has a rich history, and there are many designs that one could potentially use, including so-called *optimal designs* (Fedorov 1972). After the above quote, Hinkelmann and Kempthorne go on to state “One of the most important criteria is that of optimality, or better, variance optimality. By this we mean maximum precision (in some sense) in estimating linear combinations of treatment effects.” This may be true for physical experiments when the traditional assumptions are met, but it is almost certainly not the case for complicated simulation experiments. Instead, other criteria are more important than the number of design points when assessing the quality of an experimental design.

An underlying principle for optimal design is that sampling is expensive—the goal is to take no more samples than absolutely necessary. This does not translate well to the simulation world. For example, the time required for the total sampling effort is not necessarily proportional to the size of the design for several reasons. First, some design points may take a short time to run, while others may take orders of magnitude more time. Second, the speed of the computing hardware affects things: there may not be a practical difference (from the analysts’ perspective) between experiments that take one hour to complete versus several days to complete, in the context of the larger study. Third, if the experiment is being conducted on a computing cluster, as we typically do, we can collect orders of magnitude more data in the same amount of time required for an experiment executed on a single processor. Finally, with the large volumes of data generated by many simulation experiments, achieving statistical significance is less of a concern than with physical experiments. Parsimonious fits are desirable, thus we may eliminate many statistically significant effects from the fitted models if they have little or no practical importance.

Selecting a design is an art, as well as a science. Clearly, the number of factors and the mix of different factor types (binary, qualitative, or discrete with a limited number of levels, discrete with many levels, or continuous) play important roles. But these are rarely cast in stone—particularly during exploratory analysis. The experimenter has control over how factors are grouped, how levels are determined, etc. Even if these are specified, different experimenters may prefer different designs. The choice of design should consider a breadth of issues, such as the anticipated complexity of the response, the time required to run the simulation, the processing resources available, the ease of introducing changes to the model parameters, and more (see Kleijnen et al. (2005) for an expanded discussion). Therefore, simulation analysts need a portfolio of designs.

Of the multitude of designs we have used, in hundreds of simulations studies, the closest to an all-purpose class of designs when the factors of interest are mostly continuous and there is considerable *a priori* uncertainty about the response are Latin hypercube (LH) designs. The primary reasons that we have found LHs to be good all-purpose designs are:

- *Design Flexibility.* We can readily generate an LH for any combination of continuous k factors and desired number of design points $n \geq k$. Indeed, we have found that as long as n is much larger than k , simple rounding enables us to generate reasonable designs even if some factors are discrete valued with fewer than n levels.
- *Space-Filling.* LHs sample throughout the experimental region—not just at corner points. Specifically, if we look at any group of factors, we will find a variety of combinations of levels. As Santner, Williams, and Notz (2003) say, space-filling designs “allow one to fit a variety of models and provide information about all portions of the experimental region.”
- *Analysis Flexibility.* The resultant output data allow us to fit many different models to multiple performance measures. In particular, these designs permit us to simultaneously screen many factors for significance and fit very complex meta-models to a handful of dominant variables. This flexibility also extends to visual investigations of the data (Sanchez and Lucas 2002), as we get many “cameras on the landscapes” of relationships between inputs and outputs.

Latin Hypercube designs are discussed in Chapter 11; indeed, randomly generated LHs have been used in many computational studies over the years. For any given combination of the number of factors k and the number of design points n ($n \geq k$), there are $(n!)^{k-1}$ possible *lattice* LH designs. For lattice LH designs, the number of levels for factor i (ℓ_i) is set to n for all i , meaning that every factor is sampled at n equally-spaced values. Rather than select one lattice LH at random, we prefer to use a design matrix whose columns are orthogonal (or nearly orthogonal) and that has good space-filling properties. Cioppa and Lucas (2007) extend work by Ye (1998) to construct and tabulate nearly orthogonal Latin hypercubes (NOLHs) that have good space-filling properties in multiple dimensions, and Hernandez, Lucas, and Carlyle (2011) develop a mixed-integer programming approach that allows NOLHs to be generated for most any k and $n > k$. Vieira et al. (2011) extend this mixed-integer programming approach to enable the construction of nearly balanced, nearly orthogonal designs involving both discrete and continuous factors.

Scatter plot matrices of four different designs are shown in Figure 12.1. These are a 2^4 factorial design, a 4^4 factorial design, a space-filling NOLH design with 17 design points, and a NOLH design with 257 design points. Each subplot within one of these four matrices represents the projection of the entire design into two dimensions, and shows the combinations of levels of factors X_i and X_j that appear in the design. For example, the top row of the each scatter plot matrix shows the projections onto X_1X_2 , X_1X_3 , and X_1X_4 , respectively; although the 2^4 factorial has 16 design points, there are only four combinations of factor values (one in each corner) for each of these projections. The two-dimensional space-filling behavior of the NOLH compares

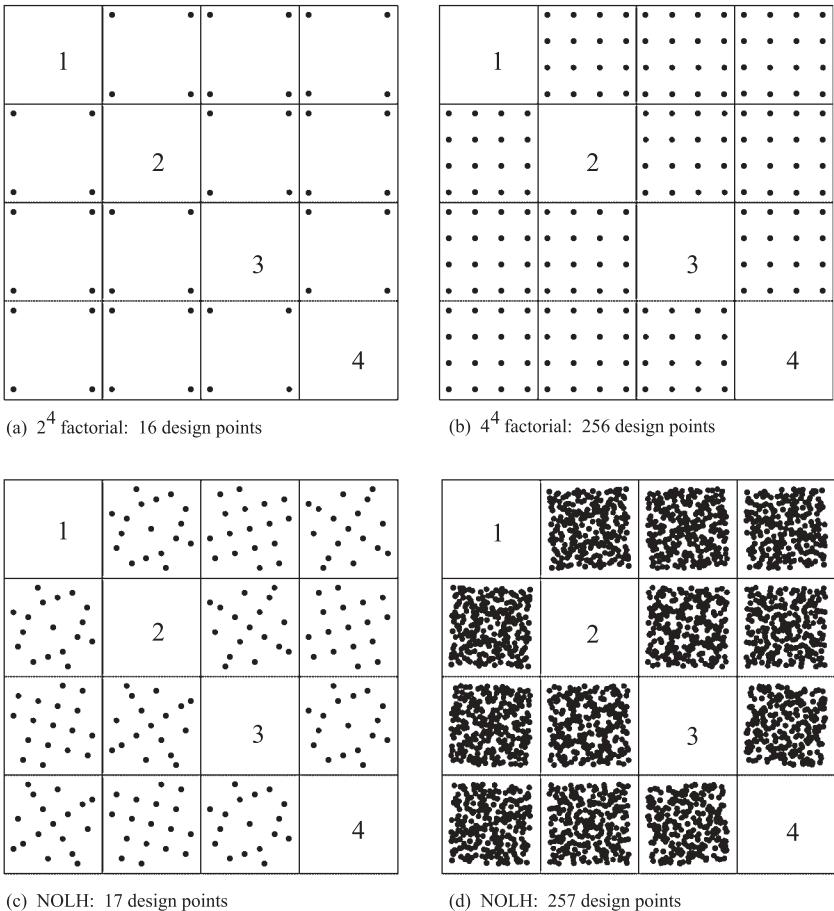


Figure 12.1 Scatter plot matrices for selected factorial and NOLH designs.

favorably with that of the 4^4 factorial for roughly 1/15 the computational effort, so experimenters concerned about the level of computational effort might prefer the latter. Alternatively, experimenters considering the use of the 4^4 factorial (and thus willing to run 256 design points) might prefer the NOLH with 257 design points (just one more)—and gain the ability to examine a much denser set of factor-level combinations, as well as explore up to 25 additional factors using the same design!

Other designs in our portfolio include two-level Resolution V fractional factorial designs for large k . These allow all main effects and two-factor interactions to be fit, can be used for a mix of qualitative and quantitative factors, and may be more useful for simulation analysts than saturated or nearly saturated designs. Augmenting these to central composite designs is also beneficial when factors are quantitative. Until recently it was difficult to find these

designs for more than about a dozen factors. The largest in Montgomery (2005) is a 2^{10-3} ; the largest in Box, Hunter, and Hunter (2005) and NIST/Sematech (2006) is a 2^{11-4} . Sanchez and Sanchez (2005) recently developed a method based on discrete-valued Walsh functions. Their approach rapidly constructs very large two-level resolution V designs—a short program generates designs up to a $2^{120-105}$ in under a minute, and can be used to generate up to a $2^{443-423}$. A variant for generating two-level resolution VII designs for up to 100 factors is also available in Sanchez (2011).

Robust design is a system optimization and improvement process, pioneered by Taguchi (1987), that emphasizes the investigation of both mean performance and performance variation (see also HK2, chapter 17). Taguchi's construction of special Resolution III orthogonal arrays has been credited as the primary trigger for the “explosive use of design for product quality improvement in monitoring in industrial processes” (Hedayat, Sloane, and Stufken 1999). The philosophy extends naturally to simulation experiments (see e.g., Ramberg et al. 1991, Sanchez 2000, Kleijnen et al. 2005). The robust design approach seeks to identify systems that perform well across a range of uncontrollable sources of variation in its environment. Experimental designs that enable this *world view* treat the controllable factors differently from the uncontrollable, or noise, factors (see Chapter 13). Analysis approaches on the resultant data focus on the distributions of the responses rather than just the mean performance.

When the number of factors is very large, sequential screening approaches may be of interest. These typically make stronger assumptions about the nature of the response surface, but are useful for quickly eliminating unimportant factors so that future experiments can focus on those that are identified as important. Sequential screening procedures can be more efficient than single-stage screening procedures. Two procedures of particular interest are the *controlled sequential bifurcation* (CSB) procedure (Wan, Ankenman, and Nelson 2006) for estimating main effects, and a variant called CSB-X (Wan, Ankenman, and Nelson 2010) for estimating main effects in the presence of two-factor interactions. These procedures have the desirable property of providing guaranteed limits on the probabilities of observing false positives and false negatives when screening for important factors. Sequential approaches we find particularly useful for simulation experiments are *fractional factorial controlled sequential bifurcation* (FF-CSB) and a variant called FFCSB-X (Sanchez, Wan, and Lucas 2009), and the *hybrid method* (Shen, Wan, and Sanchez 2009). Although these methods are heuristic, they nonetheless have been shown to have very good properties in terms of both efficiency and effectiveness. Unlike CSB and CSB-X, these latter procedures do not require a priori knowledge of the direction of factor effects, which makes them suitable for screening factors in simulation models of complex systems where little subject matter expertise exists. Screening experiments are often followed up with more detailed experiments involving those factors identified as important.

Books that discuss experimental designs for simulation include Santner, Williams, and Notz (2003), Law (2007), and Kleijnen (2007). Note that their goals for those performing simulation experiments may differ from those in this chapter. Orthogonal arrays that can be used for qualitative factors appear in Hedayat, Sloane, and Stufken (1999); these are catalogued online at (<http://www2.research.att.com/~njas/oadir/>). Spreadsheets and software for our portfolio of designs, including newer designs not referenced in the above books, are available at (<http://harvest.nps.edu>). This site is regularly updated as new designs become available.

For experiments where it is very time consuming to run a single replication, there are other single-stage designs (often used for physical experiments) that require fewer runs than fractional factorial designs. Some of these designs appear in the above references; others can be found in experimental design texts, such as HK1, HK2, Box, Hunter, and Hunter (2005), Montgomery (2005), and Ryan (2007).

12.2.5 Implementing Large-Scale DOE

Given a simulation model, a design, and input data (factors and factor settings), running the experiments to generate the output data can be a nontrivial task. There are several ways that simulation experiments can be implemented. Unfortunately, many simulation environments are not set up to take advantage of the power of designed experiments.

Perhaps the most common approach is to launch the runs manually, for example, after changing the factor settings in a graphical user interface. This is a time-consuming and error-prone method and should be avoided if possible. A preferable approach is to use a computing script to automatically run the experiments according to a specified design and consolidate the output. This requires some programming expertise and may take more time initially, but the payoff is worthwhile when conducting more than a small number of runs. To harness today's computational power, running the experiments in parallel enables faster turnaround and/or a broader set of design points or iterative sets of experiments to be conducted. Software that takes advantage of cluster computing also requires more upfront work, but allows analysts to conduct large experiments even when each run is time consuming.

12.3 APPLICATION: U.S. ARMY UNMANNED AERIAL VEHICLE (UAV) MIX STUDY

If the Division Commanders want a UAV at their level and have nothing now, we ought to give it to them.

*LTG SCOTT WALLACE, Commanding General,
U.S. Army V Corps, cited in SINCLAIR (2005)*

12.3.1 Study Overview

In 2000, the U.S. Army initiated a comprehensive transformational modernization program known as the Future Combat System (FCS) (Feickert and Lucas 2009). Three multiyear, multibillion dollar investment paths were outlined to identify new technologies for current operations, determine flexible and rapidly deployable force structures, and develop and field forces with these new technologies. Each of these paths heavily relied on UAVs and their yet-to-be-determined capabilities (Office of the Secretary of Defense 2005). The Army Chief of Staff tasked the Training and Doctrine Command (TRADOC) to evaluate the new technologies and develop an investment strategy regarding UAVs. In this section, we present some of the findings from a study conducted in 2006 to support TRADOC's effort.

Separate working groups were tasked to address different components of this endeavor. Subject matter experts developed tactical scenarios that could be linked to UAV mission requirements combined with projected UAV performance capabilities. Our portion of this study involved a simulation tool called ASC-U (short for Assignment Scheduling Capability for UAVs; see Ahner, Buss, and Ruck 2006). ASC-U employs a discrete event simulation, coupled with the optimization of a linear objective function, to determine a schedule for UAV missions. This scheduling problem is not trivial. Planning had to take into account over 300 UAVs of five different types, trying to fulfill over 21,000 missions of 17 different types, within a 15-day period. As if the problem was not complex enough, the UAVs were launched and recovered from mobile sites that had a variety of control capabilities. Moreover, the missions had limited time windows in which they could be completed—and they were not known at the beginning of the simulation, but arose dynamically.

The ASC-U simulation tool is capable of modeling different scenarios with potentially unlimited varieties of UAVs. In this study, we evaluated five types of UAVs then under consideration by the Army that differed based on the platform's capabilities and operational requirements (*Defense Industry Daily* 2005). Platoons were to operate Class I airframes. These were lightweight, portable, hand-launchable UAVs with minimal range and endurance. The sizes and capabilities of the vehicles increased in subsequent classes, and the fifth type was a division-controlled asset known as the Extended Range Multipurpose (ERMP) UAV. The unmanned aircraft carry out a variety of missions, including reconnaissance, relaying communications, surveilling threats, and engaging targets.

ASC-U was originally developed as a planning tool for generating flight schedules. However, we showed that by leveraging state-of-the-art design of experiments, ASC-U could also be used as a screening tool to assess alternative UAV mixes. For details of this study, we refer the reader to Nannini (2006). We highlight key DOE issues and present some of the analysis in this section. The bottom line is that we found that the Army could reduce the projected

cost of the UAV fleet by billions of dollars without sacrificing performance (Bauman 2007).

12.3.2 Study Goals

We began with several specific goals. First, ASC-U requires user-specified parameters that control the optimization algorithm, and these parameters can influence both the simulation run time and the solution quality. Consequently, we sought to determine appropriate values for these parameters. Second, we were interested in identifying which UAV capabilities significantly influenced the performance measures. Third, we were interested in finding out whether there were any important thresholds, interactions, or nonlinear effects—the proverbial “knee in the curve.”

12.3.3 Experimental Setup

The study goals guided our choice of experimental design. Additionally, design considerations were driven by the following three issues.

1. *Analytic Flexibility.* We wanted to *develop a broad understanding* of this simulation with little or no prior knowledge about the nature of the response surfaces. Consequently, we needed a design with analytic flexibility with regard to a large and diverse set of responses and their fitted metamodels.
2. *Many Factors.* Prior to our study, a baseline ASC-U scenario had been developed that closely matched prototype UAVs and potential Army missions. This baseline scenario used single values for operating time, air speed, operating radius, and transition times based on current and future projected UAV capabilities (Unmanned Aircraft 2005). Unfortunately, while this could provide a point estimate of anticipated performance, there was no consideration of either random variation or uncertainty about settings for the input factors. Conducting a broad exploratory analysis allowed us to examine whether the stated requirements, such as minimum UAV speed or endurance, were appropriate. After much consideration, we identified 26 simulation and UAV performance factors requiring investigation. The factors consisted of the following: optimization interval, time horizons, air speed, operating time, operating radius, and transition time. While the optimization interval is a single value for the entire scenario, the other five factors are unique for each type of UAV. Consequently, we needed a design capable of accommodating more than two dozen factors.
3. *Multiple Responses.* One primary measure was the proportion of potential missions covered (partial coverage was permitted). Another key measure, mission value, was a weighted function of coverage based on

subjective assessments of mission importance. We also looked at other aggregate performance measures, such as the mission coverage by UAV type, UAV utilization, mission package utilization, coverage delay, coverage by task type, and utilization of the ground control stations. Consequently, we needed a design that permitted parallel analysis for many responses.

Taken together, these goals and issues led us to use space-filling NOLH designs (Cioppa and Lucas 2007). We used these in an iterative manner—as we typically do in large-scale simulation studies. Invariably, lessons learned from one set of experiments guide or shape subsequent experiments.

Early experiments included the optimization interval as a factor. This is a user-specified value that determines how often (in simulated time) the ASC-U software stops and attempts to reoptimize the existing UAV assignments. The optimization interval was set to one simulated hour (for convenience) during small-scale verification runs. Larger intervals would be desirable for practical considerations, in terms of reducing both the run time for creating schedules and the disruption caused by frequent schedule changes. It came as a great surprise to the model developers that the optimization interval completely dominated the UAV schedule and its quality: Mission value and the optimization interval had a correlation of 0.97. Moreover, mission coverage exceeded 77% in all cases when the optimization interval was 1 hour. In contrast, optimization intervals greater than 1 hour dramatically degraded mission coverage, dropping it to as low as 49%. Intervals larger than 1 hour also resulted in a lower correlation between the two primary performance measures (mission coverage and mission value). Consequently, in later experiments, we fixed the optimization interval at 1 hour.

Our final NOLH experiment had 257 design points. Since ASC-U is a deterministic simulation, we required only one run per design point. At approximately 3 hours per run, this called for over 700 hours of processing time. We employed approximately 60 2.8 GHz Pentium 4 computers to make runs in parallel, enabling overnight completion of the computational experiment. We augmented these data with some from earlier experiments, resulting in 272 design points. The maximum magnitude of pairwise correlations between columns in the design matrix was 0.0372.

12.3.4 Results

When dealing with such large designs, the analysis is more akin to data mining than to traditional ANOVA. Our primary analytical techniques are descriptive statistics, stepwise regression, and partition trees. We also use a variety of graphical techniques, including predication and interaction profilers, along with histograms, boxplots, scatter plots, contour plots, and parallel plots. The plots and model fits in this section were generated using JMP, which is an interactive statistics package developed by the SAS corporation (JMP, Version

8 1989–2009). Our intent is to highlight several types of statistical and graphical approaches that we have found useful for assessing the results of large-scale simulation experiments. Details about the background, systems, analysis, and implications can be found in Nannini (2006).

12.3.5 Descriptive Statistics

We begin by presenting histograms of the performance measures. JMP histograms are accompanied by a box-and-whisker plot to help identify outliers. The bracket alongside the box-and-whisker plot identifies the range spanning the densest 50% of the data. Figure 12.2 shows that many of the performance

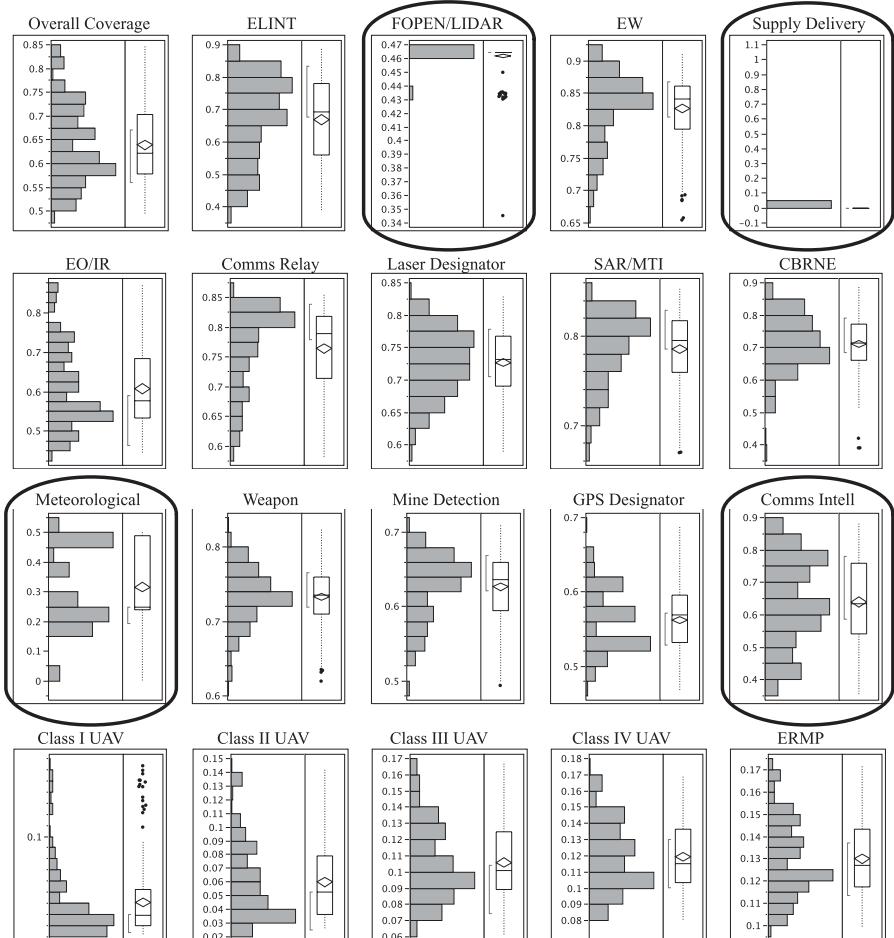


Figure 12.2 Histograms of 20 response variables for Mission Coverage.

measures appear relatively bell-shaped and symmetric, with the baseline scenario falling in the center of the data. However, several measures have multi-model distributions. Of particular interest are the coverage distributions for the meteorological and communication intelligence payloads. Two other measures demonstrate minimal to no variation in their coverage. The measure for the foliage penetrating radar/light detection and ranging (FOPEN/LIDAR) payload centered on 46.4% coverage, with little variation. Surprisingly, the simulation solution did not allocate unmanned aircraft to any supply delivery missions.

12.3.6 Interactive Regression Modeling

Our process consisted of a mixed stepwise technique, followed by standard least squares regression. We inspected trends in first order, first order with interactions, and second-order polynomial models. We evaluate the step history using the *F*-test to evaluate the significance of each term added to the model and construct a final model consisting of nine main effects, one interaction term, and one second-order polynomial term. As Figure 12.3 shows, we quickly reach a point where adding statistically significant terms has essentially no practical impact on the model fit. As we mentioned earlier, this can easily occur when there are large data sets, so it is important not to rely on statistical significance alone. In order to facilitate interpretation and explanation, parsimonious models are desirable.

Figure 12.4 displays the actual value by predicted plot and summary data for the final model consisting of nine main effects, one two-factor interaction, and one second-order polynomial. The 11 terms account for 92% of the variance within the model (see also Fig. 12.3). Two predictors accounted for nearly 68% of the variance within the model: Class IV operating radius at 53% (see also Fig. 12.3), and Class I operating time for another 15%. All of the factors were highly significant (p -value < 0.0001).

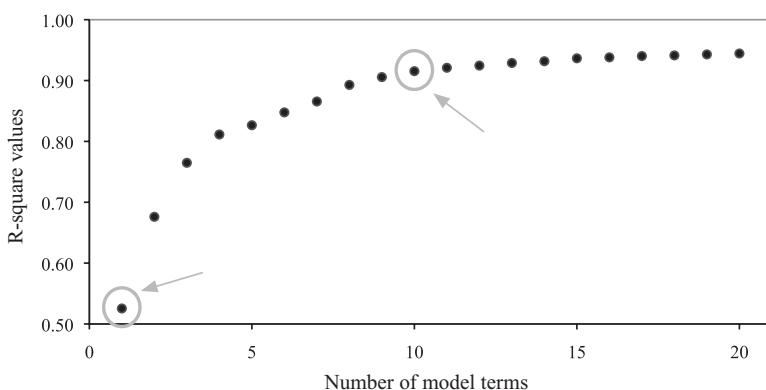


Figure 12.3 Diminishing returns in R^2 as terms are added to the fitted model for Mission Value.

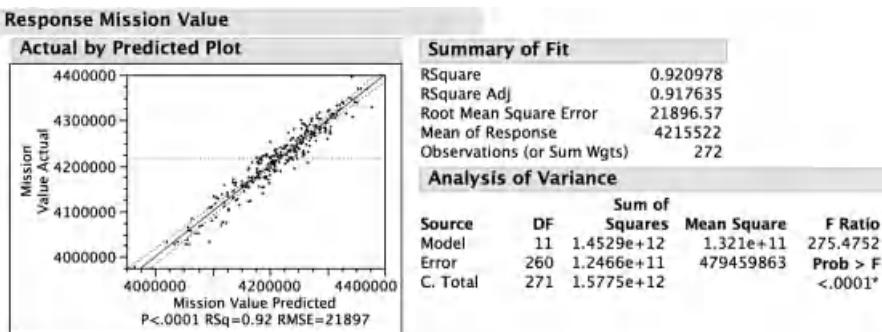


Figure 12.4 Final fitted model for Mission Value.

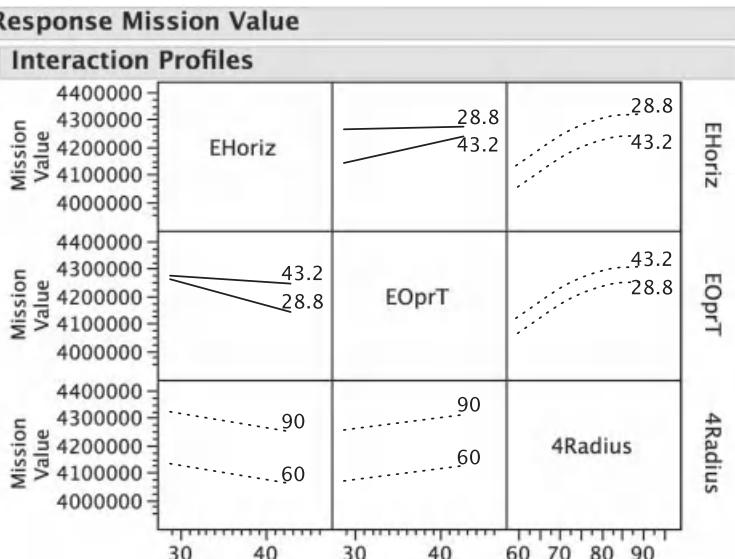


Figure 12.5 Interaction profiler for final fitted model of Mission Value.

The single interaction term in our final fitted model involves the time horizon and operating time for the ERMP, the largest of the UAVs. Figure 12.5 is a portion of the interaction profiler generated by JMP. In this plot, the *y*-axis represents the mission value, and the *x*-axes represent each term's range. The smaller plots display the effects of pairs of factors on the response. JMP displays solid curves for interaction terms that are included in the model, and the magnitude of the interaction is reflected by how far from parallel the curves are. These plots are useful tools for conveying the meaning of interactions to decision makers.

Looking at the middle row, we see ERMP time horizon crossed with ERMP operating time on the center-right plot. The top line represents the highest

mission values when the horizon time is set to the ERMP UAV's lowest level of 28.8 hours. This finding seems counterintuitive. It implies that when planning for ERMP UAV missions we should be myopic in our approach and consider a set of mission areas that become available sooner rather than a larger set of potential mission areas over a broader time window. Upon further analysis, we found that this process could be explained in the following way. If the ERMP time horizon is greater than 28.8 hours, the optimization process attempts to allocate the UAV resource to missions of greater value within the extended future time horizon. In doing so, the process skips available missions of lesser value that are closer to the current simulation time. The cumulative and potentially larger payoff generated from serving multiple missions is lost.

Additional insights gained from Figure 12.5 involve the combined effects of Class IV operating radius and ERMP time horizon. The center-top plot indicates that relatively high mission values are obtained when the Class IV operating radius is set at its maximum level of 90 km and ERMP time horizon is set at its lowest level of 28.8 hours. However, we also observe diminishing returns for the effect of Class IV operating radius on mission value. The center-left plot on the interaction profile indicates this effect. When ERMP time horizon is set to its lowest level of 28.8 hours, the effect of Class IV operating radius on mission value diminishes around 85 km, when all other effects remain constant.

The insights gained from exploring the interaction between ERMP time horizon and operating time, Class IV operating radius, and Class I operating time are significant. In the scenario examined by our study, the results indicate that overall mission success can be exploited by a relatively small ERMP time horizon and platform operating time. The next section develops an alternate regression model using mission coverage as the response variable.

The regression model developed for mission value provides an initial assessment to the solution of ASC-U. However, it is difficult to relate the numerical value provided by this measure to a tactical measure of effectiveness that military commanders can apply in an operational environment. Therefore, we developed a regression model for the overall mission coverage for the scenario.

Our analysis for the mission coverage followed the same path as the development process for the mission value model described previously. We evaluated the history generated in the stepwise regression of the full quadratic model in order to construct a final model consisting of ten main effects and one second-order polynomial. The actual value by predicted plot and the statistical report for the regression model for mission coverage is displayed in Figure 12.6.

All of the factors within our final model for mission coverage were significant (p -value < 0.0001). The 11 terms account for 92% of the variance within the model. Two predictors accounted for nearly 74% of the variance within the model: Class I operating time at 54%, and Class IV operating radius for another 20%. This result was opposite that of the mission value model, which

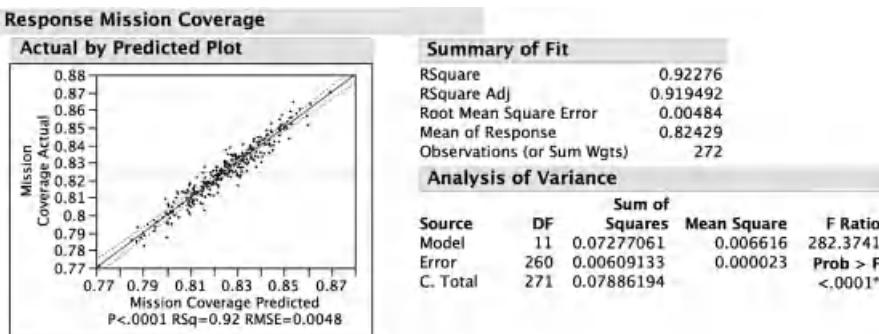


Figure 12.6 Final fitted model for Mission Coverage.

had Class IV operating radius accounting for 53% and Class I operating time for another 15%. This finding indicates that mission value does not necessarily depict equivalent coverage.

Initially, mission value and mission coverage seem to imply a similar measure of effectiveness. Mission value is derived from a value rate that is dependent on the mission area and required sensor package. This “weighting” system may result in a scenario with a mission value that does not correspond to equivalent mission coverage. For example, mission areas with more valuable sensor package requirements may be served in such a way as to limit multiple servicing of other mission areas. The cumulative effect results in a high mission value, but relatively low mission coverage. This was an important finding for the user community of ASC-U. While the simulation developers recognized that ASC-U demonstrates consistency “characterized by increases in UAV performance resulting in corresponding increases in mission coverage and overall mission value,” (Ahner, Buss, and Ruck 2006), understanding the difference between the two measures helps the user community evaluate the impact of their assigned value rates relative to mission areas and aircraft payloads.

12.3.7 Regression Trees

We explored the significant factors further using JMPs partitioning platform. JMP allows the analyst to generate regression trees as a method of exploratory modeling. Regression trees employ a binning and averaging process. The software’s algorithm evaluates all of the predictor values in order to determine the optimum split in the tree. The predictor partition value that generates the highest reduction in total sum of squares is selected to create a new branch in the tree (Sall, Creighton, and Lehman 2005).

The regression tree in Figure 12.7 indicates that when Class III aircrafts have an operating radius of greater than or equal to 41.9 km, the mean coverage for meteorological missions is increased by 16% for the scenario used in

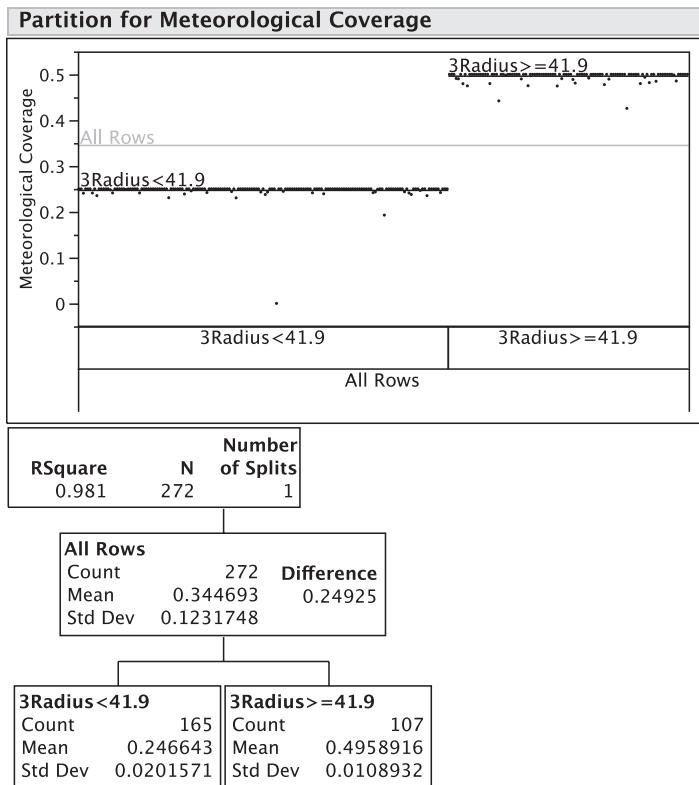


Figure 12.7 Regression tree for Meteorological Coverage.

the study. This also corresponded to nearly a doubling of the meteorological coverage of the baseline scenario.

The regression tree in Figure 12.8 displays the first three splits for communication intelligence coverage. The third split on the right is represented by ERMP time horizon ≥ 36.3 hours and falls under Class IV operating radius ≥ 73.4 km. The result is a 2% increase to communication intelligence coverage, relative to the baseline scenario. This corresponds to 9% of the variation of the performance measure.

Similar trees were constructed for each of the other performance measures. Displaying details of this information for multiple performance measures in a consolidated manner is difficult, so we focused on the initial splits. Specifically, we identified the significant factor, the level of the factor identifying the branch with the increased mean, the percent increase over the range of the examined performance measure, and the percent increase in coverage. Table 12.1 displays the results of the exploration.

The most significant increase in percent coverage occurred for the meteorological sensor package, with a 15.12% increase to the mean when Class III

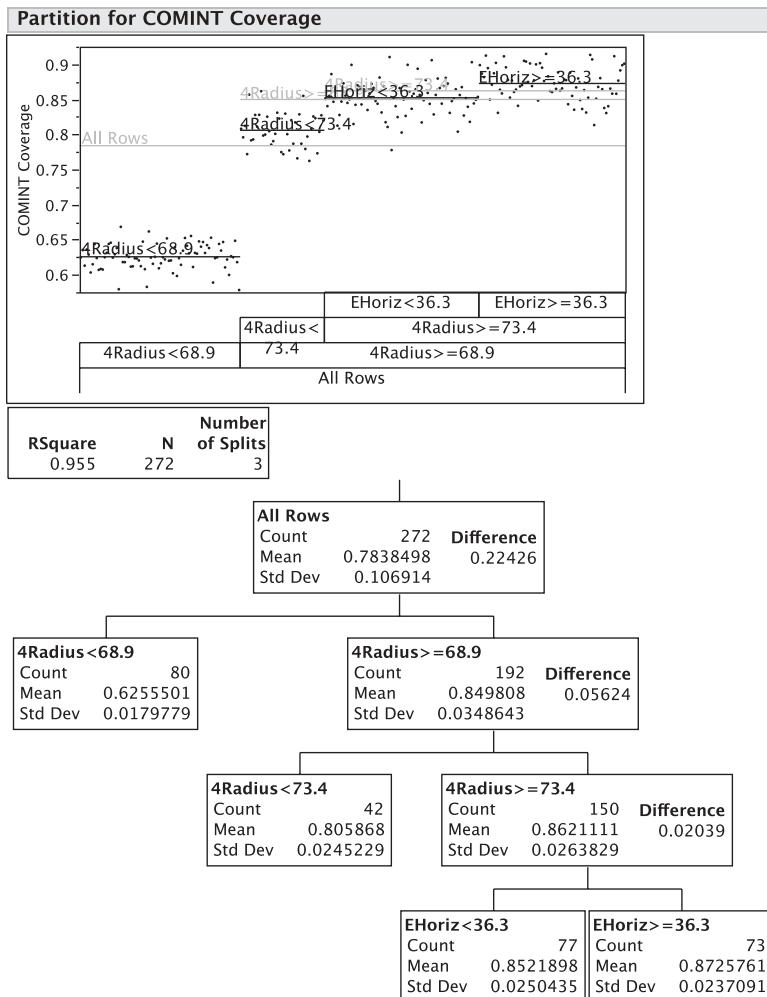


Figure 12.8 Regression tree for Communication Intelligence Coverage.

operating radius is set at greater than or equal to 41.9 km. This result corresponds to a 30.24% increase over the entire range of percent coverage for meteorological missions. While the other splits did not cause a substantial increase to the mean of the other performance measures, we note that five splits resulted in a 15% or greater increase over the range of the measure studied. Class IV operating radius and Class I operating time with seven splits each represented the largest number of splits for the 21 performance measures. Class III operating radius produced one split. It was the most significant factor to influence the mean of meteorological percent coverage. Interestingly, Class I operating time was the most significant factor for percent coverage by type

Table 12.1 Impacts of First Splits in Regression Trees for Multiple Responses

Response	Factor in First Split of Regression Tree for Mission Value	Base Scenario Factor Setting	Partition Creating Positive Increase on Mean Mission Value	Percent Increase Over Range of Response	Associated Percent Increase in Coverage
Mission value	4Radius	75 km	≥68.9 km	8.21	0.82
Overall coverage	10prT	0.830 hours	≥0.787 hour	8.28	N/A
Supply delivery	N/A	N/A	N/A	N/A	N/A
Meteorological	3Radius	40 km	≥41.9 km	30.24	15.12
COMINT	4Radius	75 km	≥68.9 km	19.59	6.60
Weapon	EoprT	36 hours	≥38.8 hours	15.54	3.33
ELINT	4Radius	75 km	≥68.9 km	15.43	3.82
Class I UAV	10prT	0.83 hour	≥0.816 hour	11.31	1.06
Class II UAV	10prT	0.83 hour	≥0.816 hour	14.36	0.50
Class III UAV	10prT	0.83 hour	≥0.894 hour	13.00	0.57
Class IV UAV	10prT	0.83 hour	≥0.789 hour	10.38	0.21
ERMP	10prT	0.83 hour	≥0.787 hour	7.35	0.16
GPS designator	Ehoriz	36 hours	<30.7 hours	29.16	6.11
Laser designator	4Radius	75 km	≥77.6 km	15.01	1.70
EO/IR/LR	10prT	0.83 hour	≥0.770 hour	9.41	1.02
SAR/MTI	4Radius	75 km	≥67.6 km	7.29	0.65
Comms relay	Ehoriz	36 hours	<36.7 hours	6.85	0.60
EW	EoprT	36 hours	≥32.3 hours	6.35	1.37
Mine detection	4Radius	75 km	≥66.2 km	5.90	0.61
CBRNE	4Radius	75 km	≥64.2 km	5.27	1.47
FOPEN/LIDAR	EoprT	36 hours	<42.9 hours	1.63	0.03

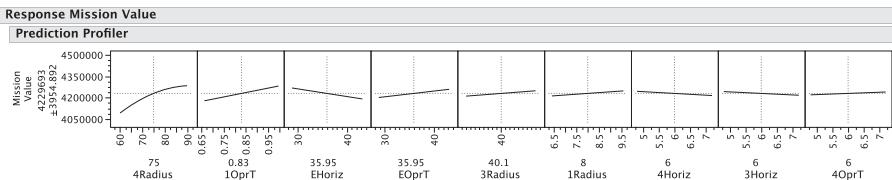


Figure 12.9 Prediction Profiler for final model of Mission Value.

of all the UAVs. Finally, no UAVs were ever allocated to supply delivery missions. While this is interesting, there were only five such missions in the entire scenario, so the analyst should be cautious about its significance.

12.3.8 Other Useful Plots

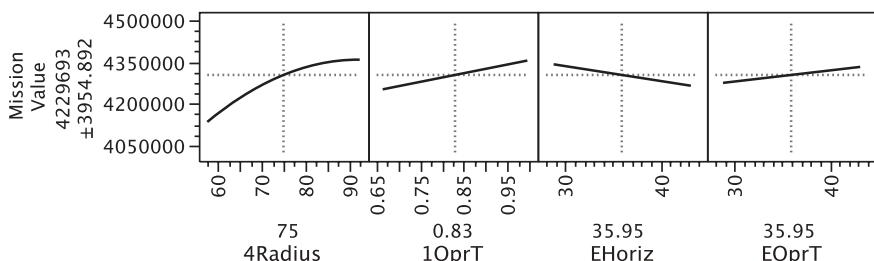
A Prediction Profiler appears in Figure 12.9. Each subplot shows the effect of changing a single factor's value while holding the rest constant. The steepness of the curves indicate that Class IV operating radius has the strongest effect on mission coverage. Class I operating time, the ERMP operating time, and the ERMP time horizon demonstrate lesser effects. The remaining main effects display a relatively weak effect on percent coverage, although all these terms are, nonetheless, highly significant (p -values < 0.0001). Plots like these are often more informative to decision makers than lists of the fitted regression parameters.

The Prediction Profiler can also be used interactively, allowing the analyst to set the factor values to other ranges within their original ranges and observe the effects. To illustrate, portions of the Prediction Profiler for two different combinations of factor settings appear in Figure 12.10. On the left hand side appears a point estimate and half-interval width for predicting the expected Mission Value using the fitted model. If interactions are present, then the slopes of the lines or curves in individual subplots will shift accordingly. For example, the subplots showing the marginal effects of the ERMP time horizon and the ERMP operating time are much flatter in the lower plot than in the upper plot.

We also produced a contour plot (Figure 12.11) to show the joint effect of Class I operating time and Class IV operating radius on mission coverage. The midpoints of each range are the baseline settings for the scenario. Filled contour regions depict the mission coverage as the two factors vary. Note that because the data include variations in all other factors, these contour plots will show more variation than contour plots based on varying two factors while holding all others constant. Region 1 corresponds to mission coverage below 80.0%, which is considered undesirable. Region 2 shows that it is possible to achieve 80% or better coverage when both factors are set below the base scenario settings. Region 3 shows that coverage above 84% can be achieved

Response Mission Value

Prediction Profiler



Response Mission Value

Prediction Profiler

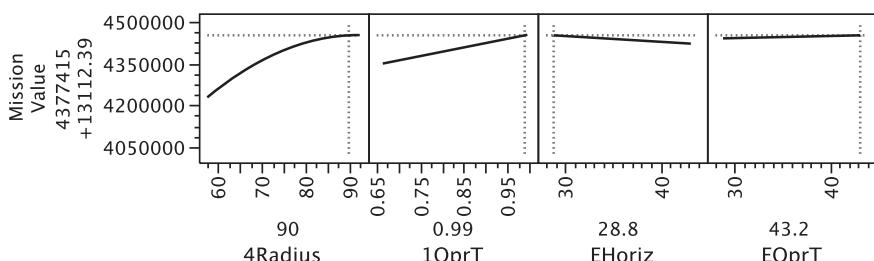


Figure 12.10 Portions of the Prediction Profiler for final model of Mission Value, two different settings of factor levels.

Contour Plot for Mission Coverage

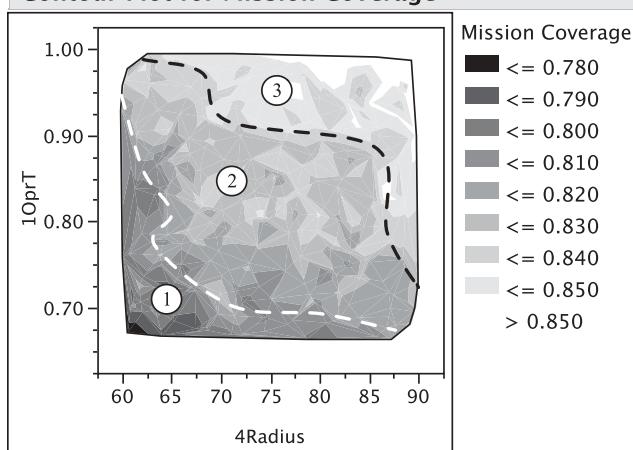


Figure 12.11 Contour plot of mission coverage by Class I operating time and Class IV operating radius.

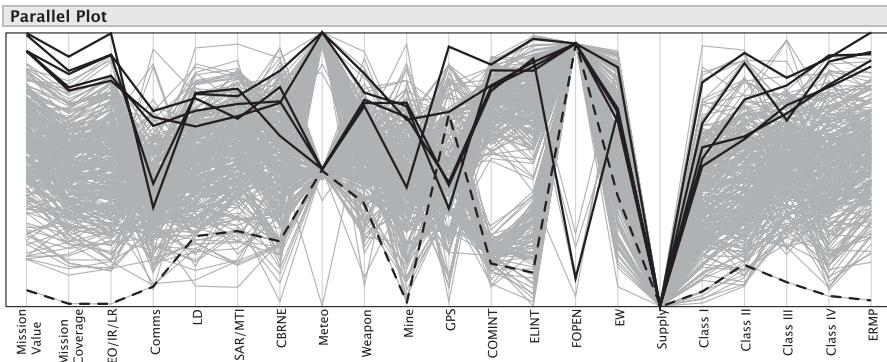


Figure 12.12 Parallel plot of 21 response variables, with traces highlighted for six design points.

(in some cases) even if one factor is below its baseline setting, provided that the other is sufficiently high above. In other words, it is not necessary for both these factors to be at high levels in order to achieve effective coverage. This suggests that trade-offs can be made. In this particular example, increasing the Class I operating time by only a few minutes may allow suitable coverage to be achieved while maintaining (or even reducing) the Class IV operating radius. Insights such as these may be of interest to program managers trying to properly scope platform requirements during system acquisition programs.

Parallel plots provide a visual representation of the association between multiple variables. Figure 12.12 displays a parallel plot to illustrate the benefits of reviewing more than one or two performance measures. The plot allows us to explore the relative results of all of the measures simultaneously. Here, 272 individual lines connect the response values for each performance measure according to the 272 different design points. Traces for the design points associated with the five highest mission values are highlighted with solid black lines (points 11, 14, 43, 85, and 253); the dotted black line shows the trace for design point 140, which has the lowest mission value. We note that while the design points with solid black lines represent the highest mission values within our DOE, they result in moderate or low values for other performance measures. For example, design points 85 and 253 are separated from design points 11, 14, and 43 within the bimodal distribution of Meteo, the meteorological coverage. Additionally, design point 43 represents one of the lowest outcomes for FOPEN coverage. The dotted line shows that design point 140 results in one of the highest FOPEN coverages, and also does relatively well with regard to the GPS response, even though it has the lowest overall mission value. Again, we point out that all design points result in zero coverage for Supply (the supply delivery mission coverage). Parallel plots can be expanded to include factors, as well as performance measures, which may help the analyst visually identify interesting combinations of factor levels. A nice feature with the JMP

implementation of parallel plots is the ability to interactively select rows to highlight.

12.3.9 Summary

The UAV modeling . . . harvested \$6 billion in savings and 6000 to 10,000 billets, that's a brigade's worth of soldiers. Over 20 years that allowed us to avoid a cost of \$20 billion.

MICHAEL F. BAUMAN (2007), Director of the U.S. Army Training and Doctrine Command Analysis Center.

Budget overruns, technical challenges, and ongoing operations eventually forced Secretary of Defense Robert Gates to formally cancel the Future Combat Systems program in 2009 (Office of the Secretary of Defense 2009). However, many of the technologies have been and are being folded into the Army and other defense and national security organizations (Osborn 2009). Indeed, UAVs are becoming omnipresent on the battlefield, being procured in record numbers, and there are ambitious plans for their expanded use in the future. Moreover, many of the recommendations from this study were enacted, such as the cancellation of the Class II and Class III UAVs. In addition, this study and others like it have demonstrated to senior leaders the potential of DOE to assist analysts in quickly and efficiently assessing procurement options. The Department of Defense spends enormous sums of money building and experimenting with simulations to assist in making, explaining, and defending choices for equipping our men and women in uniform. Doing it better saves both money and lives.

In this chapter, we have encapsulated the results of one specific defense application, but we could have shown dozens more. For example, three national security studies are summarized in Lucas et al. (2007). The first example explores equipment and employment options for protecting critical infrastructure, the second case considers nonlethal weapons within the spectrum of force-protection options in a maritime environment, and the final application investigates emergency (police, fire, and medical) responses to an urban terrorist attack. The results illustrate that valuable insights into defense and homeland security operations can be gained by the use of well-designed experiments.

12.4 PARTING THOUGHTS

One test is worth a thousand expert opinions.

BILL NYE, the Science Guy

Experiments are at the core of science, yet even within the scientific community, there is often a lack of awareness regarding the power of experimental

designs—especially designs developed for high-dimensional simulation experiments. Without this knowledge, analysts are likely to find themselves in one of a few undesirable situations. First, if they explore their models haphazardly, rather than using designed experiments, they may miss important insights or draw incorrect conclusions. Second, analysts who try to use brute-force computation to investigate a complex model will find themselves surprisingly limited in the number of factors and number of levels they are able to explore, even if they have massive computing resources available. Third, analysts who are familiar with performing physical experiments may unnecessarily restrict themselves to designs more suitable for situations involving a small to moderate number of factors and simple response surfaces. This can be problematic because the insights obtainable by analyzing experimental data depend critically on the design.

We believe that a paradigm shift is on its way, but that it cannot be fully realized without integrating state-of-the-art experimental designs, high-performance computing, new modeling environments, and innovative analytical techniques to gain deeper insights from simulation models. The interaction of these technologies offers new and exciting opportunities for both theoretical and applied statisticians. There will continue to be a need for statisticians to develop new experimental design approaches, along with data-analytic and graphical methods, that facilitate the analysis of complex systems, and to educate others about the benefits and capabilities of designed experiments. There will also be new opportunities for statisticians to apply experimental design to important and interesting problems that were beyond our capabilities only a few short years ago.

REFERENCES

- Ahner, D.K., A.H. Buss, and J. Ruck (2006). Assignment scheduling capability for unmanned aerial vehicles—A discrete event simulation with optimization in the loop approach to solving a scheduling problem. In: *Proceedings of the 2006 Winter Simulation Conference*, L.F. Perrone, F.P. Wieland, J. Liu, B.G. Lawson, D.M. Nicol, and R.M. Fujimoto (eds.). Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc., pp. 1349–1356.
- Bauman, M.F. (2007). Speech, Naval Postgraduate School, March 1.
- Box, G.E.P., W.G. Hunter, and J.S. Hunter (2005). *Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building* (2nd ed.). New York: Wiley.
- Cioppa, T.M. and T.W. Lucas (2007). Efficient nearly orthogonal and space-filling Latin hypercubes. *Technometrics*, **49**(1), 45–55.
- Defense Industry Daily (2005). Four FCS UAV sub-contracts awarded (updated). Available via <<http://www.defenseindustrydaily.com/2005/10/four-fcs-uav-subcontracts-awarded-updated/index.php>> [accessed August 15, 2010].
- Fedorov, V.V. (1972). *Theory of Optimal Experiments*. New York: Academic Press.

- Feickert, A., and N. J. Lucas (2009). Army future combat system (FCS) “spin-outs” and ground combat vehicle (GCV): Background and issues for Congress (RL32888). Available via <<http://www.fas.org/sgp/crsweapons/RL32888.pdf>> [accessed August 15, 2010].
- Goldsman, D., S.-H. Kim, and B.L. Nelson (2005). Statistical selection of the best system. In: *Proceedings of the 2005 Winter Simulation Conference*, M.E. Kuhl, N.M. Steiger, F.B. Armstrong, and J.A. Joines (eds.). Piscataway, NJ: Institute of Electrical and Electronic Engineers, Inc., pp. 178–187.
- Hammersley, J.M. and D.C. Handscomb (1964). *Monte Carlo Methods*. London & New York: Chapman and Hall.
- Hedayat, A.S., J. Sloane, and J. Stufken (1999). *Orthogonal Arrays: Theory and Applications*. New York: Springer-Verlag.
- Hernandez, A.S., Lucas, T.W., and M. Carlyle (2011). Enabling nearly orthogonal Latin hypercube construction for any non-saturated run-variable combination. Working paper, Operations Research Department, Naval Postgraduate School, Monterey, CA.
- Hinkelmann, K. and O. Kempthorne (2008). *Design and Analysis of Experiments, Vol. 1: Introduction to Experimental Design* (2nd ed.). vol. 1. Hoboken, NJ: John Wiley & Sons.
- Hoeber, F.P. (1981). *Military Applications of Modeling: Selected Case Studies*. Alexandria, VA: Military Operations Research Society.
- JMP, Version 8 (1989–2009). SAS Institute Inc., Cary, NC.
- Kleijnen, J.P.C. (2007). *Design and Analysis of Simulation Experiments*. New York: Springer.
- Kleijnen, J.P.C., S.M. Sanchez, T.W. Lucas, and T.M. Cioppa (2005). A user’s guide to the brave new world of designing simulation experiments. *INFORMS Journal on Computing*, **17**(3), 263–289.
- Law, A.M. (2007). *Simulation Modeling and Analysis* (4th ed.). New York: McGraw-Hill.
- Lucas, T.W., S.M. Sanchez, F. Martinez, L.R. Sickinger, and J.W. Roginski (2007). Defense and homeland security applications of multi-agent simulations. In: *Proceedings of the 2007 Winter Simulation Conference*, S.G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J.D. Tew, and R.R. Barton (eds.). Piscataway, NJ: Institute of Electrical and Electronic Engineers, Inc., pp. 138–149.
- Markoff, J. (2008). Military supercomputer sets record. *New York Times*. Available via <<http://www.nytimes.com/2008/06/09/technology/09petaflops.html>> [accessed August 14, 2010].
- Mitroff, I.I. and T.R. Featheringham (1974). On systemic problem solving and the error of the third kind. *Behavioral Science*, **19**(6), 383–393.
- Montgomery, D. (2005). *Design and Analysis of Experiments* (6th ed.). Hoboken, NJ: Wiley.
- Nannini, C.J. (2006). Analysis of the Assignment Scheduling Capability for UAVs (ASC-U) simulation tool. Master’s thesis, Naval Postgraduate School, Monterey, CA.,
- NIST/Sematech (2006). e-Handbook of Statistical Methods. Available via <<http://www.itl.nist.gov/div898/handbook>> [accessed August 8, 2010].

- Office of the Secretary of Defense (2005). Unmanned aircraft system roadmap 2005–2030. Available via <<http://www.fas.org/irp/program/collect/uavroadmap2005.pdf>> [accessed August 15, 2010].
- Office of the Secretary of Defense (2009). Future Combat System (FCS) program transitions to Army Brigade Combat Team Modernization. Available via <<http://www.defense.gov/Releases/Release.aspx?ReleaseID=12763>> [accessed August 15, 2010].
- Osborn, K. (2009). FCS is dead; programs live on: U.S. Army to dissolve flagship acquisition effort. *Defense News*. Available via <<http://www.defensenews.com/story.php?i=4094484&c=FEA&s=CVS>> [accessed December 10, 2010].
- Ramberg, J.S., S.M. Sanchez, P.J. Sanchez, and L.J. Hollick (1991). Designing simulation experiments: Taguchi methods and response surface metamodels. In: *Proceedings of the 1991 Winter Simulation Conference*, B.L. Nelson, W.D. Kelton, and G.M. Clark (eds.). Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc., pp. 167–176.
- Ryan, T.P. (2007). *Modern Experimental Design*. Hoboken, NJ: Wiley.
- Sacks, J., W.J. Welch, T.J. Mitchell, and H.P. Wynn (1989). Design and analysis of computer experiments (includes comments and rejoinder). *Statistical Science*, **4**, 409–435.
- Saeger, K. and J. Hinch (2001). Understanding instability in a complex deterministic combat simulation. *Military Operations Research*, **6**(4), 43–55.
- Sall, J., L. Creighton, and A. Lehman (2005). *JMP Start Statistics: A Guide to Statistics and Data Analysis using JMP and JMP IN Software* (3rd ed.). Belmont, CA: Brooks/Cole-Thomson Learning.
- Sanchez, P.J. (2011). Efficient generation of resolution VII fractional factorial designs. Working paper.
- Sanchez, S.M. (2000). Robust design: Seeking the best of all possible worlds. In: *Proceedings of the 2000 Winter Simulation Conference Winter*, J.A. Joines, R.R. Barton, K. Kang, and P.A. Fishwick (eds.). Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc., pp. 69–76.
- Sanchez, S.M. and T.W. Lucas (2002). Exploring the world of agent-based simulation: Simple models, complex analyses. In: *Proceedings of the 2002 Winter Simulation Conference*, E. Yuc̄es̄an, C. Chen, J.L. Snowdon, and J. Charnes (eds.). Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc., pp. 116–126.
- Sanchez, S.M. and P.J. Sanchez (2005). Very large fractional factorial and central composite designs. *ACM Transactions on Modeling and Computer Simulation*, **15**(4), 362–377.
- Sanchez, S.M., H. Wan, and T. Lucas (2009). Two-phase screening procedure for simulation experiments. *ACM Transactions on Modeling and Computer Simulation*, **19**(2), 1–24.
- Santner, T.J., B.J. Williams, and W.I. Notz (2003). *The Design and Analysis of Computer Experiments*. New York: Springer-Verlag.
- Shen, H., H. Wan, and S.M. Sanchez (2009). A hybrid method for simulation factor screening. *Naval Research Logistics*, **57**, 45–57.
- Sinclair, E. J., Brigadier General. (2005). Presentation at the 2005 UAVS Symposium. Available via <<http://www.quad-a.org/Symposiums/UAV-05/Presentations/>> [accessed June 6, 2006].

- Taguchi, G. (1987). *System of Experimental Design*, vol. 1 and 2. White Plains, NY: UNIPUB/Krauss International.
- Tukey, J.W. (1986). *The Collected Works of John W. Tukey. In Philosophy and Principles of Data Analysis: 1965–1986*, vol. IV. L.V. Jones (ed.) Monterey, CA: Wadsworth & Brooks/Cole, p. 773.
- Vieira, H., S.M. Sanchez, K.H. Kienitz, and M.C.N. Belderrain (2011). Generating and improving orthogonal designs by using mixed integer programming. *European Journal of Operation Research*, **215**, 629–638.
- Wan, H., B. Ankenman, and B. Nelson (2006). Controlled sequential bifurcation: A new factor-screening method for discreteevent simulation. *Operations Research*, **54**, 743–755.
- Wan, H., B. Ankenman, and B.L. Nelson (2010). Improving the efficiency and efficacy of controlled sequential bifurcation for simulation factor screening. *INFORMS Journal on Computing*, **22**(3), 482–492.
- Ye, K.Q. (1998). Orthogonal column Latin hypercubes and their application in computer experiments. *Journal of the American Statistical Association*, **93**, 1430–1439.

Robust Parameter Designs

Timothy J. Robinson and Christine M. Anderson-Cook

Some of the essentials of robust parameter design (RPD) are discussed in HK2 (chapter 17). The context for this class of problems considers designed experiments in order to understand the influence and role of factors with different levels of control in their final intended environment, production. The basic problems considered in this chapter are how to understand the influence of noise factors, those factors that cannot be fully controlled during production, on the response, and how to mitigate their impact through strategic choices of the control factors, those factors whose levels can be set and controlled during production. There are three divergent approaches to this problem—each with their own associated design classes and corresponding analyses. We outline the key characteristics of the designs and analyses for each approach, and then discuss how computer-generated designs can be helpful to expand the class of available designs for standard and nonstandard situations.

13.1 INTRODUCTION

As with other applications of designed experiments, understanding the influence of various factors on both the mean and variance of the response are important considerations. For RPD, there are quite specific goals for each of these aspects. Typically for the mean, the goal is to obtain response values as close to a desired target value as possible, where this optimization might involve a particular value or maximizing or minimizing the response. For the variance, we are interested in reducing its impact from environmental factors, differences in raw materials, or usage by the consumer. Hence, RPD problems naturally have a dual focus—optimizing the mean of the response

Design and Analysis of Experiments: Special Designs and Applications, First Edition. Edited by Klaus Hinkelmann.

© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

and minimizing the variance of the response around that target in the presence of uncontrollable factors during production that inflate the variance.

RPD was introduced and popularized in the United States during the 1980s by Genichi Taguchi (see Taguchi 1986). Taguchi (1986) postulated that there are two types of factors which operate on a system: control factors (the set of which is commonly denoted by \mathbf{x}) and noise factors (the set of which is commonly denoted by \mathbf{z}). The control variables are factors whose levels remain unchanged in the process once they are set. Noise variables, on the other hand, are factors whose levels change randomly within the process and may cause unwanted variation in the response, y . Examples of noise factors include environmental variables such as temperature and humidity, raw material suppliers, tolerances in the control factors, consumer usage, and product aging. The goal of RPD is to design the process in such a way as to operate at levels of the control variables (i.e., parameters) that make the process as insensitive as possible to the random fluctuations in the noise factors. While proper experimental design is essential in RPD, it is important to keep in mind that RPD refers to the designing of a process and not simply to a designed experiment.

The three main approaches to solving this problem are:

1. *The Taguchi-based approach*, which constructs an optimality criterion that combines contributions from the mean and variance into a single performance measure known as the signal-to-noise ratio. With this approach, the influence of the mean and variance are mixed into a single number to be optimized. Historically, this was the first method proposed, but because the optimization of this single criterion is difficult given the mixed contributions from the mean and variance, it is generally not considered one of the preferred approaches.
2. *The dual model response surface approach*, which develops separate responses for both the mean and the variance, and then uses response surface methodology approaches to simultaneously optimize these two objectives. This approach incorporates a level of flexibility that allows the user to select the relative importance of these two objectives of the problem and weigh them accordingly.
3. *The single model response surface approach*, which uses a single model based on the mean of the response, and then looks at contributions to being on target and the propagation of the variance from noise factors through terms in the model. This approach is unified in that the optimization stems from a single model, using different ways depending on the emphasis of the mean and variance components.

Since the approaches consider the RPD problem differently, the nature of the designs to estimate the mean and variance contributions are quite distinct. We now consider each of the methods in more detail, and identify common choices of associated designs, which are well suited to the requirements of the analysis.

Table 13.1 Engel Data from HK2, Example 17.5

A	B	C	D	E	F	G	% Shrinkage for Noise Factors (M, N, O)			
							-1	-1	1	1
-1	-1	-1	-1	-1	-1	-1	-1	1	-1	1
-1	-1	-1	1	1	1	1	0.3	2.5	2.7	0.3
-1	1	1	-1	-1	1	1	0.5	3.1	0.4	2.8
-1	1	1	1	1	-1	-1	2.0	1.9	1.8	2.0
1	-1	1	-1	1	-1	1	3.0	3.1	3.0	3.0
1	-1	1	1	-1	1	-1	2.1	4.2	1.0	3.1
1	1	-1	-1	1	1	-1	4.0	1.9	4.6	2.2
1	1	-1	1	-1	-1	1	2.0	1.9	1.9	1.8

13.2 TAGUCHI SIGNAL-TO-NOISE RATIO APPROACH

Taguchi's approach to RPD exploits the use of a crossed array experiment. We illustrate the crossed array with data from Engel (1992, see also HK2, example 17.5). Although the data in HK2 appeared within the context of SAS code in table 17.2, it is helpful to view it in a crossed format in Table 13.1. The experimental design in the control factors (factors A–G) is a 2^{7-4} factorial design and the experimental design in the noise factors (factors M–O) is a 2^{3-1} factorial design. Note that the *crossing* of these two orthogonal designs implies that every combination of the noise factors occurs with every combination of the control factors. A critical assumption in Taguchi's philosophy is that an appropriate response, reflecting process quality, can be chosen in such a manner as to minimize the impact of control-by-noise interactions.

By assuming that the process mean and variance at the i^{th} setting of the control factors is the same across the level combinations of the noise factors (i.e., no control-by-noise interaction), one can view the set of responses at each control factor setting as *replicates* to be used for process variance estimation. Taguchi proposed combining the sample mean and sample variance at each control factor setting into a single performance measure known as the *signal-to-noise* ratio (SNR). The formulation of the SNR is based on the notion of squared error loss,

$$E_z(y-T)^2 = \text{Var}(y; \mathbf{x}, \mathbf{z}) + [\mu(y; \mathbf{x}) - T]^2, \quad (13.1)$$

with T representing the targeted value of the response mean, and E_z denoting the fact that expectation is taken over the noise factor settings. While there

are many possible SNRs, the various formulations depend upon the goal of the experiment. If, for example, the goal is to minimize the process mean and $T = 0$, the squared error loss function reduces to $E_z(y - 0)^2$ and one might seek to maximize

$$SNR_s = -10 \log \sum_{i=1}^{n_z} \frac{y_i^2}{n}, \quad (13.2)$$

with SNR_s denoting the SNR for *smaller is better*. Note that the summation in Equation (13.2) is done over the n_z response values in the outer array (e.g., $n_z = 4$ for each of the control factor settings in Table 13.1). If the engineer were interested in maximizing the process mean, y_i^2 in Equation (13.2) could be replaced by $1/y_i^2$.

In many situations, one wishes to determine the levels of the control factors settings, which yield $y = T$ with deviations in either direction being undesirable (i.e., *target is best*). In these settings, Taguchi suggested that the set of control factors could be partitioned into those factors influencing both the process mean and variance, \mathbf{x}_1 , and those control factors affecting only the process mean, \mathbf{x}_2 . Those factors comprising \mathbf{x}_2 are commonly referred to as *tuning* or *adjustment* factors. When tuning factors exist, process optimization can take place with a two-step process. First, levels of \mathbf{x}_1 are chosen to maximize the SNR. Next the levels of \mathbf{x}_2 are chosen to produce $y = T$. Assuming that the tuning factors do indeed yield $y = T$, the squared error loss expression in Equation (13.1) reduces to $Var_z(y)$, and a reasonable SNR would be

$$SNR_T = -10 \log s^2, \quad (13.3)$$

with SNR_T denoting the SNR for *target is best*.

To illustrate Taguchi's approach, observe in Figure 13.1a,b the control factor main effect plots and the SNR plots for the data in Table 13.1. Suppose for this data set, the goal is to achieve a target reduction in shrinkage of 1.5%. Upon computing sample means and sample variances for each of the control factor settings, it is apparent that the variance is a function of the mean. Taguchi's proposed SNR for situations where the variance is a function of the mean is $-10 \log(\bar{y}^2 / s^2)$. In Figure 13.1a, it is apparent that factors A, D, and G are most influential in terms of their effects on the mean, and in Figure 13.1b, factor F appears to have the greatest influence on the SNR. Consequently, one might suggest choosing F at its low level in order to maximize the SNR and then using factors A, D, and G to adjust the mean to a target shrinkage of 1.5%.

While we have provided four basic formulations for the SNR, many other SNR formulations exist, and guidelines for selecting SNRs along with illustrative examples can be found in Phadke (1989) and Phadke and Taguchi (1987).

Certainly, an obvious criticism of Taguchi's two-step approach is that there are many situations in which the SNR is a function of both x_1 and x_2 .

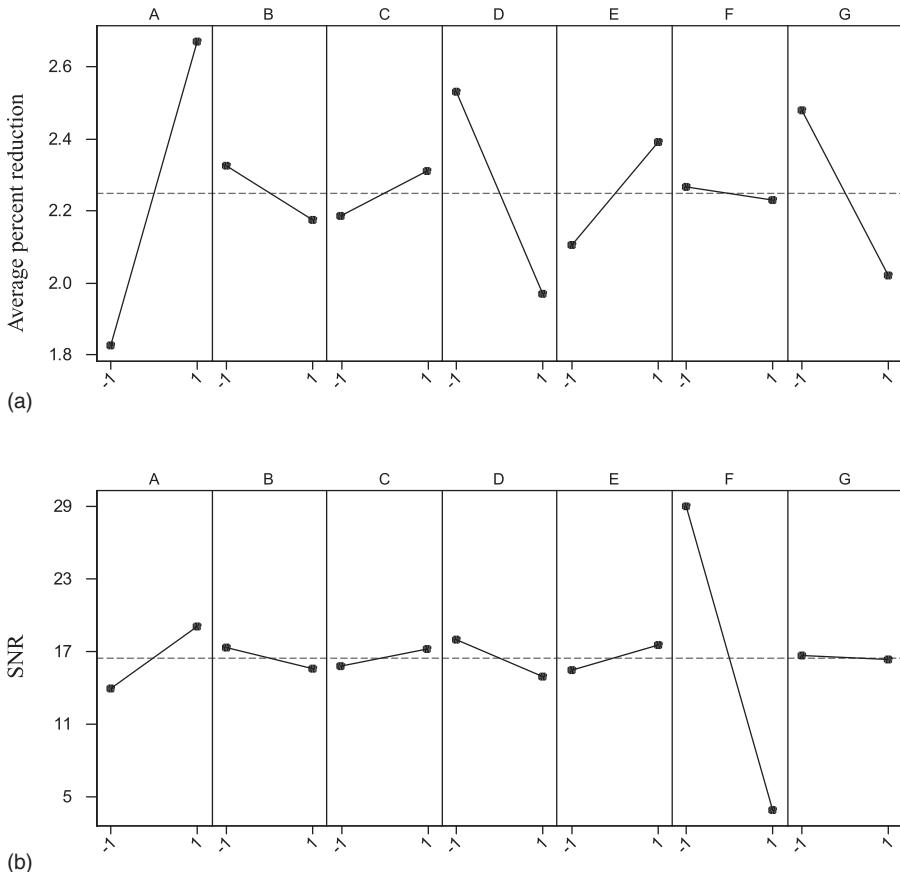


Figure 13.1 (a) Main effect plots for Engel data; (b) SNR plots for Engel data.

Consequently, the SNR does not allow for the uncoupling of those control factors that simultaneously affect the process mean and variance. This uncoupling is vital to understanding the system. Other criticisms of Taguchi's methods include his use of crossed array designs, which are often unnecessarily large, and his analytical methods. Since the late 1980s, many important works have been published which more effectively address the robust design problem in terms of experimental design and analysis. Nair (1992) provided a historical account of robust design up until 1990, and Robinson, Borror, and Myers (2004) summarized the more recent developments in RPD. While Taguchi's contributions were historically very important to the development of the area of RPD, it is generally not recommended to use the Taguchi designs and analyses, as there are more efficient and advantageous approaches. We consider these in the following sections.

13.3 DUAL MODEL RESPONSE SURFACE METHODOLOGY

13.3.1 Overview

A primary message from the panel discussion in Nair (1992) was the need to develop models for the process mean and variance separately instead of reliance on single number summaries, such as the SNR. By understanding the effect of the factors on the two characteristics of interest individually, better choices about the trade-offs can be made and more accurate models constructed since the effects of two sets of changes are not being intermixed. The idea of model building within the context of economical experimentation lends itself naturally to the use of response surface methodology (RSM). During the past 20 years, many RSM approaches have been effectively utilized in RPD applications. In this section, we outline the dual response surface approach in which response surface models are built individually for the process mean and variance. Using the estimated expressions for the mean and variance, numerical optimization methods are easily employed for finding optimal settings of the control factor variables.

13.3.2 Designs for Dual Response Modeling

An important consideration in RPD is the type of experiment utilized in collecting the data, and the crossed array was the design of choice in Taguchi's approach for RPD. Steiner and MacKay (2005) distinguish two types of Taguchi experiments: desensitization and robustness experiments. In desensitization experiments, the levels of the noise factors are changed in a way as to mimic the range of their fluctuations in the actual process. Essentially, the high and low levels of the noise factors are set at the high and low ends of the range of the variables that are observed in production. It is important that the range of the levels set in the experiment accurately portray the actual anticipated fluctuations in regular production. The data in Table 13.1, for example, come from a desensitization experiment. In robustness studies, replicate experimental runs are conducted at each control factor setting, and the variability across replicates is attributed to fluctuations due to noise factors. If the data from Table 13.1 had come from a robustness study, the noise factors, M–O, would not be specifically identified, and one would simply view the four replicates at each control factor setting as replicates. It is important to note that the success of a robustness study hinges on whether or not the important noise factors are acting upon the process over the course of the replicates. More specifically, it is essential that the variability exhibited over the replicates at each control factor setting appropriately mirrors the magnitude and type of variation that would exist in the process if the process permanently operated at each of the given control factor settings.

Asilahijani, Steiner, and MacKay (2010) (henceforth referred to as ASM) outline a four-step process for planning a crossed array RPD experiment. First,

the dominant sources of process variation must be explicitly identified. If the sources of variation cannot be identified, the engineer needs to understand how the process variation changes across time. ASM suggest anecdotal types of methods for identifying primary sources of variation, such as brainstorming and observational studies. Once the dominant sources of variation have been identified, the experimenter must determine how to mimic their effects in an experimental setting. If factors influencing the process variance can be identified, the challenge then becomes to determine the levels of these factors (i.e., the noise factors). Generally speaking, the levels of the noise factors are set at the extremes of their observed levels in the process. If the noise factors cannot be identified, the experimenter must measure the response enough times over time or location to accurately reflect the magnitude of the process variability. ASM suggest that the choice of the control factors should be the third step. Generally, the control factors are selected based upon engineering and/or process knowledge and experience. Montgomery (2009, chapter 1) suggests that activities such as brainstorming and consulting with a variety of experts are useful for selecting design factors and their associated level ranges.

Once the noise and control factors are identified, the RPD experimental design structure must be selected. It is widely accepted that crossed arrays result in a large number of experimental runs and thus can be costly to run. To address run sizes, one might wish to try fractionated crossed arrays provided that the control-by-noise interactions are not aliased with any active effects. The restriction on fractionating within each crossed array may lead to less than ideal structures. Another approach for reducing the economic impact of the crossed array is to incorporate restricted randomization. Box and Jones (1992) point out that crossed arrays run as split-plot experiments should be done in a manner in which the noise factor level combinations comprise the inner array while the control factor combinations make up the outer array. The reason for this is that the inner array effects are estimated with less precision than the outer array effects. Since the control-by-noise interaction effects are exploited in robust design, it is important to estimate these terms with the greatest degree of precision possible. The split-plot set-up suggested by Box and Jones (1992) addresses this issue. If one wishes to completely randomize the run order, the combined array offers an appealing approach and these designs along with their associated methods of analysis are described in Section 13.4. Before discussing the notion of combined arrays, however, we give a general overview of the type of modeling used when crossed-arrays are selected.

13.3.3 Analysis with Dual Response Modeling

When the experimental design is a crossed array, Box (1988) and Vining and Myers (1990) (henceforth referred to as VM) proposed the use of separate regression models for the process mean and variance. In these models, the

sample means at each of the control factor settings serve as data for the means model and the sample variances at these locations serve as the response for the variance model. Given data from a crossed array, there are a number of potential approaches to directly modeling the mean and variance as a function of the control factors. An “off-the-shelf” model for the mean is linear in the control factor terms and can be written as

$$\text{Means model: } \bar{y}_i = \mathbf{x}'_i \boldsymbol{\beta} + g^{1/2}(\mathbf{x}'_i; \boldsymbol{\gamma}) \varepsilon_i, \quad (13.4)$$

where \mathbf{x}'_i and \mathbf{x}''_i are $1 \times k$ and $1 \times l$ vectors of means model and variance model regressors, respectively, expanded to model form, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the associated vectors of means model and variance model parameters, respectively, g is the underlying variance function and the ε_i denote the random errors for the mean function, assumed to be uncorrelated with mean 0 and variance 1. Note that terms for the means model are denoted by \mathbf{x}'_i , whereas terms for the variance model are distinguished by the asterisk (i.e., the regressors comprising \mathbf{x}''_i). This allows for the fact that the process mean and variance may or may not depend upon the same set of regressors.

Just as there are numerous strategies for modeling the process mean, various approaches have been utilized for estimating the underlying variance function. Bartlett and Kendall (1946) demonstrated that if the means model errors are normal and replication exists, the variance structure can be modeled log-linearly as follows

$$\text{Variance model: } \ln(s_i^2) = \mathbf{x}''_i \boldsymbol{\gamma} + \eta_i, \quad (13.5)$$

where η_i denotes the model error term for the natural logged variances. The model parameters in Equations (13.4) and (13.5) can then be estimated via estimated weighted least squares (EWLS). The EWLS estimates of the process mean and variance functions are then given by

$$\text{Estimated process mean: } \hat{\mu}_{y_i} = \mathbf{x}'_i \boldsymbol{\beta} \quad (13.6)$$

$$\text{Estimated process variance: } \hat{\sigma}_{y_i}^2 = \exp(\mathbf{x}''_i \boldsymbol{\gamma}). \quad (13.7)$$

Once the expressions in Equations (13.6) and (13.7) have been obtained, the goal becomes to find operating conditions for the control factors such that the mean is as close as possible to the target while simultaneously keeping process variance at a minimum. Work in the literature has also been done using generalized linear models for estimating the variance function (see Lee and Nelder 1998 and Myers, Brenneman, and Myers 2005). Chipman (1998) proposes a Bayesian analysis approach for product arrays.

To illustrate the dual response approach, consider once again the crossed array data in Table 13.1. Using the information in the main effect plots (see Figure 13.1a), we specify the following models for the mean and variance

$$\text{Means model: } \bar{y}_i = \beta_0 + \beta_1 A + \beta_2 D + \beta_3 G + g^{1/2}(F; \gamma) \varepsilon_i \quad (13.8)$$

$$\text{Variance model: } E(s_i^2) = g(F; \gamma) = \gamma_0 + \gamma_1 F \quad (13.9)$$

where the variance model is fit using gamma regression with a log-link. Fitting the variance model first, we obtain the following process variance estimate:

$$\text{Estimated variance: } \hat{\sigma}_i^2 = \exp(-2.715 + 2.084F). \quad (13.10)$$

From Equation (13.10), the estimated process variance is $\exp(-0.631) = 0.532$, when $F = 1$, and $\exp(-4.799) = 0.008$, when $F = -1$. Consequently, using $\hat{v}_i^{-1} = \exp(-2.715 + 2.084F)$ as weights, half of the observations are weighted with a value of $1/0.532 = 1.88$ and the other half with a value of $1/0.008 = 125$ in an EWLS fit to the means model given in Equation (13.8). The resulting estimate of the process mean is given by

$$\text{Estimated mean: } \hat{\mu}_i = 2.23 + 0.654A - 0.206D + 0.081G. \quad (13.11)$$

Notice that the signs and magnitudes of the slopes in the estimated equations of Equations (13.10) and (13.11) agree with the conclusions obtained from the main effect plots in Figure 13.1a,b. Specifically, we see that variance increases as F increases and the SNR is maximized when F is at its low value. Also, factor A appears to have greater influence on the mean than the other factors with a slight positive trend in the mean across G and a slight negative trend in the mean across D .

An important assumption underlying the modeling so far described is that the user is able to describe the structure in the underlying mean and variance via parametric models. It is often the case that the response varies in a highly nonlinear fashion when the design factors and parametric specifications are inadequate. Vining and Bohn (1998) and Anderson-Cook and Prewitt (2005) propose nonparametric methods for describing the relationship between the response and design factors when parametric models are inadequate. While nonparametric methods can be quite useful, they generally require a larger number of design runs. To address the costly increase in design size when parametric models are inadequate, Pickle et al. (2008) propose a semi-parametric approach for modeling the mean and variance in RPD.

13.4 SINGLE MODEL RESPONSE SURFACE METHODS USING COMBINED ARRAYS

13.4.1 Overview

While crossed arrays are useful in many applications, if there are a large number of control and noise factors, the design, even if highly fractionated, may be too costly to run due to the large number of runs required. While

fractionating the control and noise factor designs can reduce design size, the price paid is that important control-by-control interactions may not be estimable without problematic aliasing issues. To address the concerns surrounding the crossed arrays, Welch et al. (1990) and Myers, Khuri, and Vining (1992) (henceforth referred to as MKV) propose a new model and design structure for RPD. The designs focused on the use of *combined arrays* in which the control and noise factors appear in a single experimental design. MKV suggest the following model for the response at the i^{th} setting of the control factors \mathbf{x}_i (assume there are r_x control factors) and the j^{th} setting of the noise factors \mathbf{z}_j (assume there are r_z noise factors):

$$y(\mathbf{x}_i, \mathbf{z}_j) = \beta_0 + \mathbf{x}'_i \boldsymbol{\beta} + \mathbf{x}'_i \mathbf{B} \mathbf{x}_i + \mathbf{z}'_j \boldsymbol{\gamma} + \mathbf{x}'_i \Delta \mathbf{z}_j + \varepsilon_{ij}, \quad (13.12)$$

where β_0 is the intercept, $\boldsymbol{\beta}$ is an $r_x \times 1$ vector, $\boldsymbol{\gamma}$ is an $r_z \times 1$ vector, and \mathbf{B} an $r_x \times r_x$ matrix containing second-order coefficients in the control variables. The matrix Δ is $r_x \times r_z$ and contains control-by-noise interaction coefficients. The model errors are multivariate normal $(\mathbf{0}, \sigma^2 \mathbf{I})$ and the noise factors are also multivariate normal $(\mathbf{0}, \mathbf{V})$ where $\mathbf{V} = \text{diag}\langle \sigma_l^2 \rangle$ with $l = 1, 2, \dots, r_z$. The model involves first- and second-order terms in the control variables as well as linear terms in the noise variables and the all important control-by-noise interactions. The model for the process mean is found by taking the unconditional expectation of $y(\mathbf{x}_i, \mathbf{z}_j)$ in Equation (13.12) to obtain

$$E_{\varepsilon,z}[y(\mathbf{x}_i, \mathbf{z}_j)] = \beta_0 + \mathbf{x}'_i \boldsymbol{\beta} + \mathbf{x}'_i \mathbf{B} \mathbf{x}_i. \quad (13.13)$$

The process variance is then found by taking the unconditional variance of $y(\mathbf{x}_i, \mathbf{z}_j)$ in Equation (13.12) and is given by

$$\text{Var}_{\varepsilon,z}[y(\mathbf{x}_i, \mathbf{z}_j)] = (\boldsymbol{\gamma}' + \mathbf{x}'_i \Delta) \text{Var}_z(\mathbf{z}_j) (\boldsymbol{\gamma}' + \mathbf{x}'_i \Delta)' + \sigma^2, \quad (13.14)$$

where $\text{Var}_z(\mathbf{z}_j)$ denotes the variance-covariance matrix of \mathbf{z}_j . If we assume that the scaling of the noise factor levels are chosen so that $\text{Var}_z(\mathbf{z}_j) = \sigma_z^2 \mathbf{I}_{r_z}$, then Equation (13.14) can be rewritten as

$$\text{Var}_{\varepsilon,z}[y(\mathbf{x}_i, \mathbf{z}_j)] = \sigma_z^2 (\boldsymbol{\gamma}' + \mathbf{x}'_i \Delta) (\boldsymbol{\gamma}' + \mathbf{x}'_i \Delta)' + \sigma^2 = \sigma_z^2 \mathbf{I}'(\mathbf{x}_i) \mathbf{I}(\mathbf{x}_i) + \sigma^2, \quad (13.15)$$

where $\mathbf{I}(\mathbf{x}_i) = \boldsymbol{\gamma}' + \mathbf{x}'_i \Delta$.

The magnitudes of the variances associated with the noise factors (i.e., the σ_z^2 s) reflect the amount of fluctuation induced in the response by changes in the noise factor levels in the process. The larger the variance associated with the noise factors, the more important it becomes to operate or design the process to minimize its effect. Reducing the variation via RPD is typically accomplished by exploiting the interactions between control and noise factors. As an illustration, consider the interaction plot in Figure 13.2. At the high level

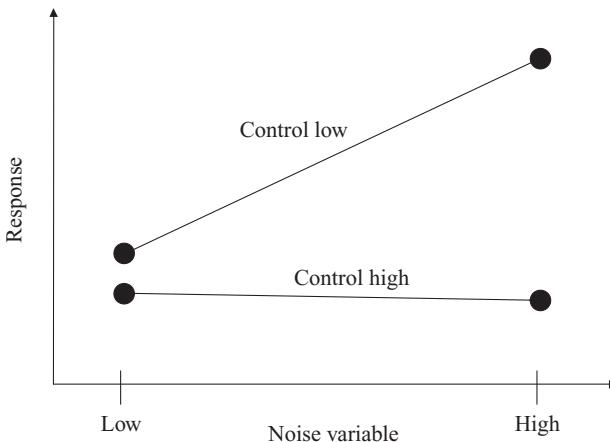


Figure 13.2 Control-by-noise factor interaction plot.

of the control factor, the response is relatively insensitive to the changes in the noise factor level. However, at the low level of the control factor, the variation in the response is large when the noise variable fluctuates between its low and high level. Here, one would be inclined to operate the process at the high level of the control factor to reduce the variation propagated to the response from changes in the noise level factor.

Note from Equation (13.15) that $\mathbf{l}(\mathbf{x}_i) = \boldsymbol{\gamma}' + \mathbf{x}_i'\boldsymbol{\Delta}$ is the vector of partial derivatives of the response model (i.e., $y(\mathbf{x}_i, \mathbf{z}_j)$) with respect to the vector of noise factors, \mathbf{z}_j . The larger the partial derivative, the greater the process variance is as a function of the given noise factor.

13.4.2 Combined Array Designs

Borkowski and Lucas (1997) propose the use of *mixed resolution* designs for use in RPD settings. The notion of design *resolution* is discussed in HK2, chapter 13. The assumed RPD response model contains linear main effects and two-factor interactions in the control factors, linear main effects in the noise factors, and all two-factor control-by-noise interactions. A mixed resolution design suitable for RPD is any 2^{k-p} fractional factorial design that satisfies the following three properties: (1) among the control factors, the design is at least resolution V; (2) among the noise factors, the design is at least resolution III; (3) none of the two-factor control-by-noise interactions are aliased with any main effect or any two-factor control-by-control interactions. When second-order terms are introduced for the control factors, Borkowski and Lucas (1997) propose *composite designs*, similar in structure to the central composite designs of Box and Wilson (1951).

Table 13.2 Mixed Resolution Design with 3 Control Variables and 3 Noise Variables

Run	x_1	x_2	x_3	z_1	z_2	z_3
1	-1	-1	-1	-1	-1	1
2	1	-1	-1	1	-1	-1
3	-1	1	-1	1	-1	-1
4	1	1	-1	-1	-1	1
5	-1	-1	1	1	-1	-1
6	1	-1	1	-1	-1	1
7	-1	1	1	-1	-1	1
8	1	1	1	1	-1	-1
9	-1	-1	-1	-1	1	-1
10	1	-1	-1	1	1	1
11	-1	1	-1	1	1	1
12	1	1	-1	-1	1	-1
13	-1	-1	1	1	1	1
14	1	-1	1	-1	1	-1
15	-1	1	1	-1	1	-1
16	1	1	1	1	1	1
17	-2.38	0	0	0	0	0
18	2.38	0	0	0	0	0
19	0	-2.38	0	0	0	0
20	0	2.38	0	0	0	0
21	0	0	-2.38	0	0	0
22	0	0	2.38	0	0	0
23	0	0	0	0	0	0
24	0	0	0	0	0	0
25	0	0	0	0	0	0

The composite designs have the following three characteristics: (1) The factorial section of the composite design is comprised of the 2^{k-p} fractional factorial points from a mixed factorial design; (2) there are $2r_x$ axial points where r_x denotes the number of control factors (for a given control factor, the axial levels are set at $\pm\alpha$, and all other factors are set at 0); there are no axial runs for the noise factors since the second-order response model in robust design typically assumes no quadratic terms for the noise factors; (3) finally, a total of n_c center runs are present.

A composite design involving three control factors and three noise factors is provided in Table 13.2. The factorial section is a 2^{6-2} fraction with the defining relation $I = x_1x_2x_3z_1 = z_1z_2z_3 = x_1x_2x_3z_2z_3$. With this defining relation, the design is resolution III with regard to the noise-by-noise interactions and resolution IV with respect to the control-by-control interactions. There are 16 runs in the factorial portion of the design, axial points are added with $\alpha = 2.38$, and three center runs are used.

Upon close inspection of the design in Table 13.2, one notices that some two-factor control-by-control interactions are aliased with some of the control-by-noise interactions. Specifically from the first word in the defining relation, $I = x_1x_2x_3z_1$, the aliases are:

$$x_1x_2 = x_3z_1$$

$$x_1x_3 = x_2z_1$$

$$x_2x_3 = x_1z_1.$$

If the user is willing to assume (or prior knowledge indicates) that x_1z_1 , x_2z_1 , and x_3z_1 are negligible, then all control-by-control interactions can be estimated.

If the practitioner is interested in estimating all the two-factor interactions, then the control-by-control interactions and control-by-noise interactions listed above must be *dealiasing*. One very efficient method of dealiasing a fractional factorial design is to augment the design with additional runs (see e.g., HK2, section 13.6.4). A full foldover of the original design is not necessary. Adding a set of runs (often in blocks of two to maintain orthogonality) can dealias the two-factor interactions. The set of runs are duplicates of runs in the original factorial portion with the sign changed on only one of the variables—that variable must be involved in the two-factor interactions that are aliased. For example, in the above alias structure, the signs on any one of the four variables involved, x_1 , x_2 , x_3 , or z_1 , could be switched while the other variables remain unchanged.

From Equation (13.15), it is apparent that the slope vector, \mathbf{l} , plays an important role in the process variance and is the only component of the process variance influenced by the experimental design. As such, when designing an experiment for the robust design setting, it is important to take into consideration the precision for estimating this slope vector. A major goal in robust design is to obtain the control factor setting, \mathbf{x} , which yields the smallest estimated process variance. In practice, the minimum process variance is obtained by setting the estimated slope vector $\hat{\mathbf{l}}(\mathbf{x}) = \mathbf{0}$. A reasonable objective function to consider when deciding upon an optimal experimental design is the precision in which one estimates the location of minimum process variance. Borror, Montgomery, and Myers (2002) propose the use of variance dispersion plots for comparing competing designs in terms of their ability to estimate the variance of the estimated slope. While these plots are sometimes useful, Borror, Montgomery, and Myers (2002) point out that this approach assumes that the user has prior information about effect sizes in the model. Without such prior information, their approach cannot be utilized. The slope optimality criterion proposed by Myers, Myers, and Robinson (2007) does not rely on prior assumptions about effect sizes. Aggarwal and Kaul (1999) construct nonorthogonal combined array designs using optimal designs involving both noise and control

Table 13.3 Filtration Rate Data

Run Number	Factor				Filtration Rate (ga/hour)
	Z1	X1	X2	X3	
1	-1	-1	-1	-1	45
2	1	-1	-1	-1	71
3	-1	1	-1	-1	48
4	1	1	-1	-1	65
5	-1	-1	1	-1	68
6	1	-1	1	-1	60
7	-1	1	1	-1	80
8	1	1	1	-1	65
9	-1	-1	-1	1	43
10	1	-1	-1	1	100
11	-1	1	-1	1	45
12	1	1	-1	1	104
13	-1	-1	1	1	75
14	1	-1	1	1	86
15	-1	1	1	1	70
16	1	1	1	1	96

variables. The design size in the combined array can be significantly reduced if the orthogonality is sacrificed.

13.4.3 Analysis of Combined Array RPD Experiments

Once an appropriate combined array design is chosen, attention turns to the analysis. To illustrate the analysis of a combined array experiment, consider the following example taken from Myers, Montgomery, and Anderson-Cook (2009, section 10.4). A 2^4 factorial design was used to study the filtration rate of a chemical product. The three control factors are pressure (x_1), concentration of formaldehyde (x_2), and stirring rate (x_3), with two levels each, and temperature (z_1), also with two levels, is considered a noise factor. The design matrix and data are provided in Table 13.3. We assume that filtration rate, y , is measured at the batch level. The full single response model proposed by Myers, Khuri, and Vining (1992) is given by

$$y(x_1, x_2, x_3, z_1) = \beta_0 + \sum_{i=1}^3 \beta_i x_i + \sum_{i=1}^3 \sum_{i < j} \beta_{ij} x_i x_j + \gamma_1 z_1 + \sum_{i=1}^3 \delta_{1i} z_1 x_i + \varepsilon. \quad (13.16)$$

The model errors are taken to be independent and identically distributed $N(0, \sigma^2)$.

The expression for the process variation (unconditional variance) of y is found by taking the variance operator through Equation (13.16), yielding

$$\begin{aligned}\text{Var}[y(x_1, x_2, x_3, z_1)] &= \left(\gamma_1 + \sum_{i=1}^3 \delta_{1i} x_i \right)^2 \text{Var}(z_1) + \text{Var}(\epsilon) \\ &= \left(\gamma_1 + \sum_{i=1}^3 \delta_{1i} x_i \right)^2 \sigma_{z_1}^2 + \sigma^2.\end{aligned}\quad (13.17)$$

It is essential to keep in mind that the noise variation manifests itself in the process and not in the experiment. The portion due to noise variation is actually the variation in the average response for a fixed setting of temperature (z_1). This is easily illustrated by writing the conditional expectation of the response

$$E[y|z_1] = \beta_0 + \sum_{i=1}^3 \beta_i x_i + \sum_{i=1}^3 \sum_{i < j} \beta_{ij} x_i x_j + \gamma_1 z_1 + \sum_{i=1}^3 \delta_{1i} z_1 x_i. \quad (13.18)$$

The model in Equation (13.18) represents the average filtration rate for a given setting of temperature. Thus, if replicated observations were taken from the process, the model in Equation (13.18) would reflect the filtration rate that we would expect over the experimental replications. Since the noise variable varies at random in the process by an amount of $\sigma_{z_1}^2$, the conditional mean has process variation associated with it. The amount of variation associated with the conditional mean can be found by taking the variance operator through Equation (13.18), yielding

$$\text{Var}[E(y|z_1)] = \sigma_{z_1}^2 \left(\gamma_1 + \sum_{i=1}^3 \delta_{1i} z_1 x_i \right)^2. \quad (13.19)$$

Practically speaking, for a fixed level of temperature in the process, if we were to observe the response over several batches in production, the average response varies by the amount given by the expression in Equation (13.19). This fluctuation in the conditional mean is due to temperature fluctuating at random in the process.

Table 13.4 presents the SAS analysis resulting from a fit of the regression model in Equation (13.16). Keeping significant terms, the estimated response is given by

$$\hat{y}(\mathbf{x}, z_1) = 70.06 + 10.81z_1 + 4.94x_2 + 7.31x_3 - 9.06x_2 z_1 + 8.31x_3 z_1. \quad (13.20)$$

Note that the regression parameter estimates do not change upon fitting the reduced model since the columns of the model matrix are pairwise orthogonal. The estimated error variance (i.e., mean squared error) does, however, change when using the reduced model and the new mean squared error is 19.51. We assume here that after centering and scaling, z_1 fluctuates at random in the process according to a Normal distribution with mean zero and variance

Table 13.4 SAS Regression Analysis of Filtration Rate Data

Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model		10	5603.12500	560.31250	21.92	0.0016
Error		5	127.81250	25.56250		
Corrected Total		15	5730.93750			
Root MSE		5.05594	R-Square	0.9777		
Dependent Mean		70.06250	Adj R-Sq	0.9331		
Coeff Var		7.21632				
Parameter Estimates						
Variable	DF	Parameter Estimate	Error	t Value	Standard	Pr > t
Intercept	1	70.06250	1.26398	55.43	<.0001	
x1	1	1.56250	1.26398	1.24	0.2713	
x2	1	4.93750	1.26398	3.91	0.0113	
x3	1	7.31250	1.26398	5.79	0.0022	
z1	1	10.81250	1.26398	8.55	0.0004	
x1x2	1	1.18750	1.26398	0.94	0.3906	
x1x3	1	-0.18750	1.26398	-0.15	0.8879	
x2x3	1	-0.56250	1.26398	-0.45	0.6749	
x1z1	1	0.06250	1.26398	0.05	0.9625	
x2z1	1	-9.06250	1.26398	-7.17	0.0008	
x3z1	1	8.31250	1.26398	6.58	0.0012	

$\sigma_{z_1}^2 = 1$, where the high and low levels of z_1 are coded at $\pm\sigma_{z_1}$. If we assume that the fitted model in Equation (13.20) is adequate, then it is reasonable to suggest that the true relationship between y , \mathbf{x} , and z_1 is

$$y(\mathbf{x}, z_1) = \beta_0 + \gamma_1 z_1 + \beta_2 x_2 + \beta_3 x_3 + \delta_{12} z_1 x_2 + \delta_{13} z_1 x_3 + \varepsilon. \quad (13.21)$$

Since $E(z_1) = 0$ and $E(\varepsilon) = 0$, a model for the process mean is

$$E_{e,z_1}[y(\mathbf{x}, z_1)] = \beta_0 + \beta_2 x_2 + \beta_3 x_3. \quad (13.22)$$

Replacing the parameters in (13.22) by their estimates in Equation (13.20), the estimated response surface for the process mean is

$$\widehat{E}_{e,z_1}[y(\mathbf{x}, z_1)] = 70.06 + 4.94x_2 + 7.31x_3. \quad (13.23)$$

The unconditional process variance of the response is given by

$$\text{Var}[y(\mathbf{x}, z_1)] = \sigma_{z_1}^2 (\gamma_1 + \delta_{12}x_2 + \delta_{13}x_3) + \sigma^2. \quad (13.24)$$

Replacing the parameters in Equation (13.24) by their estimates in Equation (13.20), the estimated response surface for the process variance is

$$\begin{aligned} \widehat{\text{Var}}[y(\mathbf{x}, z_1)] &= \sigma_{z_1}^2 (10.81 - 9.06x_2 + 8.31x_3)^2 + \hat{\sigma}^2 \\ &= 136.42 + 82.08x_2^2 + 69.06x_3^2 - 195.88x_2 + 179.66x_3 - 150x_2x_3 \end{aligned} \quad (13.25)$$

To get the final expression for $\widehat{\text{Var}}[y(\mathbf{x}, z_1)]$, we substituted $\sigma_{z_1}^2 = 1$ and $\hat{\sigma}^2 = 19.51$ (the residual mean square from the fitted response model given in Eq. 13.20).

To find optimal operating conditions in the “target is best” scenario, one might overlay the response surfaces for the mean and variance models as provided in Figure 13.3. Another approach would be to utilize a methodology such as nonlinear programming to find the optimal control factor settings such that the estimated squared error loss

$$\widehat{E}[y(\mathbf{x}, z_1) - T]^2 = \{\widehat{E}[y(\mathbf{x}, z_1)] - T\}^2 + \widehat{\text{Var}}[y(\mathbf{x}, z_1)], \quad (13.26)$$

is minimized. As an illustration, suppose the target mean is 75. Substituting the estimated expressions from Equations (13.23) and (13.25) into Equation (13.26), we have

$$\begin{aligned} \widehat{E}[y(\mathbf{x}, z_1) - T]^2 &= \{70.06 + 4.94x_2 + 7.31x_3 - 75\}^2 + \\ &\quad 136.42 + 82.08x_2^2 + 69.06x_3^2 - 195.88x_2 + 179.66x_3 - 150x_2x_3. \end{aligned} \quad (13.27)$$

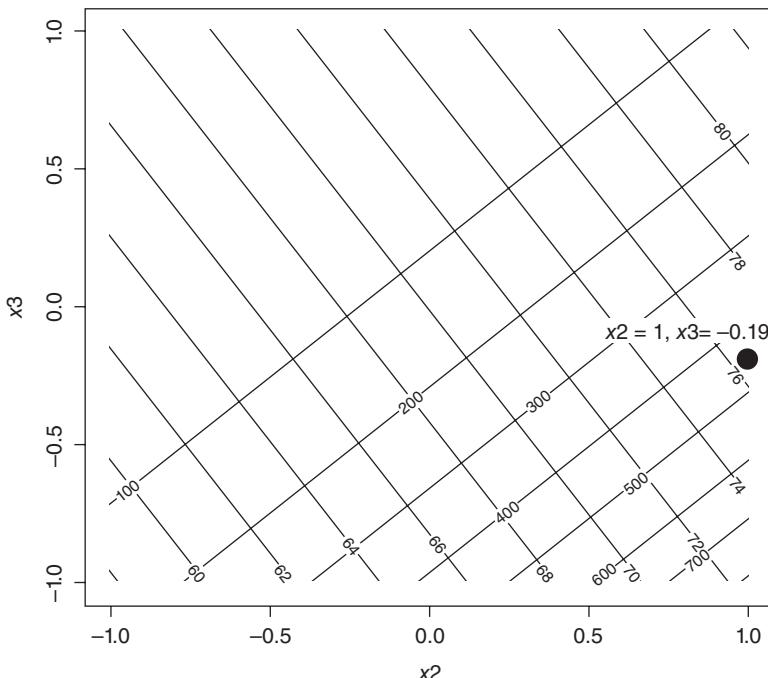


Figure 13.3 Overlaid contours of the estimated process mean (Eq. 13.23) and estimated process variance (Eq. 13.25).

Using the expression above and the reduced gradient method in EXCEL's solver tool, the optimal control factor settings are found to be $x_2 = 1.0$ and $x_3 = -0.19$, and these settings are indicated by the solid circle in Figure 13.3. At these control factor settings, the estimated process mean is 74.13, the estimated process variance is 20.14, and the estimated mean squared error (using the expression in Eq. 13.27) is 20.89.

13.4.4 Analysis of Combined Arrays with Multiple Responses

Thus far, our discussion has focused on RPD problems in which process quality is characterized by a single response. In the case of multiple responses, if one is willing to treat as negligible the correlation between the responses or the uncertainty in the parameter estimates, desirability methods (see Derringer and Suich 1980) can be used to combine the responses into a single desirability index. Miro-Quesada, Del Castillo, and Peterson (2004) propose a Bayesian predictive posterior approach for linear response models. Rajagopal, Del Castillo, and Peterson (2005) extend the work of Miro-Quesada, Del Castillo, and Peterson (2004) to situations in which there is uncertainty in the model form and uncertainty in the distributions of the model errors.

In addition to desirability functions, standard response surface approaches can be applied to balance multiple responses from a RPD. For example, in lower dimensional design spaces, overlaying contour plots of the different responses allows for the selection of an optimal combination of inputs to attain an acceptable combination of expected response values. This is similar to the plot shown in Figure 13.3, but would allow for greater flexibility of the responses to be considered. Alternately, the responses can be prioritized into primary and secondary, and then the primary response can be optimized subject to satisfying at least some threshold performance level for the secondary response. This can be achieved either graphically or using nonlinear programming procedures. See Myers, Montgomery, and Anderson-Cook (2009, section 6.6) for more details on these general approaches. Finally, a suite of possible locations in the input space that involve different weightings for the trade-offs between the response performances can be considered using the Pareto frontier approach (see Lu, Anderson-Cook, and Robinson 2010). This approach finds all combinations of inputs that are not dominated by any other input combination, and then allows the user to select a particular combination that best matches the trade-offs between the responses for the purpose of the study. See Kasprzak and Lewis (2001), Gronwald, Hohm, and Hoffmann (2008), and Trautmann and Mehnen (2009) for more details of how this approach has been used in other disciplines for optimization.

Bayesian methods are particularly useful in the RPD setting since they easily allow the user to specify non-standard performance criteria of the process such as exceedance probabilities, response quantiles (see Robinson, Anderson-Cook, and Hamada 2009), and as shown in the aforementioned papers, an ability to effectively handle multiple responses. Robinson et al. (2010) illustrate Bayesian methods in the RPD setting with nonstandard performance criteria.

13.5 COMPUTER GENERATED COMBINED ARRAYS

Recall from Section 13.4.2, for first-order and first-order plus two-factor interaction RPD models, *mixed resolution designs* (see Borkowski and Lucas 1997) are popular choices when the sample size permits. When the design size does not permit the running of a mixed resolution design or a composite design or when there are restrictions on the design region, a popular approach to design selection for the combined array is via a computer generated experiments. Computer-generated designs rely on the user providing an objective function that reflects (1) the goal of the experiment, (2) the number of experimental runs, (3) the design region, (4) a model which relates the response to the set of design factors, and (5) whether blocking is needed. Using this information, the computer uses a search algorithm and a candidate set of design points (generally a fine grid in the design region) to select an appropriate design. When the design is not fully saturated (i.e., the number of experimental runs,

N , is greater than the number of model parameters, p), it allows for estimation of experimental error variance upon presuming negligible higher order interactions. Computer-generated designs have increased in popularity in a broad number of design applications and this approach to design selection offers flexibility and convenience for the practitioner. Software packages, such as SAS JMP, MINITAB, Design Expert, and the PROC OPTEX procedure in SAS, all offer tools for design selection that enable the user to specify a given model, design size, blocking factors, and an objective function that reflects the primary goal of the experiment.

In many industrial experiments, experimentation is approached sequentially. Specifically, the study often begins with a screening design where the goal is to reduce a large list of candidate factors to a list with only the most important factors. After screening, a larger design is implemented for prediction and optimization purposes. Papers by Bingham and Li (2002), Bingham and Sitter (2003), Wu and Hamada (2009), and Li and Nachtsheim (2000) deal with the screening phase of RPD through attempts at maximizing the capability of a design to estimate models that contain control-by-noise interactions. Hypothesis testing for the evaluation of statistically significant effects is one of the focal points of variable screening. When the standard mixed resolution design calls for more runs than what the user's budget allows for, the D -criterion is an intuitive choice for design selection due to its focus upon precise parameter estimation (for some discussion of design optimality, see HK2, section 1.13). The D -criterion is based upon the notion that a good experimental design is characterized by desirable properties of its moment matrix. Assuming a linear response model,

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

where \mathbf{X} is the $N \times p$ design-model matrix for an experiment with N observations, $\boldsymbol{\beta}$ is the $p \times 1$ vector of coefficients expanded to model form, and $\boldsymbol{\epsilon}$ denotes the $N \times 1$ set of model errors assumed to $\text{NID}(0, \sigma^2)$, the moment matrix is given by

$$\mathbf{M} = \frac{\mathbf{X}'\mathbf{X}}{N}.$$

The D -optimal design is the set of design points that maximizes the determinant of the scaled moment matrix, that is,

$$|\mathbf{M}| = \frac{|\mathbf{X}'\mathbf{X}|}{N^p}. \quad (13.28)$$

Under the assumption of independent normal model errors with constant variance, the variance-covariance matrix of the estimates of the model parameters is $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$, and hence the determinant of $\mathbf{X}'\mathbf{X}$ is inversely

proportional to the square of the hyper volume of the joint confidence region on all of the regression coefficients. Consequently, the D -value for a given design is a global measure of the overall precision in the estimation of the regression model parameters, with large D -values implying greater precision. When designs are compared in terms of their respective D -values, we often speak of D -efficiency, where the D -efficiency of a particular design ξ_0 is defined as

$$D_{\text{eff}} = \left(\frac{|\mathbf{M}(\xi_0)|}{\underset{\xi}{\text{Max}} |\mathbf{M}(\xi)|} \right)^{1/p}, \quad (13.29)$$

and ξ denotes the set of all possible designs.

When using the D -criterion, the user inherently assumes that all model terms are of equal importance for precision in estimation. For standard orthogonal designs, this assumption is often manifested by all terms having equal variances. For nonstandard designs, however, the D -optimal design results in some terms being estimated with less precision than others. As such, it is important to examine the distribution of precision across the model terms. Examining this distribution is especially important in the RPD setting, where one set of terms influences the process mean (i.e., control factor and control-by-control factor interactions) and the other set of terms (i.e., noise factor and control-by-noise factor interactions) influences the process variance.

Villafranca, Zunica, and Zunica (2007) and Myers, Myers, and Robinson (2007) suggest the use of D_s -criteria for choosing optimal screening experiments in the RPD setting. The subscript "s" implies that interest is focused on a *subset* of parameters. Atkinson and Donev (1992) discuss the general use of D_s -optimality. In general, suppose the β vector (p -dimensional) is partitioned as

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \begin{matrix} s \\ p-s \end{matrix},$$

where β_1 contains the s parameters of interest, and β_2 contains the remaining regression coefficients. With D_s -optimality, the determinant criterion is applied to the variance—covariance matrix of the s coefficients of interest. Myers, Myers, and Robinson (2007) define D_s -mean and D_s -variance for situations when the user is interested only in the mean response surface and only in the variance response surface, respectively. The information matrix as used in (13.28) is partitioned as

$$\frac{\mathbf{X}'\mathbf{X}}{N^p} = \mathbf{M} = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix},$$

and the inverse \mathbf{M}^{-1} partitioned as

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{M}^{11} & \mathbf{M}^{12} \\ \mathbf{M}^{21} & \mathbf{M}^{22} \end{bmatrix}.$$

The matrix \mathbf{M}^{11} is the $s \times s$ dispersion matrix, apart from the scaling by N , of the model coefficients of interest. For D_s -optimality, we seek to maximize $|(\mathbf{M}^{11})^{-1}| = |\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21}|$. This is maximized when $\mathbf{M}_{12} = 0$ and $\mathbf{M}_{11} = \mathbf{I}_s$. For an orthogonal design, $\mathbf{M}_{12} = \mathbf{0}$ and $\mathbf{M}_{11} = \mathbf{I}_s$. Thus, the D_s -efficiency for an arbitrary design is calculated as

$$D_s\text{-eff} = \left[\frac{|\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21}|}{1} \right]^{1/s}. \quad (13.30)$$

It is important to note that when the D_s -criterion is associated with the mean parameters (or the variance) parameters, its use implies interest solely in the mean (or variance).

In order to jointly consider parameters influencing the process mean and the process variance, Das et al. (2011) proposed the following weighted optimality criterion

$$D_{RPD(w)} = w(D_s\text{-mean}) + (1-w)(D_s\text{-variance}), w \in [0, 1], \quad (13.31)$$

where D_s -mean is the D_s -criterion value with the subset of parameters affecting the process mean), that is, the intercept, the slope coefficients associated with the control factor main effects and the slope coefficients associated with the control \times control interaction terms), and D_s -variance is the D_s -criterion value with the subset of parameters affecting the process variance (i.e., the slope coefficients associated with the noise factor main effects and the slope coefficients associated with the control \times noise interaction terms). This criterion allows one to consider the combined precision in the estimation of the mean and variance parameters using a weighted average. For instance, one might be interested in estimating the mean parameters with 50% precision and the variance parameters with 50% precision(e.g., $w = 0.5$), or the mean parameters with 75% precision and variance parameters with 25% precision (i.e., $w = 0.75$), and so on. Das et al. show that designs selected using the objective function in Equation (13.31) are often superior in terms of both D_s -mean and D_s -variance efficiency than the overall D -optimal designs. Once a design has been selected based on a specified criterion, then the analysis follows the methods discussed in Section 13.4.3, since the created design is a combined array, and the approach fits with the single response surface model analysis.

13.6 RPD INVOLVING QUANTITATIVE AND QUALITATIVE FACTORS

There are often experimental situations that involve both quantitative and qualitative factors. It is further possible that some of the qualitative factor(s) may also be considered uncontrollable factors (e.g., different suppliers, different operators, and different brands of equipment). Brenneman and Myers (2003) point out that the variation between levels of categorical noise factors can be modeled via a multinomial distribution. Specifically, using a multinomial distribution to describe the variation among the levels of a categorical factor, expressions for the process mean and variance take on the forms

$$E_{\mathbf{z}}[y(\mathbf{x}, \mathbf{z})] = \sum_{i=1}^{r_z+1} p_i y(\mathbf{x}, \text{Category} = i), \quad (13.32)$$

and

$$\begin{aligned} \text{Var}_{\mathbf{z}}[y(\mathbf{x}, \mathbf{z})] &= \sum_{i=1}^{r_z+1} p_i \left(y(\mathbf{x}, \text{Category} = i) - \sum_{i=1}^{r_z+1} p_i y(\mathbf{x}, \text{Category} = i) \right)^2 + \sigma^2 \\ &= \sum_{i=1}^{r_z+1} p_i (y(\mathbf{x}, \text{Category} = i) - E_{\mathbf{z}}[y(\mathbf{x}, \mathbf{z})])^2 + \sigma^2, \end{aligned} \quad (13.33)$$

respectively. For simplicity of presentation, we assume here a single noise variable with $r_z + 1$ categories for which the $P(\text{Category } i) = p_i$ = probability that an experimental unit comes from Category i of the noise factor. In many situations, the probability of observing a given category of the noise variable will be known or well estimated. As an example, if the noise variable represents “supplier of raw material” and there are three suppliers, $p_1 = 0.7$, $p_2 = 0.15$, $p_3 = 0.15$ would imply that 70% of the raw material comes from supplier 1, and suppliers 2 and 3 each provide 15% of the total raw material. In this example, $r_z = 2$ would denote the number of indicator variables used in the response model to describe the three suppliers.

Note that the multinomial modeling of the categorical noise variable implies that the process variance not only depends on levels of the dispersion effects but also upon the proportions associated with each of the categories. The implication of this is that the process variance can be reduced by not only exploiting the control-by-noise factor interactions, but also by adjusting the proportions associated with the individual categories. While Brenneman and Myers (2003) consider fixed proportions (i.e., the p_i), Robinson, Brenneman, and Myers (2006) propose methods for nonfixed proportions, and also consider situations in which the levels of the control variable are not all “equally favorable” in the process. These control factors are known as *nonuniform*

control variables. An example of a nonuniform control variable is “line speed” where higher speeds are more desirable than lower speeds because of higher throughput. Although a slower line speed may result in the best quality in terms of minimum variance, the slower speed may not be desirable from a cost perspective. To achieve a trade-off between the nonuniform control factor and quality, Robinson, Brenneman, and Myers (2006) formulate an objective function involving the geometric mean of a desirability function for the nonuniform control variable and a desirability function for quality.

13.7 CONCLUSIONS

RPD focuses on understanding the influence of both control and noise factors during production through a designed experiment where the noise factors can be controlled. The different approaches outlined in this chapter reflect different paradigms that can be used for experimentation and analysis. In this context, it is important to match the selected design with the analytic approach, since the strengths of the designs are intended to complement the requirements of the analysis.

The key to obtaining a process tuned to give both a desirable mean response and minimal variation around that response is to consider the impact of the design factors on both the mean and variance. The strategies for doing this are varied and quite flexible. In general, combining the two characteristics of the process into a single measure, such as the SNR, is thought to be unappealing since it makes it more difficult to discern the mechanism driving these changes and leads to modeling a more complex function. Both the Dual Response and Single Response approaches presented in Sections 13.3 and 13.4 provide good techniques for understanding the control factor and control-by-noise interaction effects. These two groups of terms in the various models are the drivers of change in the responses of interest and must be estimated well. Consequently, a well-chosen experimental design is essential to finding optimal operating conditions for the process. The influence of different terms in the model(s) leads to different emphases during the design selection phase, and if computer-generated methods are going to be used to create the designs, it is important to carefully select design criteria that match the goals of the analysis.

RPD is an important part of understanding and optimizing a process or production environment, which exploits the additional control that is possible in a nonproduction environment to manipulate the noise factors to more clearly understand their impact on the response. We have talked about control and noise factors as being clear and obvious designations, but in reality, it may be helpful to think about these factors as lying on a continuum where all factors could be controllable, only with potentially differing costs associated with them in production. By understanding how the factors influence both the mean and variance of the response, we can make determinations whether the production response can be adequately optimized with the readily controllable

factors, or whether applying additional controls to a designated “noise” factor (and hence expensive-to-control factor) may be necessary to obtain a satisfactory result for the overall process.

REFERENCES

- Aggarwal, M.L. and R. Kaul (1999). Combined array approach for optimal designs. *Communications in Statistics—Theory and Methods*, **28**, 2655–2670.
- Anderson-Cook, C.M. and K. Prewitt (2005). Some guidelines for using nonparametric methods for modeling data from response surface designs. *Journal of Modern Applied Statistical Methods*, **4**, 106–119.
- Asilahijani, H., S.H. Steiner, and R.J. MacKay (2010). Reducing variation in an existing process with robust parameter design. *Quality Engineering*, **22**, 30–45.
- Atkinson, A. and A.N. Donev (1992). *Optimum Experimental Design*. Oxford: Clarendon Press.
- Bartlett, M.S. and D.G. Kendall (1946). The statistical analysis of variance heterogeneity and the logarithmic transformation. *Journal of the Royal Statistical Society, Series B*, **8**, 128–150.
- Bingham, D. and W. Li (2002). A class of optimal robust parameter designs. *Journal of Quality Technology*, **34**, 244–259.
- Bingham, D. and R.R. Sitter (2003). Fractional factorial split-plot designs for robust parameter experiments. *Technometrics*, **45**, 80–89.
- Borkowski, J.J. and J.M. Lucas (1997). Designs of mixed resolution for process robustness studies. *Technometrics*, **39**, 63–70.
- Borror, C.M., D.C. Montgomery, and R.H. Myers (2002). Evaluation of statistical designs for experiments involving noise variables. *Journal of Quality Technology*, **34**, 54–70.
- Box, G.E.P. (1988). Signal-to-noise ratios, performance criteria, and transformations. *Technometrics*, **30**, 1–17.
- Box, G.E.P. and S. Jones (1992). Split-plot designs for robust product experimentation. *Journal of Applied Statistics*, **19**, 3–26.
- Box, G.E.P. and K.B. Wilson (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society, Series B*, **13**, 1–45.
- Brenneman, W.A. and W.R. Myers (2003). Robust parameter design with categorical noise factors. *Journal of Quality Technology*, **35**, 335–341.
- Chipman, H. (1998). Handling uncertainty in analysis of robust design experiments. *Journal of Quality Technology*, **30**, 11–17.
- Das, D., T.J. Robinson, S.S. Wulff, and W.R. Myers (2011). Optimal screening designs for robust parameter design. *Journal of Statistical Theory and Applications*, **10**, 279–304.
- Derringer, G. and R. Suich (1980). Simultaneous optimization of several response variables. *Journal of Quality Technology*, **12**, 214–219.
- Engel, J. (1992). Modeling variation in industrial experiments. *Applied Statistics*, **41**, 579–593.

- Gronwald, W., T. Hohm, and D. Hoffmann (2008). Evolutionary pareto-optimization of stably folding peptides. *BMC-Bioinformatics*, **9**, 109.
- Kasprzak, E.M. and K.E. Lewis (2001). Pareto analysis in multiobjective optimization using the collinearity theorem and scaling method. *Structural Multidisciplinary Optimization*, **22**, 208–218.
- Lee, Y. and J.A. Nelder (1998). Joint modeling of mean and dispersion. *Technometrics*, **40**, 168–175.
- Li, W. and C.J. Nachtsheim (2000). Model-robust factorial designs. *Technometrics*, **42**, 345–352.
- Lu, L., C.M. Anderson-Cook, and T.J. Robinson (2010). Optimization of designed experiments based on multiple criteria utilizing a Pareto frontier, Los Alamos National Laboratory Technical Report LA-UR 10-04283.
- Miro-Quesada, G., E. Del Castillo, and J.J. Peterson (2004). A Bayesian approach for multiple response surface optimization in the presence of noise variables. *Journal of Applied Statistics*, **31**, 251–270.
- Montgomery, D.C. (2009). *Design and Analysis of Experiments, 7th Edition*. New York: John Wiley & Sons.
- Myers, R.H., W.A. Brenneman, and W.R. Myers (2005). A dual response approach to robust parameter design for a generalized linear model. *Journal of Quality Technology*, **37**, 130–138.
- Myers, R.H., A.I. Khuri, and G.G. Vining (1992). Response surface alternatives to the Taguchi robust parameter design approach. *The American Statistician*, **46**, 131–139.
- Myers, R.H., D.C. Montgomery, and C.M. Anderson-Cook (2009). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments* (3rd ed.). New York: Wiley.
- Myers, W.R., R.H. Myers, and T.J. Robinson (2007). A structural approach to design optimality in robust parameter design. Technical report, Virginia Tech Department of Statistics.
- Nair, V.N. (1992). Taguchi's parameter design: A panel discussion. *Technometrics*, **34**, 127–161.
- Phadke, M.S. (1989). *Quality Engineering Using Robust Design*. Englewood Cliffs, NJ: Prentice-Hall.
- Phadke, M.S. and G. Taguchi (1987). Selection of quality characteristics and S/N ratios for robust design. In *Conference Record, GLOBECOM 87 Meeting, IEEE*.
- Pickle, S., T.J. Robinson, J.B. Birch, and C.M. Anderson-Cook (2008). A semi-parametric approach to robust design. *Journal of Statistical Planning and Inference*, **138**, 114–131.
- Rajagopal, R., E. Del Castillo, and J.J. Peterson (2005). Model and distribution-robust process optimization with noise factors. *Journal of Quality Technology*, **37**, 210–222.
- Robinson, T.J., C.M. Anderson-Cook, and M. Hamada (2009). Bayesian analysis of split-plot experiments with non-normal responses for evaluating non-standard performance criteria. *Technometrics*, **51**, 56–65.
- Robinson, T.J., C.M. Borror, and R.H. Myers (2004). Robust parameter design: A review. *Quality and Reliability Engineering International*, **20**, 81–101.

- Robinson, T.J., W.A. Brenneman, and W.R. Myers (2006). Process optimization via robust parameter design when categorical noise factors are present. *Quality and Reliability Engineering International*, **22**, 307–320.
- Robinson, T.J., A. Pintar, C.M. Anderson-Cook, and M.S. Hamada (2010). A Bayesian approach to the analysis of split-plot product arrays and optimization in robust parameter design. Los Alamos National Laboratory Technical Report LA-UR 09-07537.
- Steiner, S.H. and R.J. MacKay (2005). *Statistical Engineering*. Milwaukee, WI: ASQ Quality Press.
- Taguchi, G. (1986). *Introduction to Quality Engineering*. White Plains, NY: UNIPUB/Kraus International.
- Trautmann, H. and J. Mehnen (2009). Preference-based pareto optimization in certain and noisy environments. *Engineering Optimization*, **41**, 23–38.
- Villafranca, R.R., L. Zunica, and R.R. Zunica (2007). D₅-Optimal experimental plans for robust parameter design. *Journal of Statistical Planning and Inference*, **137**, 1488–1495.
- Vining, G.G. and L.L. Bohn (1998). Response surfaces for the mean and variance using a nonparametric approach. *Journal of Quality Technology*, **30**, 282–291.
- Vining, G.G. and R.H. Myers (1990). Combining Taguchi and response surface philosophies: A dual response approach. *Journal of Quality Technology*, **22**, 38–45.
- Welch, W.J., T.K. Yu, S.M. Kang, and J. Sacks (1990). Computer experiments for quality control by parameter design. *Journal of Quality Technology*, **22**, 15–22.
- Wu, C.F.J. and M.S. Hamada (2009). *Experiments: Planning, Analysis, and Parameter Design Optimization* (2nd ed.). New York: John Wiley.

C H A P T E R 14

Split-Plot Response Surface Designs

G. Geoffrey Vining

14.1 INTRODUCTION

Many industrial experiments involve two classes of factors: hard-to-change and easy-to-change. For example, high-temperature furnaces often require many hours to reach equilibrium after a change in the temperature setting. On the other hand, once the temperature has reached equilibrium, the furnace can process many batches of product. Furnace temperature is a classic example of a hard-to-change factor. Factors on the level of the product batches are easy-to-change.

Agricultural experiments often involve similar types of factors. Generally, the hard-to-change factors are called whole-plot factors. Easy-to-change factors are called subplot factors. Whole plots are the experimental units for the whole-plot factors, and subplots are the experimental units for the subplot factors. The subplots are observational units for the whole-plot factors.

The basic structure for an agricultural split-plot experiment has one random allocation of the whole plots to the whole-plot factors. The design then has separate random allocations of the subplots to the subplot factors within each whole plot. Typically, the same basic subplot experimental design is run within each whole plot. The analysis of agricultural split-plot experiments is well developed (see HK1, chapter 13).

Response surface methodology (RSM) is a popular industrial sequential learning strategy, that utilizes a series of experiments. Generally, the goal of RSM is to find “optimal” operating conditions. The experimenter plans each experiment assuming that a low-order Taylor series is an adequate

Design and Analysis of Experiments: Special Designs and Applications, First Edition. Edited by Klaus Hinkelmann.

© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

approximation to the true underlying mechanism. The initial phases of RSM typically assume that a first-order model is adequate. As a result, these early phases of RSM rely upon two-level, fractional factorial designs (in some cases, very highly fractionated). In the later phases of RSM, as the experimenter nears the presumed optimal set of conditions, a first-order Taylor series model is no longer adequate to model the curvature expected in the experimental region. Popular design classes for fitting second-order models include the central composite design (CCD) (Box and Wilson 1951) and the Box–Behnken design (BBD) (Box and Behnken 1960).

Some people use the term “response surface designs” to refer to any design able to support a second-order model. Most statistical software packages follow this usage. Technically speaking, such usage is not correct since RSM is a sequential learning strategy that uses both first- and second-order models.

Historically, RSM has assumed that all the factors are equally easy (or hard) to change. As a result, the basic RSM textbooks (Myers, Montgomery, and Anderson-Cook 2009, Box and Draper 2007, and Khuri and Cornell 1996) have tended to ignore the split-plot structure that results when some of the factors are significantly harder to change than others. In the 1990s, researchers began to look at the impact of a split-plot structure on response surface designs, broadly understood. Most of the initial work focused on two-level, fractional factorial designs run as split-plots. In the 2000s, researchers brought more attention to second-order response surface designs. Jones and Nachtsheim (2009) provides a nice review of the recent literature on industrial split-plot experiments. This paper tends to be overly enthusiastic about optimal design approaches, but it does offer a reasonable balance of views.

14.2 DIFFERENCES BETWEEN AGRICULTURAL AND INDUSTRIAL EXPERIMENTATION

14.2.1 Basic Differences

Box (1999) notes that the primary differences between RSM and classical agricultural experimentation are immediacy and sequentiality. Generally, classical agricultural experimentation is restricted to one growing season per year. As a result, such experimentation needs to capture as much information as possible. Design sizes are often quite large, involve substantial replication, and require sophisticated blocking techniques. Box notes that most industrial experimenters can plan, execute, and analyze their experiments in a very short time frame, hence, the notion of immediacy. Immediacy then provides the opportunity to build experiments sequentially based upon what was learned from previous experimentation. The combination of immediacy and sequentiality leads to smaller experimental design sizes for specific phases of RSM. Time and cost often dictate either unreplicated or partially replicated designs. Blocking schemes tend to be simpler. In many cases, the blocks represent a

series of phases within RSM. The need for small experimental design sizes leads to very specialized designs for supporting second-order models. In general, three-level fractional factorial designs are not efficient in supporting the second-order Taylor series model.

Another difference between agricultural and industrial experimentation involves the nature of the factors themselves. Typically, agricultural experiments use categorical factors, such as varieties of corn or types of fertilizer. Industrial experimentation often involves quantitative factors, such as temperature and pressure. The consequences are somewhat subtle. Much of agricultural experimentation focuses on the ability to estimate contrasts. Consider a standard analysis of variance (ANOVA) model of the form

$$y_{ij} = \mu_i + \varepsilon_{ij} \quad \text{for } i = 1, 2, \dots, t \quad \text{and } j = 1, 2, \dots, n,$$

where y_{ij} is the j^{th} observed value of the response of interest for the i^{th} treatment, μ_i is the mean of the response for the i^{th} treatment, and ε_{ij} is a random error. For simplicity, this model assumes equal replication of each treatment. A contrast is a special linear combination of the treatment means,

$$\sum_{i=1}^t c_i \mu_i,$$

subject to the constraint

$$\sum_{i=1}^t c_i = 0.$$

Contrasts include as special cases all possible pairwise treatment differences. Such an emphasis on comparing treatment means makes a great deal of sense when the factors are categorical.

Industrial experiments tend to focus on a regression model (a *response surface*) as the basis for the analysis. Historically, RSM relies on Taylor series approximations as the basis for its models. Thus, the two most important models are the first-order Taylor series given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i,$$

and the second-order Taylor series given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \beta_{11} x_{i1}^2 + \dots + \beta_{kk} x_{ik}^2 + \beta_{12} x_{i1} x_{i2} + \dots + \beta_{k-1,k} x_{i,k-1} x_{ik} + \varepsilon_i.$$

The dominant design strategies for first-order models are the 2^k full factorial and the 2^{k-p} fractional factorial systems. By convention, the coded value for the low level of a factor is -1 , and the high level is $+1$. Under certain

circumstances, these designs may include “center runs,” where each factor is set to its 0 level, which is the exact midpoint between the low and high levels. Typically, time and cost prevent the experimenter from replicating the entire 2^k or 2^{k-p} design. In these situations, analysts use normal or half-normal probability plots of the estimated effects (see HK2, section 13.9). Replicating the center run provides an alternative that provides some degrees of freedom for a pure error estimate of the common variance, σ^2 .

The CCD and the BBD are the most common design strategies to support the estimation of second-order models. Both of these approaches place the design points specifically to support the second-order model. The typical CCD requires five levels for each factor. The BBD uses only three levels. Neither design approach may be viewed as a standard fractional factorial. Some software packages recommend “optimal” designs. The optimality criterion specified by the user selects the design points specifically to support the second-order model of the region of interest.

14.2.2 Classical Agricultural Split-Plot Design and Analysis

Many agricultural experiments involve two classes of factors. Some are very difficult to change and are applied either to an entire field or to very large sections of a field. Other factors are easier to change and can be applied to smaller sections of the field quite easily. For example, plows require a significant amount of time and labor to change. Consequently, one tends to use the same plowing method either for the entire field or at least for an entire row of a field. Often, we call such a factor a “whole plot” because we apply it to a large plot of land. On the other hand, the type of fertilizer used is quite easy to change. As a result, one can subdivide the whole plot, the area plowed, into “subplots.”

A consequence of this experimental structure, typically called a split-plot, is that we have more experimental units for the easy-to-change or “subplot factors” than the hard-to-change or “whole-plot factors.” The split-plot structure requires two separate sets of randomizations. One is for the whole-plot experimental units, and the second is for the subplot experimental units within each whole-plot experimental unit. These two sets of randomizations lead to two separate error terms. Further complicating the analysis is that two subplot experimental units within a specific whole-plot experiment unit are correlated. The resulting variance–covariance structure for the observed responses complicates the analysis.

An example helps to illustrate these concepts. Montgomery (2001, pp. 573–574) outlines an industrial experiment that is similar in structure to a typical agricultural experiment. The purpose is to determine the effect of three different pulp preparation methods and four different “cooking” temperatures on the tensile strength of paper. The pulp preparation methods require very large batches, while the cooking process uses much smaller batches. As a result, a split-plot approach makes imminent sense with pulp preparation as the

Table 14.1 The Data for the Paper Tensile Strength Experiment

Prep. Method	Rep 1			Rep 2			Rep 3		
	1	2	3	1	2	3	1	2	3
Temperature									
200	30	34	29	28	31	31	31	35	32
225	35	41	26	32	36	30	37	40	34
250	37	38	33	40	42	32	41	39	39
275	36	42	36	41	40	40	40	44	45

“whole-plot” factor, and the cooking temperature as the subplot factor. The key point is that a single experimental unit for pulp preparation produces several experimental units for the cooking temperature. Consequently, the observational units for the pulp preparation method are the experimental units for the cooking temperature.

Table 14.1 summarizes the experimental results. Note that the basic experiment is run three times. Each setting for the preparation method within a specific replicate is one experimental unit for this factor. As a result, there are nine experimental units for preparation method. However, there is one run for each cooking temperature within this setting of preparation method. Each of these runs for cooking temperature is an experimental unit for this factor. As a result, there are a total of 36 experimental units for cooking temperature.

If we use a completely randomized design (CRD) for the whole plots, our model becomes:

$$y_{ijk} = \mu + \alpha_i + \delta_{ij} + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk}$$

$$i = 1, 2, \dots, a; \quad j = 1, 2, \dots, b; \quad k = 1, 2, \dots, n$$

where:

μ is the overall mean;

α_i is the fixed effect due to the i^{th} “whole-plot treatment”;

δ_{ij} is the “whole-plot error”;

β_k is the fixed effect due to the k^{th} “subplot” treatment;

$(\alpha\beta)_{ik}$ is the whole-plot \times subplot treatment interaction; and

ε_{ijk} is the “subplot error.”

Table 14.2 gives the resulting expected mean squares. Clearly, the term listed as “SP error” is the appropriate error term for testing the subplot effects, and the term “WP error” is the appropriate error term for testing the

Table 14.2 Expected Mean Squares for One Whole-Plot Factor and One Subplot Factor

Whole plot	A	$\sigma^2 + n\sigma_\delta^2 + \frac{bn\sum\alpha_i^2}{a-1}$
Subplot	WP error	$\sigma^2 + n\sigma_\delta^2$
	B	$\sigma^2 + \frac{ab\sum\beta_k^2}{n-1}$
	AB	$\sigma^2 + \frac{b\sum(\alpha\beta)^2}{(a-1)(n-1)}$
	SP error	σ^2

Table 14.3 General ANOVA Table for One Whole-Plot Factor and One Subplot Factor

Source	df	MS	F
Whole-plot treat. (A)	$a - 1$	MS_A	$\frac{MS_A}{MS_{WP}}$
Whole-plot error	$a(b - 1)$	MS_{WP}	
Subplot treat. (B)	$n - 1$	MS_B	$\frac{MS_B}{MS_{SP}}$
Whole-plot treat. by subplot treat. (AB)	$(a - 1)(n - 1)$	MS_{AB}	$\frac{MS_{AB}}{MS_{SP}}$
Subplot error	$a(b - 1)(n - 1)$	MS_{SP}	

Table 14.4 ANOVA Table for the Tensile Strength Data

Source	df	SS	MS	F-Value	Pr > F
PREP	2	128.39	64.19	3.38	0.1038
WHOLE PLOT ERROR	6	113.83	18.97	.	.
TEMP	3	434.08	144.69	36.43	0.0001
PREP*TEMP	6	75.17	12.53	3.15	0.0271
SUBPLOT ERROR	18	71.50	3.97		

whole-plot effects. It is interesting to note that the interactions are all subplot effects. Table 14.3 gives the resulting ANOVA table.

Table 14.4 gives the ANOVA table for the paper tensile strength experiment. As expected, there are fewer degrees of freedom for testing the pulp preparation method (the hard-to-change factor). The net consequence is less power to see the preparation effect, which is at best marginal. Both the temperature and the interaction are significant.

14.2.3 First-Order Industrial Split-Plot Design and Analysis

Often, industrial experimenters use unreplicated 2^k or 2^{k-p} designs. In such cases, analysts typically use normal or half-normal probability plots of the estimated effects to identify the significant effects. Such an approach requires an assumption of effect sparsity.

A split-plot structure complicates the analysis of an unreplicated experiment. In this situation, the whole-plot effects have a different variance than the subplot effects. As a result, the analyst must use two normal or half-normal plots of the estimated effects: one at the whole-plot level and one at the subplot level. Often, there are too few estimated effects at the whole-plot level to make such an analysis effective. The following example illustrates the basic design structure and analysis.

Bisgaard, Fuller, and Barrios (1996, see also Bisgaard 2000) discuss an experiment on plasma. In this case, the response is the “wettability” of paper as the result of the application of a plasma treatment. Plasmas are created in low vacuum chambers. It takes a considerable amount of time to establish the vacuum. As a result, the factors applied to the chamber are hard-to-change. In this case, the whole-plot factors are:

- A: Pressure
- B: Power
- C: Gas flow rate
- D: Type of gas

There is a single subplot factor, paper type (E). Table 14.5 summarizes the experimental results. Because the design is balanced, we can use ordinary least squares to estimate the effects. The subplot effects are E and all interactions involving E. All other effects are at the whole-plot level; Table 14.6 gives the estimated effects. Figures 14.1 and 14.2 give the corresponding normal probability plots. Figure 14.1 indicates that the two most positive effects (the AD interaction and the A main effect) and the most negative effect (the D main effect) are significant. Figure 14.2 indicates that the most positive effect (the main effect of E) and the most negative effect (the AE interaction) are significant. Taken together, these results make sense.

14.2.4 Issues for Second-Order Industrial Split-Plot Designs

Consider an experiment with two whole-plot and two subplot factors. The second-order model is

$$\begin{aligned} f_w(\mathbf{z}_i)' \boldsymbol{\beta}_{wp} = & \beta_0 + \beta_1 w_1 + \beta_2 w_2 + \beta_{12} w_1 w_2 + \beta_{11} w_1^2 + \beta_{22} w_2^2 + \delta_i + \theta_1 x_1 + \theta_2 x_2 \\ & + \theta_3 x_1 x_2 + \theta_4 x_1^2 + \theta_5 x_2^2 + \theta_6 x_1 w_1 + \theta_7 x_1 w_2 + \theta_8 x_2 w_1 + \theta_9 x_2 w_2 + \varepsilon_{ij}, \end{aligned}$$

Table 14.5 The Plasma Experiment

A	B	C	D	E	
				-	+
-	-	-	-	48.6	57.0
+	-	-	-	41.2	38.2
-	+	-	-	55.8	62.9
+	+	-	-	53.5	51.3
-	-	+	-	37.6	43.5
+	-	+	-	47.2	44.8
-	+	+	-	47.2	54.6
+	+	+	-	48.7	44.4
-	-	-	+	5.0	18.1
+	-	-	+	56.8	56.2
-	+	-	+	25.6	33.0
+	+	-	+	41.8	37.8
-	-	+	+	13.3	23.7
+	-	+	+	47.5	43.2
-	+	+	+	11.3	23.9
+	+	+	+	49.5	48.2

Table 14.6 The Estimated Effects for the Plasma Experiment

Term	Whole Plot		Subplot	
	Term	Est. Coef.	Term	Est. Coef.
Constant	40.981	E		1.569
A	5.913	A*E		-2.950
B	2.112	B*E		-0.150
C	-1.694	C*E		-0.069
D	-7.550	D*E		0.512
A*B	-2.106	A*B*E		0.056
A*C	1.488	A*C*E		-0.088
A*D	8.281	A*D*E		-0.406
B*C	-0.425	B*C*E		0.450
B*D	-1.656	B*D*E		-0.094
C*D	0.837	C*D*E		0.162
A*B*C	1.431	A*B*C*E		-0.219
A*B*D	-1.650	A*B*D*E		0.137
A*C*D	-1.156	A*C*D*E		-0.131
B*C*D	0.619	B*C*D*E		0.444
A*B*C*D	3.425	A*B*C*D*E		0.125

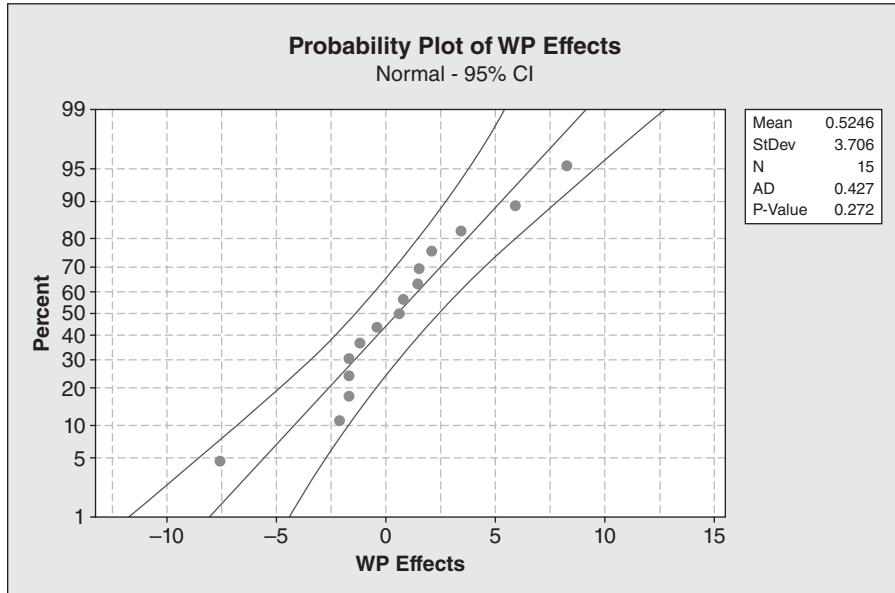


Figure 14.1 Analysis of the whole-plot effects for the plasma experiment.

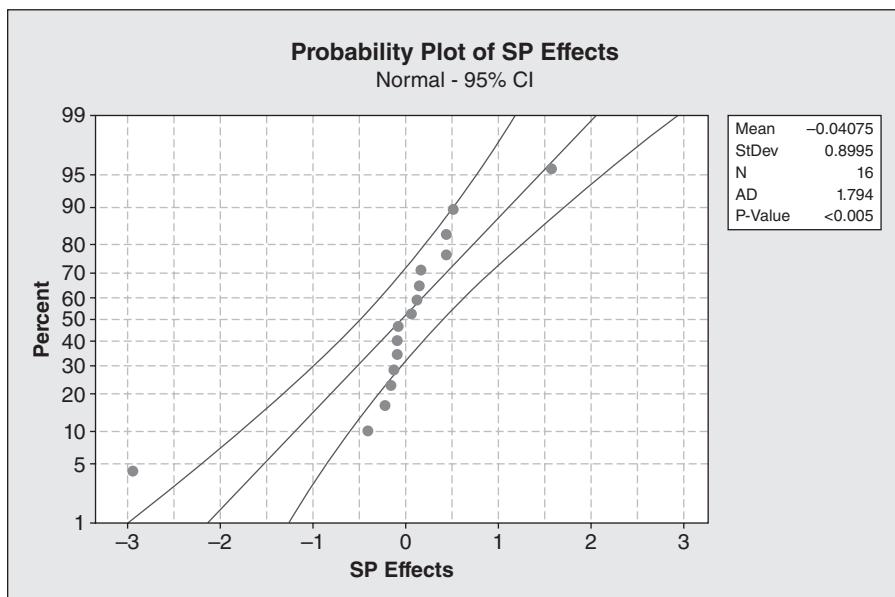


Figure 14.2 Analysis of the subplot effects for the plasma experiment.

where

the β s are the whole-plot terms;

the θ s are the subplot terms;

δ_i is the random error associated with the i^{th} whole plot; and

ε_{ij} is the random error associated with the j^{th} subplot within the i^{th} whole plot.

Typically, the whole plot by subplot interactions are subplot terms. A general model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta} + \boldsymbol{\varepsilon},$$

where

\mathbf{y} is the vector of responses;

\mathbf{X} is the overall model matrix, including both whole-plot and subplot terms;

$\boldsymbol{\beta}$ is the vector of regression coefficients;

$\boldsymbol{\delta}$ is the vector of whole-plot error terms; and

$\boldsymbol{\varepsilon}$ is the vector of subplot error terms.

Let σ^2 be the subplot error variance, and let σ_δ^2 be the whole-plot error variance. Assume the each error term has zero mean and that the two errors are independent. Thus, the variance–covariance matrix for \mathbf{y} is

$$\text{var}[\mathbf{y}] = \boldsymbol{\Sigma} = \sigma^2 \mathbf{I} + \sigma_\delta^2 \mathbf{J}\mathbf{J}',$$

where

$$\mathbf{J} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \dots & \mathbf{0} \\ \vdots & & & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1} \end{bmatrix}.$$

The ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$ is

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

The variance–covariance matrix for the OLS estimate is

$$\text{var}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\Sigma} \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \neq \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}.$$

The generalized least squares (GLS) estimate of β is

$$\hat{\boldsymbol{\beta}}_{\text{gls}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\Sigma}^{-1} \mathbf{y}.$$

The variance–covariance matrix for the GLS estimate is

$$\text{var}[\hat{\beta}_{\text{gls}}] = (\mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1}.$$

In general, the OLS estimate is not the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$. If there is no OLS–GLS equivalence, then the GLS estimate is BLUE if and only if σ^2 and σ_{δ}^2 are known.

Letsinger, Myers, and Lentner (1996) is the first major paper to address second-order response surface designs within a split-plot structure. They examine OLS, iterated reweighted least squares (IRLS), and restricted maximum likelihood (REML) (see HK2, section 1.11). Some people prefer to call REML residual maximum likelihood. The paper's basic conclusion is that OLS is an appropriate estimation technique only if each whole plot contains the exact same design in the subplot factors. Otherwise, one should prefer IRLS or REML. They concede that REML has better asymptotic properties than IRLS. However, they consider small sample situations in their simulations.

The major reason why Letsinger, Myers, and Lentner conclude that OLS is inappropriate for estimating second-order response surface designs goes back to the designs they considered. They start by rearranging a standard CCD design. Table 14.7 illustrates a typical design they consider with one hard-to-change factor (w_1) and two easy-to-change (x_1 and x_2). It is important to note that this design has two whole plots with a single subplot run each, four whole plots with four subplot runs each, and one whole plot with six subplot runs.

Table 14.7 An Example of the Letsinger, Myers, and Lentner Approach to a Split-Plot Central Composite Design

w_1	x_1	x_2
-1.682	0	0
-1	-1	-1
	1	-1
	-1	1
	1	1
0	-1.682	0
	1.682	0
	0	-1.682
	0	1.682
	0	0
	0	0
1	-1	-1
	1	-1
	-1	1
	1	1
1.682	0	0

Such an approach is radically different from the classical agricultural split-plot experiment that has exactly the same number of subplot runs within each whole plot (a property this chapter calls balance). If a factor is truly hard-to-change, then having a whole plot with only a single subplot run does not make much sense. If the economics supports only one subplot run within a whole plot, then the experimenter should utilize a completely randomized design rather than a split-plot. The lack of balance in the Letsinger, Myers, and Lentner work prevented them from seeing the possibility of OLS–GLS equivalent second-order response surface designs.

Other important early papers on split-plot second order response surface designs are Draper and John (1998) and Trinca and Gilmour (2001). Both papers immediately use generalized least squares (GLS) and restricted maximum likelihood (REML) to estimate second-order response surfaces. Neither paper questions whether OLS–GLS equivalent designs exist, probably due to Letsinger, Myers, and Lentner (1996).

14.3 OLS–GLS EQUIVALENT SECOND-ORDER SPLIT-PLOT DESIGNS AND ANALYSIS

14.3.1 Balanced Equivalent Designs

Vining, Kowalski, and Montgomery (2005) show that under certain easily achieved conditions, the OLS estimates are equivalent to the GLS coefficient estimates. In general, a design produces OLS–GLS equivalent estimates of the coefficients if there exists a nonsingular matrix \mathbf{F} such that

$$\boldsymbol{\Sigma} \mathbf{X} = \mathbf{X}\mathbf{F}.$$

Vining, Kowalski, and Montgomery derive such an \mathbf{F} for split-plot experiments. Their proof assumes that the model contains the intercept term and that the design is balanced (each whole plot must have the same number of subplot runs). The proof makes no assumption about the order of the model to be fitted. All of the other conditions are on the designs for the easy-to-change factors within each whole plot. For simplicity, we call second-order designs that meet these criteria “VKM designs.”

One strategy that achieves the necessary conditions is to use exactly the same design in the subplot factors within each whole plot. Letsinger, Myers, and Lentner (1996) called such a design a “crossed bi-randomized” design. Bisgaard (2000) would call such a design a “Cartesian product design.” Often, such experiments are needlessly large. Split-plot experiments using categorical factors, such as classical agricultural split-plots, generally follow this strategy.

Another strategy requires an orthogonal design in the subplot factors within each whole plot. One does not need to use the same orthogonal design within each whole plot, and the design does not need to be full rank. However,

VKM designs must involve the same number of subplot runs. Example designs that work include

- any fraction of a 2^k factorial design;
- whole plots consisting purely of subplot axial runs; and
- center runs.

The center run design is an example of an orthogonal design, which is less than full rank. Bisgaard's (2000) partial confounding designs also fall into this category.

These conditions explain how to modify some CCDs to produce the same estimates using OLS and GLS. For example, consider a modification of a CCD when we have two whole plot and two subplot factors. A VKM approach uses a 2^2 factorial for some of the whole plots, all four subplot axial runs within a single whole plot, and center runs in the other whole plots. Thus, all of the subplot designs within each whole plot are orthogonal. Table 14.8 gives the design for this situation.

Table 14.8 An OLS–GLS Equivalent Modification of a Central Composite Design

Whole Plot	w_1	w_2	x_1	x_2	Whole Plot	w_1	w_2	x_1	x_2
1	-1	-1	-1	-1	7	0	-1	0	0
	-1	-1	1	-1		0	-1	0	0
	-1	-1	-1	1		0	-1	0	0
	-1	-1	1	1		0	-1	0	0
2	1	-1	-1	-1	8	0	1	0	0
	1	-1	1	-1		0	1	0	0
	1	-1	-1	1		0	1	0	0
	1	-1	1	1		0	1	0	0
3	-1	1	-1	-1	9	0	0	-1	0
	-1	1	1	-1		0	0	1	0
	-1	1	-1	1		0	0	0	-1
	-1	1	1	1		0	0	0	1
4	1	1	-1	-1	10	0	0	0	0
	1	1	1	-1		0	0	0	0
	1	1	-1	1		0	0	0	0
	1	1	1	1		0	0	0	0
5	-1	0	0	0	11	0	0	0	0
	-1	0	0	0		0	0	0	0
	-1	0	0	0		0	0	0	0
	-1	0	0	0		0	0	0	0
6	1	0	0	0	12	0	0	0	0
	1	0	0	0		0	0	0	0
	1	0	0	0		0	0	0	0
	1	0	0	0		0	0	0	0

Table 14.9 Box–Behnken Design with 1 Hard-to-Change Factor and 2 Easy-to-Change Factors

Whole Plot	w_1	x_1	x_2	Whole Plot	w_1	x_1	x_2
1	-1	-1	0	4	0	0	0
	-1	1	0		0	0	0
	-1	0	-1		0	0	0
	-1	0	1		0	0	0
2	1	-1	0	5	0	0	0
	1	1	0		0	0	0
	1	0	-1		0	0	0
	1	0	1		0	0	0
3	0	-1	-1	6	0	0	0
	0	1	-1		0	0	0
	0	-1	1		0	0	0
	0	1	1		0	0	0

Box–Behnken designs are three-level designs that are formed by combining two-level factorial designs with balanced incomplete block designs. Table 14.9 gives a VKM Box–Behnken design involving one hard-to-change factor (w_1) and two easy-to-change factors (x_1 and x_2). Table 14.10 summarizes a Box–Behnken design for two hard-to-change and two easy-to-change factors.

There are several important consequences of the OLS–GLS equivalence:

1. One does not need to know the variance components to construct these designs.
2. It is possible to derive exact tests for at least some of the coefficients.
3. As a result of the first two points, it is possible to avoid such methods as REML. However, if one prefers to use REML, the VKM designs lead to better performance in the estimation and testing.
4. The VKM designs provide a basis for “pure error” (model independent) estimates of the variance components, which then can be used to test for model lack-of-fit.
5. The VKM OLS–GLS designs are easy to generate and cover most practical situations, which is contrary to the basic conclusions of Letsinger, Myers, and Lentner (1996).

The first point is nontrivial. Some authors claim that their “optimal” designs are robust to the actual variance components. Typically, these people consider ratios of σ_β^2 to σ^2 from 0.1 to 5 or 10. Ratios less than 1 are rare, and there are real examples where the ratio is over 350! As a result, the ranges in the ratios of the variances components used in checking for robustness are too limited to provide true support for the claim. The second point also is important. REML provides asymptotically efficient estimates and tests. However, response

Table 14.10 Box–Behnken Design with 2 Hard-to-Change Factors and 2 Easy-to-Change Factors

Whole Plot	w_1	w_2	x_1	x_2	Whole Plot	w_1	w_2	x_1	x_2
1	-1	-1	0	0	7	0	-1	-1	0
	-1	-1	0	0		0	-1	1	0
	-1	-1	0	0		0	-1	0	-1
	-1	-1	0	0		0	-1	0	1
2	1	-1	0	0	8	0	1	-1	0
	1	-1	0	0		0	1	1	0
	1	-1	0	0		0	1	0	-1
	1	-1	0	0		0	1	0	1
3	-1	1	0	0	9	0	0	-1	-1
	-1	1	0	0		0	0	1	-1
	-1	1	0	0		0	0	-1	1
	-1	1	0	0		0	0	1	1
4	1	1	0	0	10	0	0	0	0
	1	1	0	0		0	0	0	0
	1	1	0	0		0	0	0	0
	1	1	0	0		0	0	0	0
5	-1	0	-1	0	11	0	0	0	0
	-1	0	1	0		0	0	0	0
	-1	0	0	-1		0	0	0	0
	-1	0	0	1		0	0	0	0
6	1	0	-1	0	12	0	0	0	0
	1	0	1	0		0	0	0	0
	1	0	0	-1		0	0	0	0
	1	0	0	1		0	0	0	0

surface split-plot experiments generally use only a few runs. It is often dangerous to assume that an asymptotic procedure provides good information with small sample sizes.

14.3.2 Non-VKM Balanced OLS–GLS Equivalent Designs

Parker, Kowalski, and Vining (2007) (PKV) demonstrate the construction of other, balanced OLS–GLS equivalent designs. The focus of the VKM paper was on the \mathbf{F} matrix, which dictated their resulting designs. PKV start with the basic VKM result that a design is OLS–GLS equivalent if there exists a non-singular matrix \mathbf{F} such that

$$\mathbf{X}\mathbf{F} = \boldsymbol{\Sigma} \mathbf{X}.$$

Recall,

$$\boldsymbol{\Sigma} = \sigma^2 \mathbf{I} + \sigma_{\delta}^2 \mathbf{J}\mathbf{J}'.$$

As a result,

$$\begin{aligned} \mathbf{X}\mathbf{F} &= [\sigma^2 \mathbf{I} + \sigma_\delta^2 \mathbf{J}\mathbf{J}']\mathbf{X} \\ (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X}\mathbf{F} &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'[\sigma^2 \mathbf{I} + \sigma_\delta^2 \mathbf{J}\mathbf{J}']\mathbf{X} \\ \mathbf{F} &= \sigma^2 \mathbf{I} + \sigma_\delta^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{J}\mathbf{J}'\mathbf{X} \\ \mathbf{X}\mathbf{F} &= \sigma^2 \mathbf{X} + \sigma_\delta^2 \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{J}\mathbf{J}'\mathbf{X}. \end{aligned}$$

PKV define \mathbf{K} by

$$\mathbf{K} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{J}\mathbf{J}'\mathbf{X}.$$

Any design that meets the condition

$$\mathbf{X}\mathbf{K} = \mathbf{J}\mathbf{X},$$

is OLS–GLS equivalent. PKV use this insight to construct minimum whole-plot CCD-like designs and D -optimal designs. PKV denote the value for the subplot axial runs by α , and the value for the whole plot axial runs as β . Table 14.11 summarizes a minimum whole-plot modification of a CCD for two whole plot and two subplot factors.

PKV conclude that:

- The VKM designs are most flexible for two basic reasons. First, these designs can use any axial value for the subplot factors (α) and for the whole-plot factors (β). Second, OLS–GLS equivalence holds with model reduction. In addition, VKM designs also have the most exact tests.
- The minimum whole-plot designs establish a lower bound on the number of whole plots required. The experimenter always can add other whole plots that are consistent with OLS–GLS equivalence.
- For minimum whole-plot designs, when $\beta \neq 1$, the equivalence condition depends on α and β .
- Unlike the VKM design, minimum whole-plot designs with whole plots consisting of all subplot center runs violate the OLS–GLS equivalence.
- OLS–GLS equivalence does not hold, in general, with model reduction for minimum whole-plot designs.
- Parker, Kowalski, and Vining (2006) provides a catalog of OLS–GLS equivalent designs.

14.3.3 Unbalanced OLS–GLS Equivalent Designs

Parker, Kowalski, and Vining (2007) (PKV-2) demonstrate unbalanced OLS–GLS equivalent designs. This paper is a response to concerns about the number

Table 14.11 Example of a Minimum Whole-Plot Second-Order Design

Whole-Plot	z_1	z_2	x_1	x_2	Whole-Plot	z_1	z_2	x_1	x_2
1	-1	-1	-1	-1	6	β	0	0	0
	-1	-1	-1	1		β	0	0	0
	-1	-1	1	-1		β	0	0	0
	-1	-1	1	1		β	0	0	0
	-1	-1	0	0		β	0	0	0
2	1	-1	-1	-1	7	0	$-\beta$	0	0
	1	-1	-1	1		0	$-\beta$	0	0
	1	-1	1	-1		0	$-\beta$	0	0
	1	-1	1	1		0	$-\beta$	0	0
	1	-1	0	0		0	$-\beta$	0	0
3	-1	1	-1	-1	8	0	β	0	0
	-1	1	-1	1		0	β	0	0
	-1	1	1	-1		0	β	0	0
	-1	1	1	1		0	β	0	0
	-1	1	0	0		0	β	0	0
4	1	1	-1	-1	9	0	0	$-\alpha$	0
	1	1	-1	1		0	0	α	0
	1	1	1	-1		0	0	0	$-\alpha$
	1	1	1	1		0	0	0	α
	1	1	0	0		0	0	0	0
5	$-\beta$	0	0	0					
	$-\beta$	0	0	0					
	$-\beta$	0	0	0					
	$-\beta$	0	0	0					
	$-\beta$	0	0	0					

of whole-plot center runs for a VKM design. The requirement of balance seems to create an overabundance of subplot center runs, especially when evaluated via D -optimality. To a certain extent such a criticism is somewhat unfair since these VKM designs focused on pure error estimates of the variance components. In general, pure error estimates of the variance components require replication of points that provide little information for estimating the presumed model. As a result, they do not perform well in terms of such criteria as D -optimality.

PKV-2 develop a catalog of unbalanced designs, many of which use whole-plot center runs with a fewer number of subplot runs. The construction technique starts with a balanced design and uses the same \mathbf{K} matrix from PKV to find appropriate unbalanced designs. The goal is to start with a previously found balanced design and reduce subplot center runs such that

$$\mathbf{X}\mathbf{K} = \mathbf{J}\mathbf{X}.$$

This paper demonstrates example designs for the following cases:

- VKM CCDs;
- VKM Box–Behnken;
- minimum whole-plot designs; and
- Notz saturated designs.

One drawback to these designs is that exact tests do not exist for the coefficients.

14.4 EXACT TESTS FOR THE COEFFICIENTS

For a second-order response surface design run as a split-plot there are three general classes of tests:

- purely subplot effects;
- effectively whole-plot effects; and
- effects somewhere in between.

The design determines into which class a specific term falls. Exact tests require balanced OLS–GLS equivalent designs. For such designs, exact tests exist for the pure subplot effects and for the essentially whole-plot effects. One needs to use synthetic tests for the other effects. For example, the design given in Table 14.12 makes the pure subplot quadratic terms effectively whole-plot effects. As a result, all terms have exact tests. The design given in Table 14.13 does not allow exact tests for the subplot pure quadratic terms. All other terms do have exact tests.

The key to exact tests is to develop the appropriate error term. First, consider a residual-based error term for the pure subplot effects. Consider the model

$$\mathbf{y} = \mathbf{J}\boldsymbol{\mu}_{wp} + \mathbf{S}\boldsymbol{\gamma} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}.$$

$\boldsymbol{\mu}'_{wp} = [\mu_1, \mu_2, \dots, \mu_m]$ is the $m \times 1$ vector of means for the whole plots;

\mathbf{S} is the model matrix for the subplot terms; and

$\boldsymbol{\gamma}$ is the vector of subplot coefficients.

Let

$$\mathbf{X}^* = [\mathbf{J} \quad \mathbf{S}].$$

Define $SS_{\text{res},S}$ by

$$SS_{\text{res},S} = \mathbf{y}'[\mathbf{I} - \mathbf{X}^*(\mathbf{X}^*\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}]\mathbf{y}, \quad (14.1)$$

Table 14.12 A VKM Central Composite Design with Subplot Quadratics as Whole-Plot Effects

WP	w_1	w_2	x_1	x_2	x_3	WP	w_1	w_2	x_1	x_2	x_3
1	-1	-1	-1	-1	-1	8	0	1	0	0	0
	-1	-1	1	-1	1		0	1	0	0	0
	-1	-1	-1	1	1		0	1	0	0	0
	-1	-1	1	1	-1		0	1	0	0	0
2	1	-1	1	1	1	9	0	0	-1	0	0
	1	-1	-1	1	-1		0	0	1	0	0
	1	-1	1	-1	-1		0	0	-1	0	0
	1	-1	-1	-1	1		0	0	1	0	0
3	-1	1	1	1	1	10	0	0	0	-1	0
	-1	1	-1	1	-1		0	0	0	1	0
	-1	1	1	-1	-1		0	0	0	-1	0
	-1	1	1	-1	1		0	0	0	1	0
4	1	1	-1	-1	-1	11	0	0	0	0	-1
	1	1	1	-1	1		0	0	0	0	1
	1	1	-1	1	1		0	0	0	0	-1
	1	1	1	1	-1		0	0	0	0	1
5	-1	0	0	0	0	12	0	0	0	0	0
	-1	0	0	0	0		0	0	0	0	0
	-1	0	0	0	0		0	0	0	0	0
	-1	0	0	0	0		0	0	0	0	0
6	1	0	0	0	0	13	0	0	0	0	0
	1	0	0	0	0		0	0	0	0	0
	1	0	0	0	0		0	0	0	0	0
	1	0	0	0	0		0	0	0	0	0
7	0	-1	0	0	0	14	0	0	0	0	0
	0	-1	0	0	0		0	0	0	0	0
	0	-1	0	0	0		0	0	0	0	0
	0	-1	0	0	0		0	0	0	0	0

where $(\mathbf{X}^* \mathbf{X}^*)^-$ is a generalized inverse of $\mathbf{X}^* \mathbf{X}^*$. Vining and Kowalski (2008) establish that

$$\frac{SS_{\text{res},S}}{\sigma^2}$$

follows a χ^2 distribution with

$$df_{\text{res},S} = mn - \text{rank}[\mathbf{X}^* (\mathbf{X}^{*\prime} \mathbf{X}^*)^- \mathbf{X}^{*\prime}],$$

degrees of freedom. Thus, an appropriate error term for testing the “purely” subplot effects is

Table 14.13 Example VKM Central Composite Design with No Exact Test for the Subplot Quadratic Effects

Whole Plot	w_1	w_2	x_1	x_2	y	Whole Plot	w_1	w_2	s_1	s_2	y
1	-1	-1	-1	-1	80.40	7	0	-1	0	0	80.07
	-1	-1	1	-1	71.88		0	-1	0	0	80.79
	-1	-1	-1	1	89.91		0	-1	0	0	80.20
	-1	-1	1	1	76.87		0	-1	0	0	79.95
2	1	-1	-1	-1	87.48	8	0	1	0	0	68.98
	1	-1	1	-1	84.49		0	1	0	0	68.64
	1	-1	-1	1	90.84		0	1	0	0	69.24
	1	-1	1	1	83.61		0	1	0	0	69.20
3	-1	1	-1	-1	62.99	9	0	0	-1	0	78.56
	-1	1	1	-1	49.95		0	0	1	0	68.63
	-1	1	-1	1	79.91		0	0	0	-1	74.59
	-1	1	1	1	63.23		0	0	0	1	82.52
4	1	1	-1	-1	73.06	10	0	0	0	0	74.86
	1	1	1	-1	66.13		0	0	0	0	74.22
	1	1	-1	1	84.45		0	0	0	0	74.06
	1	1	1	1	73.29		0	0	0	0	74.82
5	-1	0	0	0	71.87	11	0	0	0	0	73.60
	-1	0	0	0	71.53		0	0	0	0	73.59
	-1	0	0	0	72.08		0	0	0	0	73.34
	-1	0	0	0	71.58		0	0	0	0	73.76
6	1	0	0	0	82.34	12	0	0	0	0	75.52
	1	0	0	0	82.20		0	0	0	0	74.74
	1	0	0	0	81.85		0	0	0	0	75.00
	1	0	0	0	81.85		0	0	0	0	74.90

$$MS_{res,S} = \frac{SS_{res,S}}{df_{res,S}}.$$

“Purely” subplot effects are those completely orthogonal to all of the whole-plot terms. Vining and Kowalski show that when the design achieves the OLS–GLS equivalence, the resulting tests are exact.

No exact tests exist for nominal subplot effects if they are not completely orthogonal to the whole-plot effects unless they are essentially whole-plot effects. However, Vining and Kowalski generate approximate tests based on Satterthwaite’s procedure (see HK1, section 9.7.7). Let X_1 be X with the column associated with the term of interest removed. The appropriate linear combination of the variance components is given by

$$\text{trace}[(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\Sigma].$$

This linear combination then becomes the error term for testing. Satterthwaite’s procedure allows the basis for determining the approximate degrees of freedom.

Vining and Kowalski (2008) also develop a residual-based error term for the whole-plot effects by considering the model

$$\mathbf{y} = \mathbf{X}_w^* \boldsymbol{\beta}_w + \mathbf{X}_s \boldsymbol{\gamma}_s + \boldsymbol{\delta} + \boldsymbol{\epsilon},$$

where

\mathbf{X}_w^* is the model matrix for all terms that are essentially whole-plot effects for at least one whole plot;

$\boldsymbol{\beta}_w$ is the vector of coefficients associated with these terms;

\mathbf{X}_s is the model matrix for all the terms that are strictly subplot effects; and

$\boldsymbol{\gamma}_s$ is the vector of coefficients for these terms.

A nominal subplot effect is a part of \mathbf{X}_w^* if there is at least one whole plot where the term does not change value. Let

- \mathbf{S}_w^* be the model matrix of nominal subplot effects that are part of \mathbf{X}_w^* ; and
- \mathbf{W} be the model matrix for the true whole-plot effects.
- Thus, $\mathbf{X}_w^* = [\mathbf{W} \quad \mathbf{S}_w^*]$.

Let \mathbf{S}_w be the matrix whose nonzero elements are the submatrices of \mathbf{S}_w^* that are essentially whole-plot effects. Finally, let $\mathbf{X}_w = [\mathbf{W} \quad \mathbf{S}_w]$. Vining and Kowalski (2008) define $SS_{\text{res},W}$ by

$$SS_{\text{res},W} = \mathbf{y}'[\mathbf{J}(\mathbf{J}'\mathbf{J})^{-1}\mathbf{J}' - \mathbf{X}_w(\mathbf{X}_w'\mathbf{X}_w)^{-1}\mathbf{X}_w']\mathbf{y}. \quad (14.2)$$

They establish that

$$\frac{SS_{\text{res},W}}{\sigma^2 + n\sigma_\delta^2},$$

follows a χ^2 distribution with

$$df_{\text{res},W} = m - \text{rank}[\mathbf{X}_w(\mathbf{X}_w'\mathbf{X}_w)^{-1}\mathbf{X}_w'],$$

if the design meets the conditions for OLS–GLS equivalence. They further establish that

$$MS_{\text{res},W} = \frac{SS_{\text{res},W}}{df_{\text{res},W}},$$

is an appropriate error term for all whole-plot effects.

Table 14.14 Data for Ceramic Pipe Example

Whole Plot	w_1	w_2	x_1	x_2	y	Whole Plot	w_1	w_2	x_1	x_2	y
1	-1	-1	-1	-1	80.40	7	0	-1	0	0	80.07
	-1	-1	1	-1	71.88		0	-1	0	0	80.79
	-1	-1	-1	1	89.91		0	-1	0	0	80.20
	-1	-1	1	1	76.87		0	-1	0	0	79.95
2	1	-1	-1	-1	87.48	8	0	1	0	0	68.98
	1	-1	1	-1	84.49		0	1	0	0	68.64
	1	-1	-1	1	90.84		0	1	0	0	69.24
	1	-1	1	1	83.61		0	1	0	0	69.20
3	-1	1	-1	-1	62.99	9	0	0	-1	0	78.56
	-1	1	1	-1	49.95		0	0	1	0	68.63
	-1	1	-1	1	79.91		0	0	0	-1	74.59
	-1	1	1	1	63.23		0	0	0	1	82.52
4	1	1	-1	-1	73.06	10	0	0	0	0	74.86
	1	1	1	-1	66.13		0	0	0	0	74.22
	1	1	-1	1	84.45		0	0	0	0	74.06
	1	1	1	1	73.29		0	0	0	0	74.82
5	-1	0	0	0	71.87	11	0	0	0	0	73.60
	-1	0	0	0	71.53		0	0	0	0	73.59
	-1	0	0	0	72.08		0	0	0	0	73.34
	-1	0	0	0	71.58		0	0	0	0	73.76
6	1	0	0	0	82.34	12	0	0	0	0	75.52
	1	0	0	0	82.20		0	0	0	0	74.74
	1	0	0	0	81.85		0	0	0	0	75.00
	1	0	0	0	81.85		0	0	0	0	74.90

As an example, an engineer conducted a second-order split-plot experiment to study the effects of:

- temperature, zone 1 of furnace (w_1 , hard-to-change);
- temperature, zone 2 of furnace (w_2 hard-to-change);
- amount of binder in the formulation (x_1 , easy-to-change); and
- grinding speed of the batch (x_2 , easy-to-change)

on the strength of ceramic pipe. Table 14.14 summarizes the data. From Equation (14.1), we obtain that $SS_{\text{res,S}}$ is 2.117755 and has 28 degrees of freedom. Thus, $MS_{\text{res,S}}$ is 0.07563. From Equation (14.2), we obtain that $SS_{\text{res,W}}$ is 27.02772 and has 6 degrees of freedom. Thus, $MS_{\text{res,W}}$ is 4.65462. All of the model terms have exact tests except the two subplot pure quadratic terms. The appropriate linear combination of the variance components for these two tests is

$$0.704878MS_{\text{res,S}} + 0.195122MS_{\text{res,W}}.$$

Table 14.15 Estimated Coefficients Using OLS and Summary of Test

Term	Estimated Coefficient	Standard Error	t	p
Intercept	74.9055	0.4968	150.77	0.0000
W_1	4.5579	0.4404	10.35	0.0000
W_2	-6.5592	0.4404	-14.89	0.0000
W_1^2	1.7381	0.8077	2.15	0.0314
W_2^2	-0.5407	0.8077	-0.67	0.5032
W_1Z_2	0.8431	0.5394	1.56	0.1180
S_1	-4.973	0.0648	-76.72	0.0000
S_2	4.0922	0.0648	63.13	0.0000
S_1^2	-2.3864	0.5486	-4.35	0.0000
S_2^2	2.5736	0.5486	4.69	0.0000
S_1S_2	-1.0394	0.0688	-15.11	0.0000
W_1S_1	1.4356	0.0688	20.88	0.0000
W_1S_2	-1.4794	0.0688	-21.52	0.0000
W_2S_1	-1.0019	0.0688	-14.57	0.0000
W_2S_2	1.9856	0.0688	28.81	0.0000

Satterthwaite's procedure yields 6.82 approximate degrees of freedom. Table 14.15 summarizes the results.

14.5 PROPER RESIDUALS FOR CHECKING ASSUMPTIONS

Standard statistical analyses typically assume the following about the random errors:

- They follow a normal distribution.
- They have constant variance.
- They are independent.

Usually, we check these assumptions via residual plots. However, in split-plot experiments, there are two error terms. Therefore, two types of residuals need to be calculated: whole-plot residuals and subplot residuals. Vining and Kowalski (2008) use the error terms for the exact tests to develop appropriate residual plots for the split-plot situation.

Recall that the subplot sums of squares residual is

$$\mathbf{y}'[\mathbf{I} - \mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}]\mathbf{y}.$$

We note that $[\mathbf{I} - \mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}]$ is an idempotent matrix. Therefore, we can define the vector of subplot residuals as

$$\mathbf{e}_s = [\mathbf{I} - \mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}]\mathbf{y}.$$

Next, recall that \mathbf{M} is a matrix that calculates the means for each whole plot, and \mathbf{Z} is the model matrix for the subplot terms. These residuals can be thought of as the “the individual data values – predicted values from the subplot model adjusted by the appropriate whole-plot mean.” Recall also that the whole-plot sums of squares residual is

$$\mathbf{y}'[\mathbf{J}(\mathbf{J}'\mathbf{J})^{-1}\mathbf{J}' - \mathbf{X}_w(\mathbf{X}_w'\mathbf{X}_w)^{-1}\mathbf{X}_w']\mathbf{y}.$$

Again, the matrix $[\mathbf{J}(\mathbf{J}'\mathbf{J})^{-1}\mathbf{J}' - \mathbf{X}_w(\mathbf{X}_w'\mathbf{X}_w)^{-1}\mathbf{X}_w']$ is idempotent. Therefore, we can define the vector of whole-plot residuals as

$$\mathbf{e}_w = [\mathbf{J}(\mathbf{J}'\mathbf{J})^{-1}\mathbf{J}' - \mathbf{X}_w(\mathbf{X}_w'\mathbf{X}_w)^{-1}\mathbf{X}_w']\mathbf{y}.$$

These residuals can be thought of as “the whole-plot means – predicted values from the whole-plot model.” We note:

- These are appropriate residuals for checking assumptions.
- We need separate plots for the subplot and the whole-plot residuals.
- We plot the subplot residuals against the predicted values.
- We plot the whole-plot residuals against the average of the predicted values for each whole plot.

Figures 14.3–14.6 illustrate these residual plots for the ceramic pipe experiment summarized in Table 14.14. These plots indicate no problems with the assumptions underlying the analysis summarized in Table 14.15.

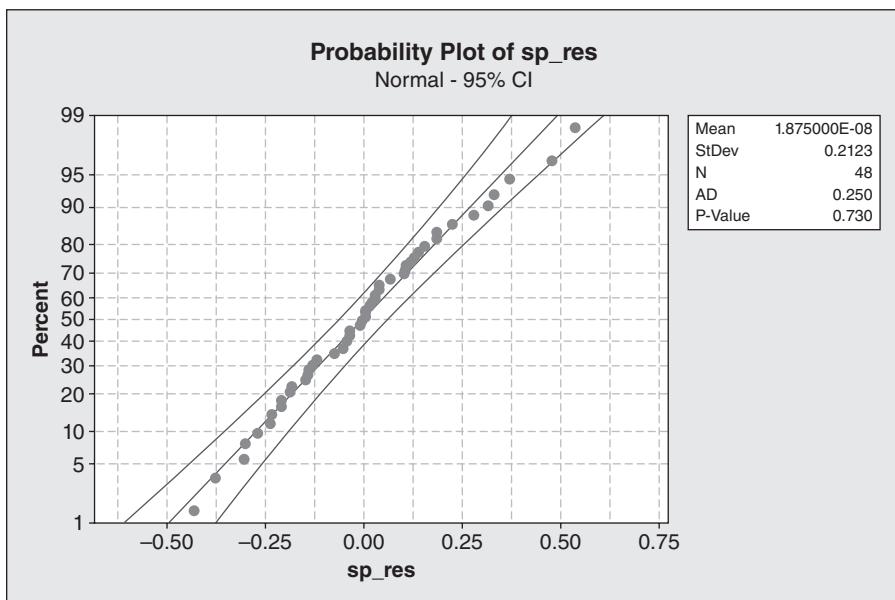


Figure 14.3 Normal probability plot of the subplot residuals for the ceramic pipe experiment.

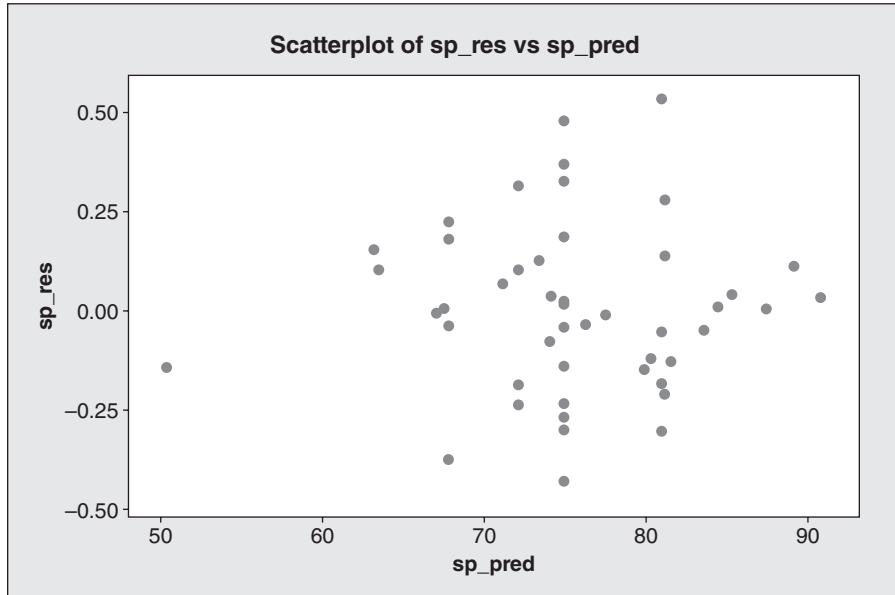


Figure 14.4 Plot of the subplot residuals versus the predicted values for the ceramic pipe experiment.

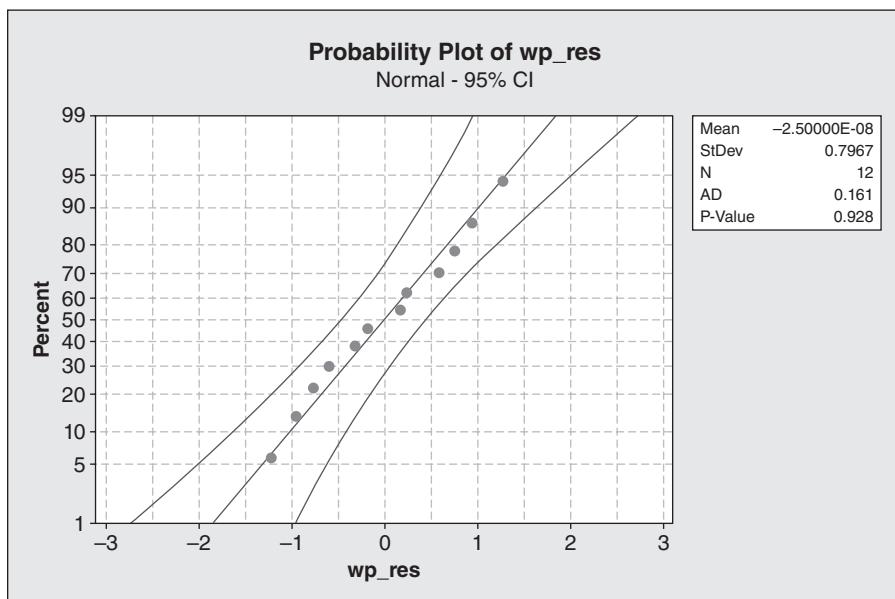


Figure 14.5 Normal probability plot of the whole-plot residuals for the ceramic pipe experiment.

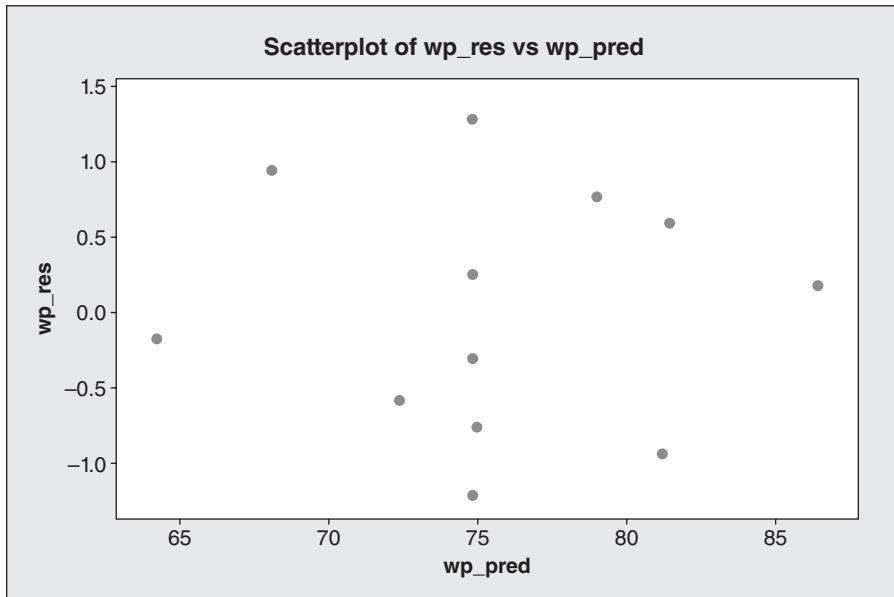


Figure 14.6 Plot of the whole-plot residuals versus the whole-plot means for the ceramic pipe experiment.

14.6 “OPTIMAL” SECOND-ORDER SPLIT-PLOT DESIGNS

D -optimality has been the dominant criterion used to generate split-plot response surface designs. Recall that

$$\text{var}[\boldsymbol{\beta}_{\text{gls}}] = (\mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1}.$$

In its simplest form, the D -optimal design is that design that maximizes the determinant of $\mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X}$, which in turn minimizes the determinant of $(\mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1}$. The interpretation of the D -optimal design is that it produces the smallest confidence ellipsoid around the estimated coefficients. Thus, in a meaningful sense, the D -optimal design produces the best joint estimates of the coefficients. Significant papers on this approach include Goos and Vanderbroek (2001, 2003), Goos (2002, 2006), Goos, Langhans, and Vanderbroek (2006), and Jones and Goos (2007). All of these papers, either implicitly or explicitly, assume REML as the basis for estimation and inference.

The designs recommended as being optimal depend heavily upon

- the assumed form of the model; and
- the shape of the experimental region of interest.

In addition, these designs are optimal only in terms of a very specific and narrow criterion, in this case the specific determinant. Generally, optimal designs are most useful for nonstandard regions, usually due to constraints, situations requiring a total number of runs not covered by classical designs, and for augmenting a set of additional runs to a previously run experiment. An example of a very irregular design region is the mixture design problem as a result of necessary constraints on the factor levels.

All of these issues are especially important in the case of split-plot response surface experiments. First, without OLS–GLS equivalence, Σ depends upon the variance components, which are unknown. As a result, the people generating the designs must perform robustness studies on the designs generated. Typically, these studies use ratios of σ_δ^2 to σ^2 of 0.1 to 5 or 10, which is often too limited to be of true practical value. Second, the designs generated are heavily dependent on the assumed model form. These designs often do not estimate subset models as the result of model reduction during analysis well and certainly not optimally. Third, D -optimal designs assume a one-shot experiment and do not deal effectively with the sequential design strategies inherent to RSM.

The greatest single issue deals with comparing designs with differing numbers of runs. The D -optimal designs for balanced designs with a fixed number of whole plots and number of subplots per whole plot is generally quite reasonable. Problems occur when the algorithms allow for differing numbers of runs either at the whole-plot or subplot level. The crucial issue becomes what is the number of runs. The first approaches defined the number of total runs by the rows of the X matrix. As a result, the important measure of design size is the number of subplot runs. The number of whole plots is irrelevant. These papers, taken to their logical conclusion, suggest that the preferred design is one subplot per whole plot, which is a completely randomized design. The D -optimal criterion essentially seeks to produce the best joint estimates of the model coefficients. The split-plot structure guarantees that the design produces less information about the whole-plot model terms than the subplot model terms. If the measure of design size is purely the number of subplot runs, then the algorithm must lead to the conclusion that a completely randomized design is optimal.

The real problem is that split-plot experiments have two sample sizes: one at the whole-plot level and one at the subplot level. In terms of the actual economics of the design, the number of whole plots is the true driver of expense. In many cases, subplot runs, up to some limit, are essentially free. The design’s goal then should be to minimize the number of whole plots subject to certain constraints.

Box (1982) outlines what he considers to be vital in the appropriate selection of an experimental design. Fundamental to Box’s perspective are two points. First, “All models are wrong; some models are useful.” Second, all scientific inquiry involves experimentation; and all scientific experimentation involves a series of experiments. Each experimental phase must build upon

what is learned from the previous phases. Both of these points stand in stark contrast to much of optimal design theory which typically assumes a single, “one-shot” experiment to estimate a known model form. According to Box, a “good” experiment design

1. provides a satisfactory distribution of information;
2. gives a fitted value as close as possible to the true;
3. provides the ability to test for lack of fit;
4. allows transformations;
5. allows for blocking;
6. allows sequential assembly (design augmentation);
7. provides an internal estimate of error;
8. is insensitive to the presence of outliers;
9. uses a near minimum number of runs;
10. provides data patterns that allow visual appreciation of the information in the data;
11. ensure simplicity of calculation;
12. behaves well when there are errors in the factors;
13. requires only a few levels for the factors; and
14. provides a check of the constancy of variance assumption.

Underlying many of these points is the concept of projection properties, which consider the design’s structure if one or more of the experimental factors proves insignificant in the analysis. Good projection properties ensure that the experimental design in the remaining factors maintains a good structure.

Box’s major point is that most classical experimental designs perform well in terms of these fourteen points in addition to performing well, although not necessarily best, in terms of the various optimality criteria. In this light, he strongly recommends the use of such classical experimental designs as the 2^k factorial system and the central composite design. The designs outlined in this chapter very much follow the Box philosophy of experimentation. They are “good” designs reflecting appropriate compromises across many competing criteria.

One may or may not agree with all of Box’s 14 points; however, they do underscore an extremely important issue. The final choice of any experimental design involves a complex compromise across many competing and often contradictory criteria. For example, the ability to detect lack-of-fit requires a certain number of experimental runs that may provide no meaningful information for estimating the presumed model. In a similar manner, pure error estimates of the σ^2 involve replication that often provides no additional information for estimating the model. Most statistical software packages that produce

optimal designs try to take these issues into consideration. In most cases, the packages outline the specifics in their help manuals.

REFERENCES

- Bisgaard, S. (2000). The design and analysis of $2^{k-p} \times 2^{q-r}$ split plot experiments. *Journal of Quality Technology*, **32**, 39–56.
- Bisgaard, S., H. Fuller, and E. Barrios (1996). Two-level factorials run as split-plot experiments. *Quality Engineering*, **8**, 705–708.
- Box, G.E.P. (1982). Choice of response surface design and alphabetic optimality. *Utilitas Mathematica*, **21B**, 11–55.
- Box, G.E.P. (1999). Statistics as a catalyst to learning by scientific method part II—A discussion. *Journal of Quality Technology*, **31**, 16–29.
- Box, G.E.P. and D.W. Behnken (1960). Some new three-level designs for the study of quantitative variables. *Technometrics*, **2**, 455–475.
- Box, G.E.P. and N.R. Draper (2007). *Response Surfaces, Mixtures, and Ridge Analysis* (2nd ed.). New York: John Wiley and Sons.
- Box, G.E.P. and K.B. Wilson (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society, Series B*, **13**, 1–45.
- Draper, N.R. and J.A. John (1998). Response surface designs where levels of some factors are difficult to change. *Australian and New Zealand Journal of Statistics*, **40**, 487–495.
- Goos, P. (2002). *The Optimal Design of Blocked and Split-Plot Experiments*. New York: Springer.
- Goos, P. (2006). Optimal versus orthogonal and equivalent-estimation design of blocked and split-plot experiments. *Statistica Neerlandica*, **60**, 361–378.
- Goos, P. and M. Vanderbroek (2001). Optimal split-plot designs. *Journal of Quality Technology*, **33**, 436–450.
- Goos, P. and M. Vanderbroek (2003). D-optimal split-plot designs with given numbers and sizes of whole plots. *Technometrics*, **45**, 235–245.
- Goos, P., I. Langhans, and M. Vanderbroek (2006). Practical inference from industrial split-plot designs. *Journal of Quality Technology*, **38**, 162–179.
- Jones, B. and P. Goos (2007). A candidate-set free algorithm for generating D-optimal split-plot designs. *Applied Statistics*, **56**, 347–364.
- Jones, B. and C.J. Nachtsheim (2009). Split-plot designs: What, why, and how. *Journal of Quality Technology*, **41**, 340–361.
- Khuri, A.I. and J.A. Cornell (1996). *Response Surfaces: Designs and Analyses* (2nd ed.). New York: Marcel Dekker.
- Letsinger, J.D., R.H. Myers, and M. Lentner (1996). Response surface methods for bi-randomization structures. *Journal of Quality Technology*, **28**, 381–397.
- Montgomery, D.C. (2001). *Design and Analysis of Experiments* (5th ed.). New York: John Wiley and Sons.
- Myers, R.H., D.C. Montgomery, and C.M. Anderson-Cook (2009). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments* (3rd ed.). New York: John Wiley and Sons.

- Parker, P.A., S.M. Kowalski, and G.G. Vining (2006). Classes of split-plot response surface designs for equivalent estimation. *Quality and Reliability Engineering International*, **22**, 291–305.
- Parker, P.A., S.M. Kowalski, and G.G. Vining (2007). Construction of balanced equivalent estimation second-order split-plot designs. *Technometrics*, **49**, 56–65.
- Trinca, L.A. and S.G. Gilmour (2001). Multistratum response surface designs. *Technometrics*, **43**, 25–33.
- Vining, G.G. and S.M. Kowalski (2008). Exact inference for response surface designs within a split-plot structure. *Journal of Quality Technology*, **40**, 394–406.
- Vining, G.G., S.M. Kowalski, and D.C. Montgomery (2005). Response surface designs within a split-plot structure. *Journal of Quality Technology*, **37**, 115–129.

Design and Analysis of Experiments for Directional Data

Sango B. Otieno and Christine M. Anderson-Cook

15.1 SUMMARY

Directional data arise in a variety of applications, and while designed experiments are moderately rare for circular, cylindrical, and spherical data, there has been some work on the design and analysis of experiments in this area. In this chapter, we describe the essential characteristics of directional data, and then examine the currently available tools to analyze data from one-way and multiway classifications. The designs considered are generally basic, typically consisting of completely randomized designs with multiple treatments, or full factorial designs for experiments involving more than one factor. The scope of the chapter is to describe the analysis methods for different problems of interest within directional data, with a stronger emphasis on methods for circular data, since these are most abundant in the directional data literature.

15.2 INTRODUCTION AND HISTORICAL BACKGROUND

Directional observations are common in biology and the earth sciences and arise naturally in the study of plant and animal behavior. Batschelet (1981) and Mardia (1972) give examples, which included the vanishing angles of 714 mallard ducks released after being displaced by varying distances from their normal habitat, the orientations of turtles after an experimental treatment, and the angles between the swimming directions of Daphnia and the plane of

Design and Analysis of Experiments: Special Designs and Applications, First Edition. Edited by Klaus Hinkelmann.

© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

polarization of light. Directional data often are divided into directed (circular) and undirected (axial) measurements. Examples of circular data include wind and paleomagnetic measurements. Pebble orientations, B-axial lineations and fold axes are examples of axial data. Directional data can be measured in two dimensions, such as wind directions, or three dimensions, such as fault lines. In higher dimensions, a p -dimensional directional sample value is denoted by a unit vector OP starting at the center O of a hypersphere of radius one finishing at a point P on the surface of the hypersphere. Although a p -dimensional unit vector does not have a physical interpretation in terms of direction, a set of vectors whose components are continuous proportions might be usefully analyzed using techniques for directional data. In this chapter we focus on 2- and 3-dimensional applications with a readily available physical interpretation.

Interest in directional data analysis dates back at least to the mathematicians and astronomers of the 18th century. Indeed, the theory of errors was developed by Gauss primarily to analyze certain directional measurements in astronomy. It is a historical accident that the observational errors were sufficiently small to allow Gauss to make a linear approximation, and, as a result, he developed a linear rather than a directional theory of errors (Mardia 1972, p. xvii). However, a major roadblock to advancements in this area has been computational complexity. Despite the rapid development of specialized methods for directional statistics over the last 30 years, software has only recently become available to make such methods easy to use. The circular package (Ulric Lund and Claudio Agostinelli, at <http://www.cran.r-project.org/>) for the R programming language, as well as the circstats package (Nicholas J. Cox) for Stata, are both available as open source and offer a variety of functions to users of these programming languages. In addition, the commercially available Oriana software (Computing Service Kovach, at <http://www.kovcomp.co.uk/oriana/index.html>) provides basic analysis functionality for Windows users. Other software include CircStat2009 toolbox (Max Planck Institute for Biological Cybernetics) for circular statistics and the DDSTAP package developed by Ashis SenGupta (Indian Statistical Institute, Calcutta).

This chapter concentrates on the design and analysis of experiments of directional data. Because many of the applications of directional data are observational in nature, there are relatively few designed experiments in the statistical literature. The types of experiments considered are also relatively simple with typically one or two linear or categorical factors being manipulated to produce a directional response. As such, the chapter focuses primarily on the analysis tools available for one- and two-way classifications that yield a circular, cylindrical, or spherical response. In Section 15.2.2, we briefly consider the case where one or more input factors are directional in nature.

15.2.1 Overview of Directional Data

Directional data are unique from more common data measured on a linear scale for two primary reasons: there are no natural maximum or minimum

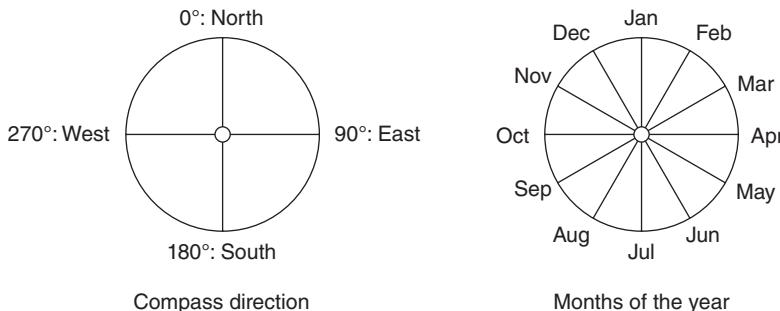


Figure 15.1 Examples of data on a circular scale.

values because of the wrap-around nature of the circle or sphere; and the position of the zero direction is arbitrary (see Figure 15.1). Hence, the analysis and interpretation of directional data requires specific data representations, descriptions, and distributions, as well as rotationally invariant methods. Excellent references for directional data include Fisher (1993), Mardia (1972) and Fisher, Lewis, and Embleton (1987).

Directional data are often characterized in terms of unit length vectors, or by representing the angular observations as points on the unit radius circle or sphere. For circular data, the angle, θ , is usually specified in degrees in the range $(-180^\circ, 180^\circ)$ or $(0^\circ, 360^\circ)$, or in radians in the range $(-\pi, \pi)$ or $(0, 2\pi)$. Cylindrical data are comprised of an angular measure, θ , as defined above and a linear measurement, y . Spherical data are represented in polar form by specifying the azimuth (or declination) and the latitude (or inclination). The azimuth, ϕ , is given in degrees with range $(-180^\circ, 180^\circ)$. The latitude (also called elevation angle), θ , is specified in $(-90^\circ, 90^\circ)$. Instead of an azimuth and latitude, a longitude angle in $(0^\circ, 360^\circ)$ and a co-latitude angle in $(0^\circ, 180^\circ)$ can also be used.

In the following sections, we assume, unless stated otherwise, that circular data are specified in radians in the interval $(0, 2\pi)$, and spherical data are specified by the pair (azimuth, latitude) both in degrees. Circular data are usually plotted with a marker for each direction plotted over the corresponding point on the unit circle. Cylindrical data combine a circular measure with a linear measure to give a location on a cylinder with a unit radius. Spherical data are conveniently represented showing a projection of the unit sphere with markers over the points corresponding to the directions.

When dealing with directional data, it can be convenient to obtain the Cartesian (rectangular) coordinates and to convert between this form and angular, polar or spherical forms. The conversion formulas are given in Table 15.1, and the relationships are illustrated in Figure 15.2.

Note, ρ is the mean resultant length, obtained from the 1st trigonometric central moment of the circular distribution function $f(\theta)$ given by

Table 15.1 Conversion Formulas between Cartesian and Polar or Spherical Co-ordinates (Azimuths and Latitudes)

	Polar to Cartesian	Cartesian to Polar ^a
Circle	$(\theta, \rho) \rightarrow (x, y)$ $x = \rho \cos \theta; y = \rho \sin \theta$	$(x, y) \rightarrow (\theta, \rho)$ $\phi = \tan^{-1}(y, x); \rho = (x^2 + y^2)^{\frac{1}{2}}$
Sphere	$(\phi, \theta, \rho) \rightarrow (x, y, z)$ $x = \rho \cos \theta \cos \phi; y = \rho \cos \theta \sin \phi;$ $z = \rho \sin \theta$	$(x, y, z) \rightarrow (\phi, \theta, \rho)$ $\theta = \arctan\left(z / (x^2 + y^2)^{\frac{1}{2}}\right);$ $\phi = \tan^{-1}(y, x); \rho = (x^2 + y^2 + z^2)^{\frac{1}{2}}$

^a $\tan^{-1}(y, x)$ denotes the inverse tangent of y/x with correction of the angle for $x < 0$.

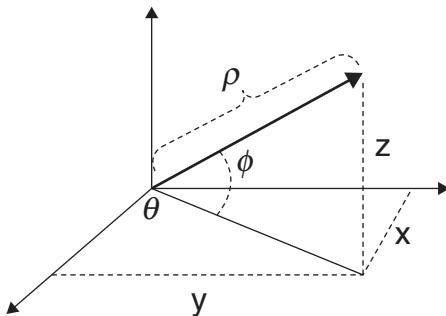


Figure 15.2 The relation between the Cartesian (x, y, z) and spherical (ϕ, θ, ρ) coordinate systems.

$$\begin{aligned}\mu &= \rho e^{i(\theta-\mu)} = \rho \cos(\theta - \mu) + i \rho \sin(\theta - \mu) \\ &= \int_0^{2\pi} \cos(\theta - \mu) f(\theta) d\theta + i \int_0^{2\pi} \sin(\theta - \mu) f(\theta) d\theta,\end{aligned}$$

where μ is the mean direction.

A circular distribution has its total probability on the circumference of a unit circle. See Jammalamadaka and SenGupta (2001, pp. 25–63) for a detailed discussion of circular probability distributions. Two frequently used families of distributions for circular data are the von Mises and Uniform distributions. The importance of the von Mises distribution is similar to the Normal distribution on the line (Mardia 1972). It was introduced by von Mises (1918), and is a symmetric unimodal distribution characterized by a mean direction μ , and concentration parameter κ , with probability density function

$$f(\theta) = [2\pi I_o(\kappa)]^{-1} \exp[\kappa \cos(\theta - \mu)] \quad 0 \leq \theta, \mu < 2\pi, \quad 0 \leq \kappa < \infty, \quad (15.1)$$

where

$$I_o(\kappa) = (2\pi)^{-1} \int_0^{2\pi} \exp[\kappa \cos(\phi)] d\phi = \sum_{j=0}^{\infty} -\frac{1}{(j)^2} \left(\frac{\kappa^2}{4}\right)^j,$$

is the modified Bessel function of order zero. See Fisher (1993, p. 50) for a series expansion and methods for evaluating $I_o(\kappa)$. κ is a concentration parameter, which quantifies the dispersion. As κ increases from zero, $f(\theta)$ peaks higher about μ . Note, we say that the circular random variable θ is symmetric about a given direction μ if its distribution has the property $f(\mu + \theta) = f(\mu - \theta)$, for all θ , where addition or subtraction is modulo 2π .

If κ is zero, then $f(\theta) = \frac{1}{2\pi}$ for all angles on the circle, and the distribution is the circular uniform distribution with no preferred direction, and the total probability is spread out uniformly on the circumference of a circle.

Given a set of circular observations $\theta_1, \dots, \theta_n$, each observation is measured as a unit vector with coordinates from the origin of $(\cos \theta_i, \sin \theta_i)$, $i = 1, \dots, n$. The resultant vector of these n unit vectors is obtained by summing them componentwise to get $R = (\sum_{i=1}^n \cos \theta_i, \sum_{i=1}^n \sin \theta_i) = (C, S)$, say. The sample circular mean, $\bar{\theta}$, is the angle corresponding to the mean resultant length ($\bar{R} = R/n$). The usual (maximum likelihood) estimate $\hat{\mu}$ of the mean direction μ is the sample mean direction $\hat{\mu} = \bar{\theta}$. The maximum likelihood estimate $\hat{\kappa}_{ML}$ of κ (given below) is the solution of the equation $A_1(\hat{\kappa}_{ML}) = \bar{R}$, where \bar{R} is the mean resultant length of the sample, and $A_1(x) = I_1(x)/I_0(x)$, the ratio of two modified Bessel functions. Approximations to solve for the estimates of κ in closed form are dependent on the size of \bar{R} , and are as follows:

$$\hat{\kappa}_{ML} = \begin{cases} 2\bar{R} + \bar{R}^3 + 5\bar{R}^5 / 6, & \bar{R} < 0.53 \\ -0.4 + 1.39\bar{R} + 0.43/(1 - \bar{R}), & 0.53 \leq \bar{R} < 0.85 \\ 1/(\bar{R}^3 - 4\bar{R}^2 + 3\bar{R}), & \bar{R} \geq 0.85. \end{cases} \quad (15.2)$$

To correct for bias (which occurs when the sample size $n \leq 15$ and $\bar{R} < 0.45$), the estimate of κ is obtained by

$$\hat{\kappa} = \begin{cases} \max(\hat{\kappa}_{ML} - 2(n\hat{\kappa}_{ML})^{-1}, 0), & \hat{\kappa}_{ML} < 2 \\ (n-1)^3 \hat{\kappa}_{ML} / (n^3 + n), & \hat{\kappa}_{ML} \geq 2. \end{cases} \quad (15.3)$$

See Fisher (1993) for more details.

Suppose for a single population of circular data located on the circumference of a unit circle, we obtain a sample of angles, $\theta_1, \dots, \theta_n$. We can treat the θ_i 's as vectors of unit length, u_i , starting at the origin and pointing in the direction of their angle. From this vector representation, we can obtain the resultant vector, $u = \sum_i u_i$. We standardize the length of the resultant vector to lie on the

interval $(0, 1)$ and hence obtain $\bar{R} = (c^2 + s^2)^{\frac{1}{2}} / n$, where $c = \sum_i \cos \theta_i$ and $s = \sum_i \sin \theta_i$. The mean or average direction for the sample, θ , is defined to be the angle of u .

$$\theta = \begin{cases} \arctan(s/c), & \text{if } c > 0 \\ \frac{\pi}{2}, & \text{if } c = 0 \text{ and } s > 0 \\ -\frac{\pi}{2}, & \text{if } c = 0 \text{ and } s < 0 \\ \pi + \arctan(s/c), & \text{otherwise.} \end{cases} \quad (15.4)$$

We use the notation “.” to denote summation over that index of the angles or resultant lengths. We define the unit vector associated with the average direction to be $u/\|u\|$.

The resultant vector has two characteristics that match the polar representation: the average direction and the standardized length (\bar{R}). This length can be thought of as a measure of the dispersion, since the circular variance $(1 - \bar{R})$ is a function of this length. See Mardia (1972) for details. In some applications, the objective of the study is to determine the optimal combination of factor levels for achieving some predetermined target direction. This location-only effect parallels the usual factor effects in analysis of variance (ANOVA) methods for linear data. A nonsignificant effect exists if $\theta_{1.} \approx \dots \approx \theta_{p.}$ for all groups. A dispersion-only measure strives to quantify differences in the spread of the directional response, in an effort to limit the range of the response. A nonsignificant measure reflects similar variation within each group, and occurs if $\bar{R}_{1.} \approx \dots \approx \bar{R}_{p.}$ for all groups. Finally, a combined location and dispersion measure considers both the mean and the spread simultaneously. A main effect is equal to zero if both of the above requirements are satisfied. Since it is generally considered best to examine location and dispersion effects on the response separately for better understanding of the nature of the relationship between factors and response, this last measure is likely of limited practical use, except in specialized situations.

For the balanced one-way situation with p groups, we obtain m sample values, $\theta_{i1}, \dots, \theta_{im}$, for the i^{th} group, with i in $1, \dots, p$. The resultant vector for each group is obtained as above, giving the following summary statistics: $\theta_{1.}, \dots, \theta_{p.}$ and $\bar{R}_{1.}, \dots, \bar{R}_{p.}$ For the two-way situation, the resultant vectors for all of the rows, columns, and cells are obtained, and summarized by their angles and lengths. Hence, $\theta_{i..}$ and $\theta_{..j}$ are the resultant angles for the i^{th} row and the j^{th} column, respectively, and $\bar{R}_{ij..}$ is the standardized length of $u_{ij..}$ for the ij^{th} cell.

Statistical theory developed by von Mises (1918) and by Fisher (1953), and extended by Watson and Williams (1956), Stephens (1962a, 1962b, 1962c, 1962d) and Mardia (1975), considers the distribution of points on the surface of a hypersphere in p dimensions. The von Mises–Fisher distribution (known as the von Mises and Fisher in two and three dimensions, respectively) is sym-

metrical about a polar axis and attains a maximum density at the pole and a minimum density at the antipole. Its parameter κ determines the clustering of the distribution of points around the polar axis (or equivalently, the degree of clustering of the distribution of vectors in the p -dimensional space). Note that higher values of κ generate a concentrated distribution around the polar axis, while a zero value generates a uniform distribution over the surface of the p -dimensional hypersphere. A unit random vector x has the $(p-1)$ -dimensional von Mises–Fisher (or Langevin) distribution $M_p(\mu, \kappa)$ if its probability density function with respect to the uniform distribution is

$$f(x; \mu, \kappa) = \left(\frac{\kappa}{2} \right)^{p/2-1} \frac{1}{\Gamma(p/2) I_{p/2-1}(\kappa)} \exp\{\kappa \mu' x\}, \quad (15.5)$$

where $\kappa \geq 0$, $\|\mu\| = 1$ and I_v denotes the modified Bessel function of the first kind and order v , defined as

$$I_v(\kappa) = \frac{1}{2\kappa} \int_0^{2\pi} \cos v\theta e^{\kappa \cos \theta} d\theta. \quad (15.6)$$

On the other hand, when observations are not directions but axes, the unit vectors x and $-x$ are indistinguishable. In this case, it is $\pm x$ that is observed. Such data, referred to as axial data on the sphere (Watson 1983; Fisher, Lewis, and Embleton 1987; Mardia and Jupp 2000), are modeled using the Watson distribution. Mardia and Jupp (2000) proposed the one-way ANOVA technique for samples of unit axes, which are assumed to come from the Watson distribution, while Figueiredo (2006) proposed a nested two-way ANOVA for a concentrated bipolar Watson distribution. The dimension of the data is p , where the notation talks in terms of $p-1$, because of the constraint to place the data on the surface of the sphere. Defined on the unit sphere in \Re^p , $S_{p-1} = \{x \in \Re^p : x'x = 1\}$, and usually denoted by $W(u, \xi)$, where $u \in S_{p-1}$ is a directional parameter, $\xi > 0$ is a concentration parameter; it has probability density function given by

$$f(x) = \left\{ F_1 \left(\frac{1}{2}, \frac{p}{2}, \xi \right) \right\}^{-1} \exp \xi (u' x)^2 \quad \text{with } x \in S_{p-1}, u \in S_{p-1}, \xi \in \Re, \quad (15.7)$$

where the normalizing constant is the reciprocal of the confluent hypergeometric function $F_1(\cdot)$, which is defined by

$$F_1 \left(\frac{1}{2}, \frac{p}{2}, \xi \right) = \frac{\Gamma(p/2)}{\Gamma(1/2)\Gamma(p-1)/2} \int_0^1 \exp(\xi t) t^{-0.5} (1-t)^{(p-3)/2} dt. \quad (15.8)$$

This distribution has a directional parameter u and a concentration parameter ξ . For $\xi > 0$, the distribution has a maximum at u , and so it is bipolar. As ξ increases, the distribution becomes more concentrated around u . For $\xi < 0$,

the distribution is concentrated around the great circle orthogonal to \mathbf{u} , and so the distribution is a symmetric girdle. For $\xi = 0$, the distribution is uniform. See Mardia and Jupp (2000, p. 181–236), for details about the properties of this distribution.

15.2.2 Existing Designs for Directional Data

Many of the previously studied data sets with a directional response are from observational sources, such as biology, astronomy, and geology, with the goal of studying location. Designed experiments involving one or more factors thought to influence the directional responses have stemmed from an analogous strategy to ANOVA used for linear data. Typically, directional data designed experiments are comprised of one or more factors (generally small numbers) with a traditional form (either categorical or on a linear scale) with a directional response (circular, cylindrical, or spherical). The structure of published experiments involving directional data is quite simple, with no known examples of methods for restricted randomization or more advanced forms. While the authors are familiar with several applications that could utilize design and analysis methods for these types of experiments, there do not appear to be published methods available.

A typical example of a directional data designed experiment from the literature is a completely randomized design (see HK1, chapter 6) in which a (von Mises) variate θ is measured for each of r “levels” of a single classification. In this case, the observations at the i^{th} level constitute the i^{th} sample. More generally, we can have $s \geq 2$ classifications, each with several levels (a cross-classified or factorial experimental layout) with a sample of several observations made at each combination of levels. An example of a designed experiment based on a two-factor factorial considers the direction of movement of animals placed at the center of a regular n -sided polygon. The explanatory variable was the location of a signal emitted from one of the vertices, and the direction in which the animal subsequently moved was recorded as that of the closest vertex (Graves 1979). As with this example, there is often limited precision of measurements, with data rounded and then recorded to the nearest 1° , 5° , 10° , 15° , or 30° .

There are a few examples of designs involving two or more factors, such as Harrison and Kanji (1988), Anderson and Wu (1995), and Anderson-Cook (2001). In these cases, the error-control designs considered are again quite elementary with the use of factorial designs as treatment designs. Although there are some industrial applications that lend themselves to experimental situations, many of the directional data applications are biological, geological, or astronomy related, and hence lead to observational studies where it is rare to be able to flexibly manipulate multiple input factor levels.

An alternate type of experiment that we consider briefly here involves one or more directional data inputs, which are thought to impact a linear response. For example, we might conduct an experiment to test reaction time and con-

sider time of day as a directional measure. Time can naturally be thought of as a directional (circular) measure, since there is a continuity of time of day that makes the designation of a new day starting and ending at midnight arbitrary. Laycock (1975) presents an overview of designs and methods for analyzing this type of experiment. An optimal design for best prediction of model parameters (as measured by D -optimality) suggests equal spacing of observations around all hours of the clock (Karlin and Studden 1966). For the analysis, assign an angle, θ , to represent times on the 24-hour clock, perhaps with midnight corresponding to 0° , 6 a.m. corresponding to 90° , 12 noon corresponding to 180° , etc. The form of the model is assumed to be

$$y = \beta_0 + \beta_1 \cos \theta + \beta_2 \sin \theta + \dots + \beta_{2m-1} \cos(m\theta) + \beta_{2m} \sin(m\theta) + \varepsilon, \quad (15.9)$$

which is functionally equivalent to the model

$$y = \beta_0 + R_1 \cos(\theta - \theta_1^*) + \dots + R_m \cos(m\theta - \theta_m^*) + \varepsilon. \quad (15.10)$$

The first form of the model has the advantage of being able to be estimated with linear regression techniques, but can be transformed to the second using the Cartesian–Polar conversion identities to identify the amplitudes, R_i , and locations of the maximum amplitudes for the various frequencies of effects, θ_i^* , which are perhaps more immediately interpretable than the terms in the first equation. This model connects with the form used for cylindrical data considered later in the chapter.

In the remainder of the chapter, we focus on the design and analysis of experiments with linear or categorical inputs and a directional response.

15.3 ANOVA FOR CIRCULAR DATA

ANOVA for this simplest type of directional data assumes that we have conducted an experiment where the assignment of treatments to the experimental units was randomized and has replication within each of the treatment groups. Again, this simple design structure is adequate for many of the documented directional applications.

We now consider the different options that are available based on the focus of the analysis and the dispersion of the observed data. Some of the methods are sensitive to data within a group being too diffuse around the circumference of the circle, while others make adaptations depending on the size of the estimated von Mises concentration parameter, κ .

The foundation of ANOVA type techniques for circular data was laid by Watson and Williams (1956) and enhanced by Stephens (1962a, 1962b, 1969, 1972, 1982) and Upton (1974).

ANOVA for distributions other than von Mises have also been proposed by Mardia and Spurr (1973) for l -modal von Mises distributions and

Anderson-Cook (2000) for cylindrical data. Some work has been done to consider how to obtain a good design: Rao and Sengupta (1970), consider the problem of optimally allocating design points in a hierarchical design with a fixed cost, while Wu (1997) considers the choice of the optimal location of the points to be sampled on a circle and develop Φ -optimal exact and approximate designs on a circle.

15.3.1 One-Way ANOVA

In this section, we consider the class of experiments with analyses to look for differences in the mean direction, the spread of the data within groups, or both for experimental data based on a completely randomized design with m groups and N total observations, with potentially unequal numbers of observations, N_i , per group. Each of the observations is a point on the unit circle or a direction. Typically, differences in the mean, or preferred, direction are most commonly of interest, but there are also situations when primary focus may be on the spread of the data within groups. Occasionally, we are interested in evaluating the hypothesis that the mean and spread of groups simultaneously have no differences, although we consider this a less desirable test, since if differences are found, then the nature of those differences may be more difficult to resolve. Note that many of the tests rely on correction factors to help adjust the test statistic to obtain a good match to the assumed distribution. After the methods have been introduced, an example is used to illustrate the various techniques.

15.3.1.1 Tests of Equal Mean Directions

We first consider testing the hypothesis

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_q \quad H_1 : \text{Not all } \mu_i \text{s are equal.} \quad (15.11)$$

This test is most directly analogous to the ANOVA for data on a linear scale. Different tests are for various combinations of conditions: whether the κ_i are assumed equal or not, whether the data are from a von Mises or other distribution, and whether the concentrations and number of replicates are large or small. Let $N = \sum N_i$. Define $\tilde{\kappa}$ to be the median of the κ_i , \bar{R}_m to be the median of the \bar{R}_i , and N_{\min} to be the smallest of the number of replicates N_i .

Likelihood Ratio Test

Assume that the data follow the von Mises distribution (see Eq. 15.1) and that all of the concentrations κ_i are equal (say, to κ). For small κ (but not too near 0), calculate

$$U = 2 \left[\left(\sum R_i \right)^2 - R^2 \right] / N, \quad (15.12)$$

and correction factor

$$c = [1 - \kappa^2 / 8 + q / (2N\kappa^2)]^{-1}. \quad (15.13)$$

We note that R is calculated as in Section 15.2.1, and R_i is the mean direction for group i . Assuming H_0 is true, cU is distributed approximately as chi-square with $q - 1$ degrees of freedom if N is large. Reject H_0 for large values of cU , Mardia and Jupp (2000, p. 137). Note that if κ values are near 0, the test will not have much power to reject H_0 except for extremely large replicates, since this level of κ spreads the probability almost uniformly over 360° . This matches our intuition from the linear data case, where for very diffuse samples, it is difficult to distinguish differences in the mean.

Multisample Watson–Williams Test

Assume that the data follow the von Mises distribution, and that all of the κ_i are equal (say, to κ). This test applies for larger values of κ than does the likelihood ratio test. Calculate

$$T_r = \frac{\left(\sum R_i - R\right)/(q-1)}{\left(N - \sum R_i\right)/(N-q)}. \quad (15.14)$$

Under H_0 , T_r is distributed approximately as F with $q - 1$ and $N - q$ degrees of freedom if N is large. Reject H_0 for large values of T_r . Mardia and Jupp (2000, p. 135) recommend this test for $\kappa \geq 1$, but Fisher (1993, pp. 126–127) suggests that κ should exceed 2 unless T_r is multiplied by $1 + 0.375/\tilde{\kappa}$.

Embedding Method

The embedding approach proposed by Harrison, Kanji, and Gadsden (1986) considers the observations as unit vectors on the plane. The total variation in \bar{R} is separated into components that represent variation between and within the q groups, analogously to ANOVA. The test assumes data from a von Mises distribution and the concentration parameter values, κ , to be large. Calculate

$$U = \frac{\left(\sum N_i \bar{R}_i^2 - N \bar{R}^2\right)/(q-1)}{\left(N - \sum N_i \bar{R}_i^2\right)/(N-q)}, \quad (15.15)$$

with correction factor,

$$c = 1 - 0.2/\hat{\kappa} - 0.1/\hat{\kappa}^2. \quad (15.16)$$

Under H_0 , cU is distributed approximately as F with $q - 1$ and $N - q$ degrees of freedom if N is large. Reject H_0 for large values of cU , Harrison, Kanji, and Gadsden (1986, pp. 135–136); Mardia and Jupp (2000, p. 139).

Heterogeneous Case

Assume that the data follow the von Mises distribution, but we do not assume that all of the κ within treatment groups are equal. This test relies on large sample sizes for the distributional assumptions to hold. Calculate

$$R_W = \left[\left(\sum \hat{\kappa}_i R_i \cos \bar{\theta}_i \right)^2 + \left(\sum \hat{\kappa}_i R_i \sin \bar{\theta}_i \right)^2 \right]^{\frac{1}{2}}, \quad (15.17)$$

and hence

$$Y_r = 2 \left(\sum \hat{\kappa}_i R_i - R_W \right). \quad (15.18)$$

Under H_0 , Y_r is distributed approximately as chi-square with $q - 1$ degrees of freedom for large N . Reject H_0 for large values of Y_r , Mardia and Jupp (2000, p. 137).

Nonparametric Test

Now we relax the assumption that the data are from the von Mises distribution, and make no distributional assumptions. We do not need to assume equal dispersions because the test calculates values to select one of two possible tests. Compute estimated dispersion values $\hat{\delta}_i$, $i = 1, 2, \dots, q$, from the data (Fisher 1993, p. 34), where

$$\hat{\delta}_i = \left[1 - \sum \cos 2(\theta_{ij} - \bar{\theta}_i) / N_i \right] / (2\bar{R}_i^2). \quad (15.19)$$

The method of testing differs depending on how similar the spread of the different groups are. This is characterized by considering the ratio of the extreme dispersions. If $\hat{\delta}_{\max} / \hat{\delta}_{\min} \leq 4$, use the P method, otherwise use the M method (Fisher 1993, pp. 116–117).

The P method is used when the $\hat{\delta}_i$ are comparable. Calculate $Y_r = (N - R_p) / \hat{\delta}_0$, where $R_p = \left[(\sum N_i \cos \bar{\theta}_i)^2 + (\sum N_i \sin \bar{\theta}_i)^2 \right]^{\frac{1}{2}}$ and $\hat{\delta}_0 = \sum N_i \hat{\delta}_i / N$.

On the other hand, if the $\hat{\delta}_i$ are not comparable, use the M method. Calculate $Y_r = 2 \left[\sum \left(1/\hat{\delta}_i^2 \right) - R_M \right]$, where $R_M = \left[\left[\sum (\cos \bar{\theta}_i) / \hat{\delta}_i^2 \right]^2 + \left[\sum (\sin \bar{\theta}_i) / \hat{\delta}_i^2 \right]^2 \right]^{\frac{1}{2}}$.

For either the P or M method, if H_0 is true, Y_r is distributed approximately as chi-square with $q - 1$ degrees of freedom for large N . Reject H_0 for large values of Y_r . Fisher (1993) recommends this test only if the number of observations in all groups exceeds 24.

15.3.1.2 Tests of Equality of Concentrations

We now consider a hypothesis that focuses on the spreads of the data within samples when comparing different treatment groups. More formally, consider testing the hypothesis

$$H_0 : \kappa_1 = \kappa_2 = \dots = \kappa_q = \kappa \quad H_1 : \text{Not all } \kappa_i's \text{ are equal.} \quad (15.20)$$

for data assumed to come from a von Mises distribution. Different tests exist depending on whether the concentrations, κ_i , and the number of replicates are large or small. Also, there are alternative tests for data not assumed to come from a von Mises distribution. Recall that $N = \sum N_i$, $\tilde{\kappa}$ is calculated as the median of the $\hat{\kappa}_i$, \bar{R}_m is the median of the \bar{R}_i , and $N_{\min} = \min(N_i)$.

Small Concentrations

Assume that the concentration parameter $\kappa < 1$. There are no constraints on the vector means Mardia and Jupp (2000, p.140) give the test statistic

$$U_1 = \sum w_i g_1^2(2\bar{R}_i) - \left[\sum w_i g_1(2\bar{R}_i) \right]^2 / \sum w_i, \quad (15.21)$$

where $w_i = 4/(N_i - 4)$, $g_1(2\bar{R}) = \sin^{-1}(2a\bar{R})$, and $a = 0.61237$. Under H_0 , U_1 is distributed approximately as chi-square with $q - 1$ degrees of freedom for large N . Reject H_0 for large values of U_1 .

Medium Concentrations

This test assumes the concentration κ are in the range $(1, 2)$ Mardia and Jupp (2000, p. 140) propose

$$U_2 = \sum w_i g_2^2(2\bar{R}_i) - \left[\sum w_i g_2(2\bar{R}_i) \right]^2 / \sum w_i, \quad (15.22)$$

where $w_i = (N_i - 3)/0.79791$, and $g_2(2\bar{R}) = \sin^{-1}[(\bar{R} - 1.0894)/0.25789]$. Under H_0 , U_2 is distributed approximately as chi-square with $q - 1$ degrees of freedom for large N . Reject H_0 for large values of U_2 .

Larger Concentrations

This test is designed for high concentrations with $\kappa > 2$. In this situation, Mardia and Jupp (2000, p. 140) recommend Bartlett's test of homogeneity. Define $v_i = N_i - 1$, $v = N - q$, and

$$d = \left(\sum \frac{1}{v_i} - \frac{1}{v} \right) / 3(q-1).$$

Then the test statistic is

$$U_3 = \left[v \log \left(\frac{N - \sum R_i}{v} \right) - \sum v_i \log \left(\frac{N_i - R_i}{v_i} \right) \right] / (d+1). \quad (15.23)$$

Under H_0 , U_3 is distributed approximately as chi-square with $q - 1$ degrees of freedom for large N . Reject H_0 for large values of U_3 .

Tangential Approach

Now we consider a different method proposed by Fisher (1986), called the tangential approach. The name comes from the idea that the deviation between two angles, say ψ and ϕ , can be measured by $\sin(\psi - \phi)$, which can be thought of as the length of the tangent to a circle at ψ . The test consists of forming a variable $d_{ij} = |\sin(\theta_{ij} - \bar{\theta}_i)|$, $i = 1, 2, \dots, q$, $j = 1, 2, \dots, N_i$, and performing a one-way ANOVA. Calculate the test statistic

$$f_r = \frac{(N-q)\sum N_i (\bar{d}_i - \bar{d})^2}{(q-1)\sum \sum (d_{ij} - \bar{d}_i)^2}, \quad (15.24)$$

where $\bar{d}_i = \sum_j d_{ij} / N_i$ and $\bar{d} = \sum \sum d_{ij} / N$. Under H_0 , f_r is distributed approximately as F with $q-1$ and $N-q$ degrees of freedom. Reject H_0 for large values of f_r . Fisher (1993, pp. 131–132) recommends this test because it is insensitive to outliers and departures from the von Mises distribution; see also Mardia and Jupp (2000, p. 139).

15.3.1.3 Test Equality of Mean Directions and Equality of Concentrations

This tests the compound null hypothesis simultaneously considering location and dispersion effects, formally stated as

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_q \text{ AND } \kappa_1 = \kappa_2 = \dots = \kappa_q, \quad (15.25)$$

against the alternative H_1 : Not all μ_i 's are equal **OR** Not all κ_i 's are equal. This test is appropriate if the κ_i are small, but a large N is required. Compute

$$U = 2\left(\sum R_i^2 / N_i - R^2 / N\right). \quad (15.26)$$

Under H_0 , U is distributed approximately as chi-square with $2(q-1)$ degrees of freedom for large N . Reject H_0 for large values of U , Mardia and Jupp (2000, p. 138).

15.3.1.4 Resampling

A common problem for inference is that the exact sampling distribution of a test statistic may be unknown. In many situations, large-sample approximations to distributions are used to allow testing. However, these approximations may be poor for small-to-moderate sample sizes. Resampling methods have been developed in many applications, including directional data, to deal with such problems. These are generally computationally intensive, but modern computers allow resampling to be done almost routinely. The flexibility to relax distributional assumptions is often a more important consideration than the computer run-time required. Hence, these methods have become increasingly popular. Fisher (1993, chapter 8) gives detailed summaries of two types of resampling: bootstrap and randomization methods.

The general approach to resampling is to draw randomly from the observed data to make new samples, but in such a way that the null hypothesis H_0 is assumed true. The newly drawn samples are used to calculate the test statistic of interest. The process is repeated many times, say 500 to 1000, to generate an estimate of the distribution of the test statistic assuming H_0 is true.

Test Equality of Mean Directions Using von Mises Distribution

This test uses the von Mises assumption and $H_0 : \mu_1 = \mu_2 = \dots = \mu_q$. Here we use the test statistic Y_r described in Equation (15.18). Calculate Y_r as described above; this will be the test statistic Y'_r . Fisher (1993, p. 124) recommends the procedure presented in his section 8.4.4, but with sampling from the von Mises distribution, using Monte Carlo simulation.

To perform this test, randomly draw observations from the von Mises distribution (Fisher 1993, p. 49) under H_0 for the q subsamples, that is, each subsample i has N_i observations drawn from the distribution with vector mean equal to 0 and concentration κ . Calculate Y_r using these randomly drawn angles. Repeat the process many times, say $N_R = 1000$ times, to approximate the distribution of the test statistic Y_r under the assumption that H_0 is true. Sort these Y_r values in ascending order, and find the cutoff value Y_r^* that is the 100α largest value. All values greater than Y_r^* are in the critical portion of the distribution. Hence, reject H_0 if $Y'_r > Y_r^*$.

Test Equality of Mean Directions: Nonparametric Case

This tests the hypothesis $H_0 : \mu_1 = \mu_2 = \dots = \mu_q$, but does not assume a specific probability distribution for the data. Here, we use the test statistic Y_r described in Equation (15.18). Calculate Y_r as described above; this will be the test statistic Y'_r . Fisher (1993, p. 115) recommends the following resampling procedure. With no assumed distribution, bootstrap resampling is used.

First, center the observed data for each subsample by subtracting the observed mean for the subsample, as $\Psi_{ij} = \theta_{ij} - \bar{\theta}_i$, $i = 1, 2, \dots, q$, $j = 1, 2, \dots, N_i$. Centering the data values ensures that the Ψ_{ij} fit H_0 . Now, for subsample 1, randomly draw N_1 values with replacement from the N_1 values of Ψ_{1j} ; sampling with replacement implies that some Ψ_{1j} values can appear more than once. Do this for each of the q subsamples. Now we have new samples (still total sample size N) that are based on the original sample data. Calculate Y_r for these randomly resampled azimuths. Repeat the process many times, say $N_R = 1000$ times, to make up a distribution of the test statistic Y_r for H_0 true. Sort these Y_r values and find the cutoff value Y_r^* that corresponds to the largest 100α of the values in the list. All values greater than Y_r^* are in the critical portion of the distribution. Hence, reject H_0 if $Y'_r > Y_r^*$, Fisher (1993).

Test Equality of Concentrations Using Tangential Method

Now test the hypothesis $H_0 : \kappa_1 = \kappa_2 = \dots = \kappa_q$, assuming the data are from the von Mises distribution. Here we use the test statistic f_r described in (Eq. 15.24). Calculate f_r as described above; this will be the test statistic f'_r . Fisher (1993, p. 132) recommends the randomization test.

If H_0 is true, then we could randomly mix all N observed angles among and within the q subsamples and would not expect substantial changes in the calculated f_r . The randomization (also called permutation) test scrambles all observations and calculates f_r . Ideally, all permutations would be used, but this quickly becomes infeasible even for moderate sample sizes. Instead, we use a random subset, say N_R , of the many permutations. For each, we calculate f_r , and the collection estimates the distribution of f_r . As above, we find the test cutoff value f_r^* that defines the critical region. Reject H_0 if $f_r' > f_r^*$, Fisher (1993, pp. 131–132, 214–218).

15.3.1.5 Example Experiment and Analysis

We now consider an example based on the directions of three groups of ants (long-legged desert ants *Cataglyphis fortis*), from an experiment described in Wehner and Müller (1985), to test for interocular transfer, data provided in Table 15.2. The control group (Set 1) consists of 11 long-legged desert ants after one eye on each ant was “trained” to learn the ant’s home direction, then covered and the other eye uncovered. Ants in two treatment groups were being tested for intraocular transfer: Set 2 (32 ants) having their naive eyes occluded while being trained to the 0° compass direction, and Set 3 (18 ants) having their naive eyes occluded for the duration of the experiment, Wehner and Müller (1985).

Table 15.3 gives basic directional summaries of the characteristics of data by group and for the entire experiment.

Figure 15.3 shows the walking directions of three groups of monocular ants that can see only the pattern of polarized light in the sky. We focus our attention on the question of whether there are any differences in the mean directions and spread of the three groups. These questions are considered both separately and with the specialized combined test.

Table 15.4 shows the results of the testing using various techniques, described earlier in the section. Based on the results, we see that all of the tests for the

Table 15.2 Directions (Measured Due North in Degrees) of Desert Ants

Set 1	11	11 3 -22 -1 -7 27 -2 15 14 -13 0
Set 2	32	-46 10 -3 9 19 -5 49 24 14 14 4 -14 -4 30 4 -172 24 -6 -32 -128 -68 -12 8 21 10 -11 -12 25 24 -7 18 -3
Set 3	18	22 32 -25 4 43 108 47 -49 -67 -19 -14 4 -2 140 82 6 -21 19

Table 15.3 Summary of Desert Ants Data

Data Set	n	$\bar{\theta}$ (in radians)	$\hat{p} = \bar{R}$	R	$\hat{\kappa}_{ML}$	$\hat{\kappa}$
Set 1	11	1.531	0.974	10.709	19.175	14.288
Set 2	32	1.566	0.814	26.061	3.049	2.769
Set 3	18	1.372	0.679	12.230	1.886	1.827
All Sets	61	1.510	0.801	48.841	2.870	2.731

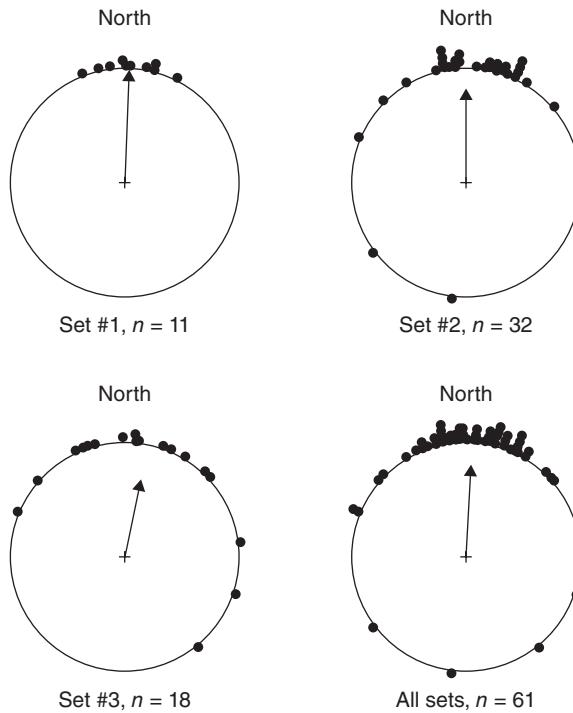


Figure 15.3 Circular plots showing directions (with mean resultant vectors) for three sets of desert ants.

Table 15.4 One-Way ANOVA for Desert Ants Data Using Different Tests

Procedure	Value of Test Statistic	Approximate Distribution	p-Value
<i>Tests of equal mean directions</i>			
Likelihood ratio test	0.511	χ^2_2	0.775
Watson–Williams test	0.437	$F_{2,58}$	0.648
Embedding method	1.075	$F_{2,58}$	0.348
Heterogeneous case	0.646	χ^2_2	0.724
<i>Tests of equality of concentrations</i>			
Medium concentration	25.968	χ^2_2	0
Large concentration	13.590	χ^2_2	0.001
Tangential approach	2.779e + 32	$F_{2,58}$	0
<i>Test equality of mean directions and equality of concentrations</i>			
Variant of likelihood test	1.708	χ^2_4	0.789

equality of the group means lead to the same conclusion at the 1, 5, and 10% levels. We would accept the null hypothesis of no difference between the groups. Examining the plot of individual group resultant lengths shown in Figure 15.3, we see that the mean directions do all appear to point in quite similar directions. Given that the differences in these mean directions are not large relative to the spread of observations within each group, we can see that the different tests have drawn sensible conclusions.

Considering the test for potential differences in the spread, as measured through the von Mises concentration parameter, κ , we see that the null hypothesis is rejected. This result again matches the plots of the data in Figure 15.3, where we see that the concentration of data for the three groups are markedly different, as summarized by the length of the resultant vectors. Finally, the combined test of means and concentrations does not yield a significant result, which indicates that this test, which combines the comparison of mean directions and concentrations, did not see sufficient differences to allow us to reject the joint null hypothesis of no differences in either of these characteristics. As noted earlier, this combined test is somewhat confusing as the relative weights of the two contributions are not clarified.

As illustrated by the multitude of tests for different aspects of circular data and the different stipulations based on potential values of the concentration parameters, evaluating the patterns from a completely randomized design with multiple treatments has more complexity than for the standard linear ANOVA. We now consider the analyses for more complicated designs and the implications on analysis for these situations.

15.3.2 Multiway ANOVA

As Fisher (1993, pp. 133–134) noted, the tools available for the analysis of multifactor experiments with a circular response is relatively limited in the directional data literature. Typically, the goal of the experiment is to consider a number of factors that are thought to have an impact on a circular response, and the goal is to gain a better understanding of the effect and relative importance of the different effects. The methods primarily focus on testing for factor main effects and interactions for the mean direction based on a designed factorial experiments with all combinations of two or more factors, with assumed equal number of replicates per combination. The analysis approaches are also appropriate for cross-classified data from observational studies, although this is not the primary focus of this chapter. Central to the appropriateness of the models and analysis is the assumption that the dispersions across all of the groups are equal, or at least quite similar. While obtaining tests with an appropriately approximated distribution is one key aspect of the problem, another important consideration is to define and interpret what is meant by an interaction term in the circular data context. A limited amount of work has been done to consider differences in the dispersion of the groups formed by the cross-classification.

Underwood and Chapman (1985) developed an extension of the Watson and Williams (1956) ANOVA type decomposition of total variation ($T.V.$) into additive contributions from the main, interaction and residual effects for the two-way scenario. In the following, R_{ijk} is the resultant length for a particular combination of i, j, k , and if a “dot” appears, then we have taken the resultant length over all levels of that index, and \bar{R} is obtained by dividing ΣR_{ijk} by the number of observations that went into the resultant length.

The total variation can be partitioned into row, column, interaction, and residual variation as indicated in Equation (15.27):

$$\begin{aligned} T.V. &= 2\kappa(N - R_{..}) \\ &= R.V. + C.V. + I.V. + Resid.V. \end{aligned} \quad (15.27)$$

The form for the different contributions is as follows:

$$\text{Row: } R.V. = 2\kappa \left(\sum_{i=1}^p R_{i..} - R_{..} \right) \quad (15.28)$$

$$\text{Column: } C.V. = 2\kappa \left(\sum_{j=1}^q R_{.j} - R_{..} \right) \quad (15.29)$$

$$\text{Interaction: } I.V. = 2\kappa \left(\sum_{i=1}^p \sum_{j=1}^q R_{ij.} - \sum_{i=1}^p R_{i..} - \sum_{j=1}^q R_{.j} + R_{..} \right) \quad (15.30)$$

$$\text{Residual: } Resid.V. = 2\kappa \left(N - \sum_{i=1}^p \sum_{j=1}^q R_{ij.} \right) \quad (15.31)$$

We are careful not to call the contributions sums of squares, since although the row, column, and residual terms are all guaranteed to be non-negative, there is no such guarantee for the interaction term. Anderson and Wu (1995) show that assuming a von Mises distribution, if the data are fairly widely spread within a row–column combination (say with $\kappa \leq 4$) then the chance of observing a negative interaction term is not trivial. There are also difficulties with this approach with the interpretation of the interaction term, as it does not appear to have an equivalent interpretation to the linear case.

To test the various effects, an F -statistic of the form

$$\left(1 + \frac{3}{8\hat{\kappa}} \right) \frac{\text{Effect Term / df}_{\text{Effect}}}{\text{Resid.V. / df}_{\text{Residual}}} \sim F_{\text{df}_{\text{Effect}}, \text{df}_{\text{Residual}}}, \quad (15.32)$$

is used, where $\hat{\kappa}$ is the maximum likelihood estimate of the von Mises concentration parameter. The multiplicative coefficient was obtained by Stephens (1969) to improve the distributional properties of the statistic for more disperse samples.

A second approach proposed by Harrison, Kanji, and Gadsden (1986) and supplemented by Harrison and Kanji (1988) suggests an alternative, which resolves the problem of a potential negative interaction term, using

$$SS_{\text{Total}} = SS_{\text{Row}} + SS_{\text{Column}} + SS_{\text{Interaction}} + SS_{\text{Residual}} \quad (15.33)$$

where the definitions for the various terms for a two-way factorial experiment with p rows, q columns and m replicates per cell combination are given as

$$SS_{\text{Total}} = pqm(1 - \bar{R}^2) \quad (15.34)$$

$$SS_{\text{Row}} = qm \left(\sum_i \bar{R}_i^2 - p \bar{R}^2 \right) \quad (15.35)$$

$$SS_{\text{Column}} = pm \left(\sum_j \bar{R}_j^2 - q \bar{R}^2 \right) \quad (15.36)$$

$$SS_{\text{Interaction}} = m \left(\sum_i \sum_j \bar{R}_{ij}^2 - q \sum_i \bar{R}_i^2 - p \sum_j \bar{R}_j^2 + pq \bar{R}^2 \right) \quad (15.37)$$

$$SS_{\text{Residual}} = m \left(pq - \sum_i \sum_j \bar{R}_{ij}^2 \right). \quad (15.38)$$

The cost of guaranteeing a positive interaction term is that the analysis now considers both the mean and dispersion in a combined measure. Hence a zero effect for any of the main or interaction effects is achieved only when there are no differences in either the mean directions or the resultant vector lengths. See Anderson and Wu (1995) for more details. As with the first approach, this method is also quite sensitive to the assumption that the concentrations of the different cells of the crossed design are similar, as well as that the data within each cell are not too spread around the circle. Similar, to the first approach, tests for the different effects use an F -distribution with a correction factor to improve upon the distributional assumptions:

$$\left(1 + \frac{1}{5\hat{\kappa}} + \frac{1}{10\hat{\kappa}^2} \right) \frac{SS_{\text{Effect}} / df_{\text{Effect}}}{SS_{\text{Residual}} / df_{\text{Residual}}} \sim F_{df_{\text{Effect}}, df_{\text{Residual}}}. \quad (15.39)$$

A third alternative was proposed by Anderson and Wu (1995), which is considered a strategy for a 2^k factorial design, where interest lies in differences in only the mean directions, and is based on the likelihood ratio test for data from a von Mises distribution with a common dispersion for all groups. The test for the row effects have the form

$$\chi_{\text{row}}^2 = 2\hat{\kappa} \sum_i R_i (1 - \cos(\theta_i - \theta..)), \quad (15.40)$$

which is asymptotically distributed as χ^2_{p-1} , with an equivalent form for the column effects.

The test statistics for the interaction effect is adapted to take into account the “wrap-around” nature of directional data, to have the form

$$\chi^2_{\text{interaction}} = 2\hat{k} \sum_l R_l(1 - |\cos(\theta_l - \theta \cdot \cdot)|), \quad (15.41)$$

where the subscript l indicates the different levels of the constructed interaction combinations. For example, for the two-way interaction effect, the pseudo-factor would be defined with the θ_{11} and θ_{22} combinations forming the low level of the pseudo-factor, and θ_{12} and θ_{21} for the high level. Then the effect of the interaction could be obtained by looking at the difference in the average effect of the low and high levels of the pseudo effect. For a 2^k factorial experiment with a larger number of factors and interactions, pseudo-factors can be created for all of the two- and higher-order interactions by looking at the product of the factor effects involved in that interaction, to obtain the appropriate designations of the low and high levels. A half-normal plot of all the factor effects based on the original main effects and the created pseudo-factors can be constructed using the square roots of the various effects and comparing them to the square root of the quantiles of a chi-squared distribution with 1 degree of freedom. Effects with large values that fall far from a straight line are influential effects on the response. This approach will enable us to rank the importance of the different effects and gauge their size.

If there is interest in examining how the main and interaction effects influence the dispersion of the response, Anderson and Wu (1996) present some methods for considering this type of problem. For a replicated experiment, the resultant length for each of the factor combinations can be obtained, and an estimate of the concentration parameter κ from the von Mises distribution can be calculated. Since this concentration parameter is measured on a linear scale, these responses can be analyzed using a standard dispersion analysis from an unreplicated factorial experiment once a suitable transformation has been identified. Cordeiro, Paula, and Botter (1994) suggest some alternate strategies for considering dispersion in this context.

15.4 ANOVA FOR CYLINDRICAL DATA

Cylindrical data are comprised of an angular measure, θ , and a linear measurement, y . The name, cylindrical data, arises since we can think of displaying the data on the surface of a cylinder. The angular measure gives a location on the edge of the circular base, while the linear measure defines a height above the base, at that orientation. Figure 15.4 shows a sample observation on the unit cylinder.

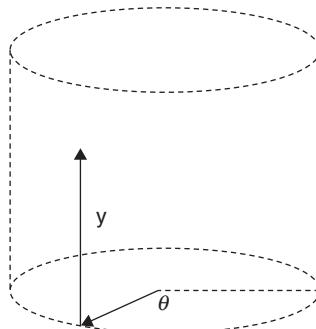


Figure 15.4 Angular and linear measures of cylindrical data.

Alternately, we can consider the data on a two-dimensional surface represented with polar coordinates, with the linear measurement y now defining the distance from the origin. This second representation suggests that we could transform to Cartesian coordinates and do a multivariate ANOVA (MANOVA) with the two responses, X_1 and X_2 . In some applications, the transformation may be natural if it is believed that the input factors are likely to impact latitude and longitude directly. For example, in an experiment considering the migration of plants or animals, it may be quite natural to think of some factors influencing the path of migration parallel and perpendicular to prevailing winds or the path of the sun from east to west. In this case, treating the cylindrical response as bivariate in Cartesian coordinates may be natural and advantageous. If this representation is appropriate, then standard MANOVA techniques are applicable and readily available.

However, in other applications, it is much more natural to think of the input variables having separate effects on direction or on the magnitude of the response. In these cases, transforming to Cartesian coordinates can unnecessarily complicate the analysis by confounding the influence of the inputs on the responses and introducing extraneous interactions and model complexity. We now consider models available for analysis of cylindrical data, as modeled by an angular and linear response.

The basic model for cylindrical data assumes a joint distribution of the angular-linear components with a marginal distribution for the angular measure as a von Mises distribution, and a conditional distribution for the linear measure given the angular measure as Normal distribution (Mardia and Sutton 1978). The density can be written as

$$f(\theta, y) = \{2\pi I_0(\kappa)\}^{-1} \exp\{\kappa \cos(\theta - \mu_0)(2\pi\sigma_c^2)^{-1/2}\} \exp[-\{(y - \mu_c)^2 / 2\sigma_c^2\}], \quad (15.42)$$

where μ_0 , κ are the parameters associated with the marginal distribution of the angle,

$$\mu_c = \mu + \{\rho(\cos \theta - \cos \theta_0)\}, \quad (15.43)$$

and with $\sigma_c^2 = \sigma^2\{1 - (\rho^2)\}$. Hence, the model parameters for the conditional distribution of $y|\theta$ are μ , the center of the linear responses, ρ , the amplitude of the differences between minimum and maximum values of y , and θ_0 , the angle associated with the maximum y value. The conditional distribution portion can be thought of as defining a plane that slices through the unit cylinder to define an elliptical disc. Alternate forms of the model have been suggested, including higher-order models, such as quadratic and cubic (Anderson-Cook and Noble 2001).

Analysis of the angular component of the bivariate response uses methods described in the previous section. The analysis of the linear response can follow a number of different approaches.

For one-way ANOVA, with m groups and N_i replicates per group and $N = \sum N_i$ total observations, we can adapt the basic model to estimate the circular-linear relationship for each group in Cartesian coordinates as

$$E(y_{ij} | \theta_{ij}) = b_{0i} + b_{1i} \cos \theta + b_{2i} \sin \theta, \quad (15.44)$$

or in polar coordinates with

$$E(y_{ij} | \theta_{ij}) = b_{0i} + r(\cos \theta - \theta_0), \quad (15.45)$$

where y_{ij} is the j th observation in the i th group, and r is the amplitude of the cosine function (Anderson-Cook 1999). The first form of the model is a linear model that can be estimated with maximum likelihood or least squares. Similarly, the null model, with no differences between groups, can be estimated with

$$E(y_{ij} | \theta_{ij}) = b_0 + b_1 \cos \theta + b_2 \sin \theta. \quad (15.46)$$

An F -statistic with $3(m - 1), N - 3m$ degrees of freedom can be constructed with

$$F_{\text{obs}} = \frac{\{SS_{\text{residuals}}(\text{null}) - SS_{\text{residuals}}(\text{full})\} / 3(m - 1)}{SS_{\text{residuals}}(\text{full}) / (N - 3m)}. \quad (15.47)$$

If differences are found between the groups, then a variety of strategies can be used to determine global similarities between the groups. The first approach looks for global similarities between the groups to reduce the number of parameters by testing hypotheses for common values for b_0, b_1, b_2 or b_0, r, θ_0 depending on the selected parameterization. The second approach looks to reduce the number of different groups by using the backwards elimination algorithm in multiple linear regression to identify groups that can be combined. See Anderson-Cook (1999) for more details.

Next, we examine the analysis approach for multiway ANOVA from a factorial experiment. Consider the results from a 2^2 factorial design with factors A and B. The linear response can be described by

$$E\{y_{ijk} | \theta_{ijk}\} = \mu^* + A_i^* + B_j^* + AB_{ij}^*, \quad (15.48)$$

where

$$\mu^* = \mu_0 + \mu_1 \cos \theta + \mu_2 \sin \theta, \quad (15.49)$$

represents the null curve, and

$$A_i^* = A_{0i} + A_{1i} \cos \theta + A_{2i} \sin \theta \quad (15.50)$$

is the curve associated with the level i (either high or low) of factor A, and similar expressions for the factor B and interaction AB. See Anderson-Cook (2001) for more details. Because the sum of cosine curves remains a cosine curve, the result for each factor combination is a distinct cosine curve, which allows mapping back to the one-way ANOVA case considered previously.

With this formulation, the main and interaction effects on the conditional response of $y|\theta$ can be estimated as separate cosine curves for each effect. For the high and low levels of the factor, the analysis suggests mirror-image cosine curves centered at zero. The amplitude and angle location of the maximum value for y are estimated from the data. For each term of the model, we have these mirror-image curves, which can then be combined to give cosine curves for each group. Model reduction and testing of hypotheses for main and interaction terms are again possible through backwards stepwise regression and using F -statistics. In this way, particular terms in the model can be removed, and the new estimates for each factor combination are the result of the remaining term curves. This simple representation allows for flexibility in modeling while obtaining an easily interpreted factor effect for all of the main and interaction terms.

15.5 ANOVA FOR SPHERICAL DATA

There are two major categories of data on the sphere. If the response has a single direction, like a radius from the center to a single point on the edge of the sphere, then it is called spherical data. If the response implies a vector across the entire diameter of the sphere with no associate end point, then the data are called axial data. The flight direction of a bird from an elevated perch is an example of spherical data, while the orientation of a fault line in a geological formation would be better characterized as axial data.

Experimental data for spherical and axial data have only occasionally appeared in the scientific literature. Once again, the most typical form of these experiments are completely randomized designs with a single factor being manipulated to form several different levels. Equal or unequal numbers of responses are obtained for each level of the factor with the goal of testing and understanding differences in the groups. The primary emphasis of the tech-

niques in the literature has the focus of evaluating differences in the means. The assumed distribution of the available tests are sensitive to differences in the concentrations (or spreads) within the groups, and corrective adjustments are common to improve the size and power of the tests. For additional methods, see Fisher, Lewis, and Embleton (1987) or Mardia and Jupp (2000). The foundation of ANOVA type techniques for spherical data was laid by Fisher (1953) and Watson and Williams (1956) and enhanced by Stephens (1962a, 1969) and Mardia (1975).

15.5.1 One-Way ANOVA for Spherical Data

Suppose that x_{i1}, \dots, x_{iN_i} , ($i = 1, \dots, q$) are q independent random samples of sizes N_1, \dots, N_q from the spherical von Mises–Fisher distribution $M_q(\mu_i, \kappa_i)$ defined in Section 15.2, for $i = 1, \dots, q$. Let $N = \sum_{i=1}^q N_i$, and the R_i and R denote the resultant length of the i^{th} sample and overall combined sample, respectively. Assuming, $\kappa_1 = \dots = \kappa_q = \kappa$ (unknown concentration), we wish to test $H_0: \mu_1 = \dots = \mu_q$ against the alternative that at least one of the means are different. Under H_0 , the vector of sample mean directions, $\bar{x}_1, \dots, \bar{x}$, are assumed equivalent, so that $R_1 + \dots + R_q \approx R$. Consider the following identity involving, respectively, the variation of the combined sample, the variation within samples, and the variation between samples.

$$2\kappa[n - R] = 2\kappa \left[\sum_i^q R_i - R \right] + 2\kappa \left[N - \sum_i^q R_i \right]. \quad (15.51)$$

15.5.1.1 A High Concentration F-Test

Using the fact that $2\kappa(N - \sum_{j=1}^q R_j)$ and $2\kappa(\sum_{j=1}^q R_j - R)$ are approximately independently distributed as chi-square with $(N - q)(p - 1)$ and $(q - 1)(p - 1)$ degrees of freedom respectively for moderately concentrated data, Watson and Williams (1956) showed that under H_0 , for an average resultant vector length of $\bar{R} \geq 0.67$,

$$\frac{\left(\sum_{i=1}^q R_i - R \right) / (q-1)(p-1)}{\left(N - \sum_{i=1}^q R_i \right) / (N-q)(p-1)} \sim F_{(q-1)(p-1), (N-q)(p-1)}, \quad (15.52)$$

where p is the dimension of the data. For $\hat{\kappa} \geq 1$, an improved approximation by Mardia and Jupp (2000, p. 223) is given by

$$\frac{\hat{\kappa}}{\hat{\gamma}} \frac{\left(\sum_{i=1}^q R_i - R \right) / (q-1)(p-1)}{\left(N - \sum_{i=1}^q R_i \right) / (N-q)(p-1)} \sim F_{(q-1)(p-1), (N-q)(p-1)}, \quad (15.53)$$

where $\hat{\kappa} = A_p^{-1}(\bar{R})$ and $\hat{\gamma}$ is obtained by replacing κ with $\hat{\kappa}$ in $\frac{1}{\gamma} = \frac{1}{\kappa} - \frac{1}{5\kappa^3}$, when $p = 3$ (spherical), and $\frac{1}{\gamma} = \frac{1}{\kappa} - \frac{p-3}{4\kappa^2} - \frac{p-3}{4\kappa^3}$, when $p \neq 3$.

15.5.1.2 The Likelihood Ratio Test

An alternate test involves considering a likelihood ratio test. Let $\theta = \kappa\mu$ be the canonical parameter of the exponential model defined in Section 15.2. Then the log-likelihood based on x_1, \dots, x_n is $l(\theta, x_1, \dots, x_n) = n[\theta^T \bar{x} - a_p(\kappa)]$, where $a_p(\kappa) = \log\left[\left(\frac{\kappa}{2}\right)^{1-p/2} \Gamma\left(\frac{p}{2}\right) I_{p/2-1}(\kappa)\right]$. To test the hypothesis of no differences among group mean directions, calculate

$$\omega = 2 \left\{ \hat{\kappa} \sum_{i=1}^q R_i - \tilde{\kappa} R - n a_p(\hat{\kappa}) + n a_p(\tilde{\kappa}) \right\}, \quad (15.54)$$

where $\hat{\kappa}$ is the maximum likelihood estimates of κ under H_0 , and $\tilde{\kappa}$ is the estimated value of the common concentration parameter, κ , obtained by treating the different groups separately. These estimates are given by $A_p(\hat{\kappa}) = \bar{R}$, and $A_p(\tilde{\kappa}) = \frac{1}{N} \sum_{i=1}^q R_i$, respectively. Under H_0 , for large N , ω has an approximate chi-square distribution with $(q-1)(p-1)$ degrees of freedom. For small values of κ , $a_p(\kappa) \approx \kappa^2/2p$ and $A_p(\kappa) \approx \kappa/p$ gives

$$\omega \approx \frac{p}{n} \left\{ \left[\sum_{i=1}^q R_i \right]^2 - R^2 \right\} = U, \quad (15.55)$$

which has an approximate chi-square distribution with $(q-1)(p-1)$ degrees of freedom for large N . As in the circular case, this approximation is improved using a multiplicative corrective adjustment, c . Thus $cU \sim \chi^2_{(q-1)(p-1)}$, where

$$\frac{1}{c} = 1 - \frac{\kappa^2}{p(p+2)} + \frac{p(p-1)q}{4N\kappa^2}.$$

For $p = 3$, the approximation works well even for small values of N , provided κ (in practice is replaced by its MLE) is not near 0 or 1.

Other extensions of ANOVA procedures to spherical from circular data are discussed by Mardia and Jupp (2000, pp. 225–232).

15.5.2 One-Way ANOVA for Axial Data

Let $x_{i1}, x_{i2}, \dots, x_{iN_i}$ be a sample of size N_i from the population from a spherical Watson distribution (defined in Section 15.2) $W(\mathbf{u}_i, \xi_i)$, for $i = 1, \dots, k$, with

the k samples being independent. Let $N = \sum_{i=1}^k N_i$ be the total number of observations, \mathbf{u} be the directional parameter, and $\xi = \xi_1 = \xi_2 = \dots = \xi_k$ be the unknown concentration parameters, but assumed equal across the groups. Consider the identity,

$$(\mathbf{u}'\mathbf{x}_{ij})^2 = \{(\mathbf{u}'\mathbf{x}_{ij})^2 - (\mathbf{u}'\mathbf{x}_{ij})^2\} + (\mathbf{u}'\mathbf{x}_{ij})^2, \quad (15.56)$$

which can be adapted to

$$2\xi\{1 - (\mathbf{u}'\mathbf{x}_{ij})^2\} = 2\xi\{1 - (\mathbf{u}'\mathbf{x}_{ij})^2\} + 2\xi\{(\mathbf{u}'\mathbf{x}_{ij})^2 - (\mathbf{u}'\mathbf{x}_{ij})^2\}, \quad (15.57)$$

where \mathbf{u} and \mathbf{u}_i are both unknown and are estimated by maximum likelihood. Note $\hat{\mathbf{u}}_i$ is the eigenvector associated with the largest eigenvalue w_i of the orientation matrix $\mathbf{T}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij}\mathbf{x}_{ij}'$, and $\hat{\mathbf{u}}$ is the eigenvector associated with the largest eigenvalue w of $\mathbf{T} = \sum_{i=1}^k T_i$.

Next, summing the identity and replacing unknown parameters with their estimates, we obtain

$$2\xi \sum_{i=1}^k \sum_{j=1}^{n_i} \{1 - [\mathbf{u}'\mathbf{x}_{ij}]^2\} = 2\xi \sum_{i=1}^k \sum_{j=1}^{n_i} \{1 - [\mathbf{u}'\mathbf{x}_{ij}]^2\} + 2\xi \sum_{i=1}^k \sum_{j=1}^{n_i} \{[\mathbf{u}'\mathbf{x}_{ij}]^2 - [\mathbf{u}'\mathbf{x}_{ij}]^2\}, \quad (15.58)$$

which can be written as

$$2\xi[n-w] = 2\xi \left[n - \sum_{i=1}^k w_i \right] + 2\xi \left[\sum_{i=1}^k w_i - w \right]. \quad (15.59)$$

Clearly, $2\xi(n-w) \geq 0$, since $2\xi(\sum_{i=1}^k w_i - w) \geq 0$ as the largest eigenvalue of $\mathbf{T} = \sum_{i=1}^k \mathbf{T}_i$ is not greater than the sums of the largest eigenvalue of \mathbf{T}_i , and $2\xi(n - \sum_{i=1}^k w_i)$ is greater or equal to zero, since it can be expressed as $2\xi \sum_{i=1}^k \sum_{j=1}^{n_i} (1 - \cos^2 \theta_{ij})$, where θ_{ij} is the angle between $\hat{\mathbf{u}}_i$ and \mathbf{x}_{ij} .

One can show that the approximate distributions under H_0 for large ξ are as follows:

$$2\xi(n_i - w_i) \sim \chi^2_{(n_i-1)(p-1)}, \text{ and thus } 2\xi \sum_1^k (n_i - w_i) \sim \chi^2_{(n-k)(p-1)}.$$

Hence, $2\xi(n-w) \sim \chi^2_{(n-1)(p-1)}$. One can show that $2\xi(n - \sum_1^k w_i)$ and $2\xi(\sum_1^k w_i - w)$ are approximately independent by combining Cochran's theorem with the identity above, thus it follows that for large ξ , $2\xi(\sum_1^k w_i - w) \sim \chi^2_{(k-1)(p-1)}$. The null hypothesis of no difference between the mean directions of the groups is rejected for large values of the statistic

Table 15.5 One-Way ANOVA

Source of Variation	Value	Degrees of Freedom
Between groups	$2\xi(\sum_{i=1}^k w_i - w)$	$(k-1)(p-1)$
Within groups	$2\xi(n - \sum_{i=1}^k w_i)$	$(n-k)(p-1)$
Total	$2\xi(n - w)$	$(n-1)(p-1)$

$$F = \frac{(n-k)\left(\sum_{i=1}^k w_i - w\right)}{(k-1)\left(n - \sum_{i=1}^k w_i\right)}, \quad (15.60)$$

which has an approximate F distribution with $(k-1)(p-1)$ and $(n-k)(p-1)$ degrees of freedom.

Table 15.5 above is a usual tabular scheme for one-way ANOVA, with the calculations for the “between-groups” sum of squares and “within-groups” sum of squares.

15.5.3 Multiway ANOVA for Axial Data

Two-way and multiway ANOVA for spherical data techniques arise from generalizing the methods for the circular case suggested by Harrison, Kanji, and Gadsden (1986), and Harrison and Kanji (1988). Figueiredo (2006) extends the one-way ANOVA based on the bipolar (axial) Watson distribution to a two-way layout.

Suppose that we have n observations from a twofold nested classification (see HK1, section 4.12), with classification 1 having r groups, indexed by $i = 1, \dots, r$, and classification 2 having s_i groups within group i of classification 1, indexed by $j = 1, \dots, s_i$. When an observation is classified into group i of classification 1 and group j of classification 2, this observation falls in cell (i, j) in row i and column j of a two-way table. Let n_{ij} be the number of observations in cell (i, j) , and let w_{ij} be the largest eigenvalue of the orientation matrix associated with these observations. Let $n = \sum_{i=1}^r n_i$ be the total number of observations, where $n_i = \sum_{j=1}^{s_i} n_{ij}$ and $n^* = \sum_{i=1}^r s_i$, where n^* is the total number of categories based on both of the classifications. Let w_i be the largest eigenvalue of the orientation matrix associated with the observations in row i and suppose that w is the largest eigenvalue of the orientation matrix associated with all observations.

Consider the following identity:

$$\begin{aligned} 2\xi(n-w) &= 2\xi \sum_{j=1}^{s_1} (n_{1j} - w_{1j}) + 2\xi \sum_{j=1}^{s_2} (n_{2j} - w_{2j}) + \dots + 2\xi \sum_{j=1}^{s_r} (n_{rj} - w_{rj}) \\ &\quad + 2\xi \left(\sum_{j=1}^{s_1} w_{1j} - w_{1.} \right) + \dots + 2\xi \left(\sum_{j=1}^{s_r} w_{rj} - w_{r.} \right) + 2\xi \left(\sum_{i=1}^r w_{i.} - w \right), \end{aligned} \quad (15.61)$$

Collecting terms, we have

$$\begin{aligned} 2\xi(n-w) &= 2\xi \left[n - \sum_{i=1}^r \sum_{j=1}^{s_i} w_{ij} \right] + 2\xi \left[\sum_{j=1}^{s_1} w_{1j} - w_{1.} \right] + \dots \\ &\quad + 2\xi \left[\sum_{j=1}^{s_r} w_{rj} - w_{r.} \right] + 2\xi \left[\sum_{i=1}^r w_{i.} - w \right], \end{aligned} \quad (15.62)$$

with corresponding assumed distributions for large ξ

$$\chi^2_{(n-1)(p-1)} = \chi^2_{(n-n^*)(p-1)} + \chi^2_{(s_1-1)(p-1)} + \dots + \chi^2_{(s_r-1)(p-1)} + \chi^2_{(r-1)(p-1)}. \quad (15.63)$$

where p is the dimension of the data.

The terms may be arranged in a variance component table as in Table 15.5. The null hypothesis is that there are no differences among the mean direction of the groups, and this is rejected for large values of the following statistic

$$F_1 = \frac{(n-n^*) \left(\sum_{i=1}^r w_{i.} - w \right)}{(r-1) \left(n - \sum_{i=1}^r \sum_{j=1}^{s_i} w_{ij} \right)}, \quad (15.64)$$

which under the null hypothesis and for large ξ is approximately distributed as $F_{(r-1)(p-1), (n-n^*)(p-1)}$.

Similarly for testing for no differences among columns within row i , the hypothesis is rejected for large values of the following statistic

$$F_2 = \frac{(n-n^*) \left(\sum_{j=1}^{s_i} w_{ij} - w_{i.} \right)}{(s_i-1) \left(n - \sum_{i=1}^r \sum_{j=1}^{s_i} w_{ij} \right)}, \quad (15.65)$$

which under the null hypothesis and for large ξ is approximately distributed as $F_{(s_i-1)(p-1), (n-n^*)(p-1)}$. Note that a rough check of the assumption of homoscedasticity (that all the Watson distributions involved have the same concentration ξ) can be derived from the fact that for large ξ , $2\xi[1-(\mathbf{u}'\mathbf{x})^2] \sim \chi^2_{p-1}$, where $\mathbf{x} \in S_{p-1}$ comes from the bipolar Watson population $W(\mathbf{u}, \xi)$. Thus, in the twofold nested case, we consider the values of $(1-w_{ik}/n_{ik})/(1-w_{jl}/n_{jl})$ and compare it with $F_{(n_{ik}-1)(p-1), (n_{jl}-1)(p-1), \alpha}$. The above analysis can be set up in the usual tabular scheme for twofold nested ANOVA, with the “between-rows” sum of squares and “within-rows” sum of squares as shown in Table 15.6.

An example of data from a design of experiment is the vectorcardiogram data of Downs, Liebman, and Mackay (1971), where the researchers were interested in investigating whether there is no interaction between the two factors: type of system (Frank system or McFee system) and the demographic

Table 15.6 Twofold Nested ANOVA.

Source of Variation	Value	Degrees of Freedom
Between rows	$2\xi(\sum_{i=1}^r w_{i\cdot} - w)$	$(r-1)(p-1)$
Between cols. within row 1	$2\xi(\sum_{j=1}^{s_1} w_{1j} - w_{1\cdot})$	$(s_1-1)(p-1)$
⋮		
Between cols. within row r	$2\xi(\sum_{j=1}^{s_r} w_{rj} - w_{r\cdot})$	$(s_r-1)(p-1)$
Within cells	$2\xi(n - \sum_{i=1}^r \sum_{j=1}^{s_i} w_{ij})$	$(n-n^*)(p-1)$
Total	$2\xi(n - w)$	$(n-1)(p-1)$

four-level factor “age-sex” (boys aged 2–10, boys aged 11–19, girls aged 2–10 and girls aged 11–19), and whether the type of system or the “age-sex” affect the result of the vectorcardiogram. Each vectorcardiogram has three unit vectors associated with it. A two-way ANOVA is illustrated by using the unit spherical vector, which represents the spatial direction of the vector of the QRS loop having the greatest magnitude; see Figueiredo (2008).

15.6 CONCLUSIONS

Designed experiments involving one or more factors thought to influence the directional responses have stemmed from an analogous strategy to ANOVA used for linear data. Typically, designed experiments for directional data are relatively simple in structure with a small number of factors, which are either categorical or on a linear scale. The form of these experiments are generally completely randomized designs with allocations to the groups in equal or unequal number of replications. Factorial (or occasionally fractional factorial) designs are most common for multiway experiments. While this simple structure for experiments may appear restrictive, relative to the options available for traditional experiments involving a linear response, the nature of many of the applications for directional data have been primarily observational. Hence, there has been relatively little demand for more complicated designs. The primary types of directional data (circular, cylindrical, or spherical) have associated techniques for ANOVA and a number of more specialized methods.

Multisample tests have been proposed in the literature for testing the hypothesis of the equality of the directional parameters. The von Mises and the von Mises–Fisher distributions are commonly used as models for circular and spherical data problems, respectively. The methods discussed in this chapter are taken from the literature, primarily Fisher (1993) and Mardia and Jupp (2000), who refer extensively to prior work. Readers seeking additional background or methods should consider these references for added information.

REFERENCES

- Anderson, C.M. and C.F.J. Wu (1995). Measuring location effects from factorial experiments with a directional response. *International Statistical Review*, **63**, 345–363.
- Anderson, C.M. and C.F.J. Wu (1996). Dispersion measures and analysis for factorial directional data with replicates. *Journal of the Royal Statistical Society. Series C*, **45**, 47–61.
- Anderson-Cook, C.M. (1999). A tutorial on one-way analysis of circular-linear data. *Journal of Quality Technology*, **31**, 109–119.
- Anderson-Cook, C.M. (2000). An industrial example using one-way analysis of circular-linear data. *Computational Statistics and Data Analysis*, **33**, 45–57.
- Anderson-Cook, C.M. (2001). Understanding the influence of several factors on a cylindrical response. *Journal of Quality Technology*, **33**, 167–180.
- Anderson-Cook, C.M. and R.B. Noble (2001). An alternate model for cylindrical data. *Nonlinear Analysis*, **47**, 2011–2022.
- Batschelet, E. (1981). *Circular Statistics in Biology*. London: Academic Press.
- Cordeiro, G., G. Paula, and D. Botter (1994). Improved likelihood ratio tests for dispersion models. *International Statistical Review*, **62**, 257–274.
- Downs, T., J. Liebman, and W. Mackay (1971). Statistical methods for vectorcardiogram orientations. In: *Vectorcardiography 2: Proceedings of XIth International Symposium on Vectorcardiography*, R.I. Hoffman and E. Glassman (eds.). Amsterdam, The Netherlands: North-Holland, pp. 216–222. Data available at <<http://www.mcs.st-andrews.ac.uk/pej/vcg.html>>.
- Figueiredo, A. (2006). Two-way analysis of variance for data from a concentrated bipolar Watson distribution. *Journal of Applied Statistics*, **33**, 575–581.
- Figueiredo, A. (2008). Two-way ANOVA for the Watson distribution defined on the hypersphere. *Statistical Papers*, **49**, 363–376.
- Fisher, R.A. (1953). Dispersion on a sphere. *Proceedings of the Royal Society (London) Series A*, **347**, 371–385.
- Fisher, N.I. (1986). Robust comparison of dispersion for samples of direction data. *The Australian Journal of Statistics*, **28**, 213–219.
- Fisher, N.I. (1993). *Statistical Analysis of Circular Data*. Cambridge: Cambridge University Press.
- Fisher, N.I., T. Lewis, and B.J.J. Embleton (1987). *Statistical Analysis of Spherical Data*. Cambridge: Cambridge University Press.
- Graves, T.S. (1979). Randomization procedures for a two way analysis of orientation data. Ph.D. Thesis, Cornell University, Ithaca, NY.
- Harrison, D. and G.K. Kanji (1988). The development of analysis of variance for circular data. *Journal of Applied Statistics*, **15**, 197–224.
- Harrison, D., G.K. Kanji, and R.J. Gadsden (1986). Analysis of variance for circular data. *Journal of Applied Statistics*, **13**, 123–138.
- Jammalamadaka, S.R. and A. SenGupta (2001). *Topics in Circular Statistics*. Singapore: World Scientific Publishing Co.

- Karlin, S. and W.J. Studden (1966). Optimal experimental designs. *The Annals of Mathematical Statistics*, **37**, 783–815.
- Laycock, P.J. (1975). Optimal design: Regression models for directions. *Biometrika*, **62**, 305–311.
- Mardia, K.V. (1972). *Statistics of Directional Data*. London: Academic Press.
- Mardia, K.V. (1975). Statistics of Directional Data (with discussion). *Journal of the Royal Statistical Society. Series B*, **37**, 349–393.
- Mardia, K.V. and P.E. Jupp (2000). *Directional Statistics*. Chichester, UK: John Wiley.
- Mardia, K.V. and B.D. Spurr (1973). Multisample tests for multimodal and axial circular populations. *Journal of the Royal Statistical Society. Series B*, **35**, 422–436.
- Mardia, K.V. and T.W. Sutton (1978). A model with cylindrical variables and applications. *Journal of the Royal Statistical Society. Series B*, **40**, 229–233.
- Rao, J.S. and S. Sengupta (1970). An optimum hierarchical sampling procedure for crossbedding data. *Journal of Geology*, **78**, 533–544.
- Stephens, M.A. (1962a). The statistics of directions. PhD thesis. University of Toronto.
- Stephens, M.A. (1962b). Exact and approximate tests for directions I. *Biometrika*, **49**, 463–477.
- Stephens, M.A. (1962c). Exact and approximate tests for directions II. *Biometrika*, **49**, 547–552.
- Stephens, M.A. (1962d). The extension of the von Mises and Fisher vector distributions to higher dimensions Technical Report. Department of Statistics, Johns Hopkins University.
- Stephens, M.A. (1969). Tests for the von Mises distribution. *Biometrika*, **56**, 149–160.
- Stephens, M.A. (1972). Multi-sample tests for the von Mises distribution. *Journal of the American Statistical Association*, **67**, 456–461.
- Stephens, M.A. (1982). Use of the von-Mises distribution to analyze continuous proportions. *Biometrika*, **69**, 197–203.
- Underwood, A.J. and M.G. Chapman (1985). Multifactorial analysis of directions of movement of animals. *Journal of Experimental Marine Biology and Ecology*, **91**, 17–43.
- Upton, G.J.G. (1974). New approximations to the distribution of certain angular statistics. *Biometrika*, **61**, 369–373.
- Von Mises, R. (1918). Über die Ganzzahligkeit der Atomgewichte und verwandte Fragen. *Physikalische Zeitung*, **19**, 490–500.
- Watson, G.S. (1983). *Statistics on Spheres*. New York: John Wiley.
- Watson, G.S. and E.J. Williams (1956). On the construction of significance tests on the circle and sphere. *Biometrika*, **43**, 344–352.
- Wehner, R. and M. Müller (1985). Does interocular transfer occur in visual navigation by ants? *Nature*, **315**, 228–229.
- Wu, H. (1997). Optimal exact designs on a circle or a circular arc. *The Annals of Statistics*, **25**, 2027–2043.

Name Index

Note: Page numbers in **bold** indicate an article or book reference.

- Aaslyung, M., 369, **375**
Abdelbasit, K. M., 149, **162**
Agarwal, S. C., 32, 41, **66**
Aggarwal, M. L., 455, **467**
Agin, M., 149, **162**
Agresti, A., 138, **162**, 370, **374**
Aguiar, A., 116, **133**
Ahner, D. K., 423, 430, **438**
Ai, M. Y., 310, **329**
Akaike, H., 287, **297**
Allison, D. B., 80, **107**
Altman, N. S., 85, **106**
Amaya-Amaya, M., 334, **377**
Anderson, C. M., 508, 519–521, **531**
Anderson, D., 340, **374**
Anderson-Cook, C. M., 451, 456, 461,
 467–469, 472, **499**, 508, 510, 523–524,
 531
Ankenman, B., 421, **441**
Anscombe, F. J., 223, **246**
Antonellis, K. J., 77–79, **107**
Armitage, P., 220, **246**
Arpat, A. B., 74, **107**
Arrow, K. J., 223, **246**
Arunachalan, V., 13, **66**
Arya, A. S., 27, 30, **66**, **69**
Asilahijani, H., 448–449, **467**
Atkinson, A. C., 275, **279**, 285, **297**, 338,
 374, 463, **467**
Attie, A. D., 85, **107**
Bai, Z., 259, **279**
Baksalary, J. K., 59, **66**
Baezinger, P. S., 114–115, **135**
Baker, R. J., 114, **133**
Ball, S. T., 115, **132**
Banerjee, T., 92, **106**
Barnard, G. A., 219, **247**
Barrett, T., 73, **107**
Barrios, E., 477, **499**
Bartky, W., 220, **247**
Bartlett, M. S., 113, **132**, 450, **467**
Basford, K. E., 116, **134**
Bates, R. L., 327, **328**
Batschelet, E., 501, **531**
Bauer, P., 241, 245–246, **247–248**
Bauman, M. F., 424, 437, **438**
Beaver, R., 360, **375**
Beazer-Barclay, Y. D., 77–79, **107**
Beckman, R., 400–401, **410**
Behnken, D. W., 472, 484, **499**
Belderrain, M. C. N., 419, **441**
Berger, J. O., 223, **247**
Berry, D. A., 223, **247**

Design and Analysis of Experiments: Special Designs and Applications, First Edition. Edited by Klaus Hinkelmann.

© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

- Besag, J., 114, **132**
 Bhatti, A. U., 115, **132**
 Biedermann, S., 141, 146, 149, **162**
 Billingsley, P., 218, **247**
 Bingham, D., 462, **467**
 Bingham, D. R., 327, **327**
 Birch, J. B., 451, **468**
 Bisgaard, S., 327, **329**, 477, 482–483, **499**
 Biswas, A., 261, **279**
 Biswas, P. C., 60, **67**
 Blackburn, P. R., 232, **249**
 Blackwell, D., 223, **246**
 Bliemer, M. C., 368, **374**
 Block, R. M., 311, 317, **327**
 Boer, M., 3, **68**
 Bohn, L. L., 451, **469**
 Booker, A., 391, **409**
 Borkowski, J. J., 453, 461, **467**
 Borror, C. M., 447, 455, **467–468**
 Bose, R. C., 110, **133**, 283–284, **297**
 Botstein, D., 3, **68**
 Botter, D., 521, **531**
 Boulton, M., 355, **376**
 Boutros, M., 79, **107**
 Bowman, D. T., 114–115, **133**
 Box, G. E. P., 117, **133**, 301–303, **327**,
 421–422, **438**, 449, 453, **467**, 472, 484,
 497, **499**
 Brannath, W., 246, **248**
 Bredie, W., 369, **375**
 Brenneman, W. A., 450, 465–466, **467–469**
 Bretz, F., 81, 93–94, **107**
 Brewster, J. F., 327, **328**
 Bristow, J. D., 213, 230, **247**
 Brittain, E., 237–238, **249**
 Bross, I., 220, **247**
 Brownman, K. W., 3, **66**
 Brown, B. J., 205, **210**
 Brown, P. O., 73, 75, **108**
 Brownie, C., 114–115, **135**
 Bruen, A., 312, **327**
 Brunner, E., 81, 93–94, **107**
 Bunch, D. S., 340, **374**
 Burakoff, S. J., 223, **247**
 Burgess, L., 331, 335, 338–341, 343, 345,
 347, 349, 351, 355–363, **374**, **377**
 Burns, C., 291, **297**
 Burton, J. W., 114–115, **133**
 Bush, S., 358–360, 362, **374**
 Buss, A. H., 423, 430, **438**
 Butler, D. G., 122, **133**
 Butler, N. A., 316–317, 321, **327**
 Byington, R. P., 209, **210**
 Cafeo, J., 405, **410**
 Callow, M. J., 81, **107**
 Carlson, F., 370, **374**
 Carlyle, M., 419, **439**
 Carriquiri, A., 85, **108**
 Carroll, R. J., 74, **108**
 CAST Investigators, 213, **247**
 CAST II Investigators, 214, 229, **247**
 Castle, R., 334, **376**
 Cavendish, J., 405, **410**
 Ceccarelli, S., 116, **135**
 Chai, F. S., 56, **66**
 Chai, G. X., 262, **281**
 Chakraborty, G., 333, **375**
 Chaloner, K., 149, **162–163**, 368, **375**
 Chan, B. S. P., 111, **133**
 Chan, W. S., 259, **281**
 Chapman, M. G., 519, **532**
 Chatterjee, K., 56, **66**, 293, **297**
 Chaudhary, B. D., 2, 65, **70**
 Chen, H., 301, 303, 309, 312, 315–317,
 322–324, **327–328**
 Chen, J., 307–308, 323, **327–328**
 Chen, Y., 307, **329**
 Cheng, C.-S., 111, **133**, 287, **297**, 305–307,
 312, 314–317, 321–327, **327–329**
 Cheng, S. W., 321, **328**
 Cheung, S. H., 259, **281**
 Chipman, H., 450, **467**
 Choe, S. E., 79, **107**
 Choi, K. C., 43–44, 50–51, **66**
 Cioppa, T. M., 416, 418–419, 421, 425,
 438–439
 Clarke, F. R., 114, **133**
 Christie, B. R., 12, 21, **69**
 Church, G. M., 79, **107**
 Churchill, G. A., 3, **70**, 82, 85, 87–88, **107**
 Clatworthy, W. H., 28, 34, 38, 48, 51–52,
 66
 Coad, D. S., 267, **279**
 Coast, J., 369, **375**
 Cockerham, C. C., 8–9, 61–64, **66**, **70**
 Collin, F., 77–79, **107**
 Collins, R., 209, **211**, 226, **249**

- Comstock, R. E., 9, 12, **66**
Conover, J. 400–401, **410**
Conover, W. J., 382, **410**
Coombs, N. E., 110–111, 127, **133**
Cooper, M., 116, **134**
Cope, L. M., 79, **107**
Corcos, A., 1, **69**
Cordiero, G., 521, **531**
Cornell, J. A., 472, **499**
Cornfield, J., 223, **247**, 278, **279**
Cox, D. R., 273, **279**
Creighton, L., 430, **440**
Cressie, N. A. C., 119, **133**
Critchlow, D. E., 370, **375**
Crossa, J., 114, 118, **134**
Cui, L., 245, **247**
Cukier, R. I., 385, **410**
Cullis, B. R., 111, 114–119, 122, 125, 127,
 133–135
Curnow, R. N., 27–30, 33, 39, 57, **66**, **68**
Currie, I. D., 114, **133**
Cutler, S. J., 187, **210**
- Das, A. 32, 46, 48–50, 54–59, **66**, **68**
Das, D., 464, **467**
Das, M. N. 32, 41, 63, **66**, **69**
Davidson, R. R., 358, 360, **375**
Davis, R. W., 73, 75, **108**
Davydov, A. A., 312, **328**
Dean, A. M., 48–49, 54–56, **66**
Dean, T., 355, **377**
Deeg, H. J., 223, **247**
Defense Industry Daily, 423, **438**
Dejaeger, S., 369, **375**
Dekkers, J. C. M., 85, **108**
Delacy, I. H., 116, **134**
Del Castillo, E., 460, **468**
Deleger, S., 278, **280**
DeMets, D. L., 168, 206, **210**, 213, 230,
 232, 234–235, **247**
Deng, H., 294, **297**
Deng, L.-Y., 287, **297**, 320–321, **328**
Derringer, G., 460, **467**
Desai, N. R., 59–61, **67**
Detre, K., 178, **211**
Dette, H., 141, 146, 149, **162–163**
Dey, A., 47–49, 51, 56, 59–60, **66–67**, **70**,
 335, 347, **375**
Dhall, S. P., 59, **67**
- Diggle, P. J., 120, **133**
Divecha, J., 41, **67**
Diyagama, D., 77, **108**
Dobbin, K., 84–85, **107**
Dodge, H. F., 220, **247**
Doerge, R. W., 3, **67**
Domrachev, M., 73, **107**
Donev, A. N., 285, **297**, 338, **374**, 463,
 467
Donovan, D. M., 365, **377**
Downs, T., 529, **531**
Dragalin, V., 252, **279**
Draper, N. R., 472, 482, **499**
Drignei, D., 404, **410**
Dros, H. A., 155, **163**
Dudoit, S., 78, 81, **107**
Dunnett, C. W., 220, **247**
Dunwell, J. M., 3, **67**
Durban, M., 114, **133**
Durbán Reguera, M. L., 114, **133**
Durham, S., 258–259, **281**
Dutilleul, P., 113, **136**
Dutkowski, G. W., 116, **133**, **135**
- Ebbutt, A., 275, **279**
Eccleston, J. A., 51, 54, **67**, **69**, 111, **133**,
 155, **163–164**
Eckert, S. R., 113–114, **135**
Ederer, F., 187, **210**
Edgar, R., 73, **107**
Edwards, J., 80, **107**
Eeuwijk, F., 3, **68**
Efron, B., 254, 263, **279**
Eisele, J., 262, **279**
Elfving, G., 140, 145, **163**
Ellenberg, S. S., 178, 206, **210**
Embleton, B. J. J., 502, 507, 525, **531**
Engel, J., 445, **467**
Engler, J., 275, **279**
Ennis, D. M., 369, **375**
Erskine, W., 116, **135**
Ettenson, R., 333, **375**
Evangelista, C., 73, **107**
- Falconer, D. S., 2, **67**
Fan, J., 391, **410**
Fang, K. T., 321, **328**, 402, **410**
Fanta, S., 51, **70**
Farrar, S., 355, **377**

- Featheringham, T. R., 415, **439**
 Feder, M., 323, **328**
 Federer, W. T., 30, **67**, 111, 114, **133**
 Fedorov, V. V., 285, **297**, 418, **438**
 Feickert, A., 423, **439**
 Ferrara, J. L. M., 223, **247**
 Ferrini, S., 332, 368, 370, **375–376**
 Fiebig, D. G., 334, 351, 370, **375–376**
 Figueiredo, A., 507, 528, 530, **531**
 Finn, A., 369, **375**
 Fisch, C., 213, 230, **247**
 Fisher, L. D., 241, **247**
 Fisher, N. I., 503, 505, 507, 511–512,
 515–516, 518, 525, 530, **531**
 Fisher, R. A., 245, **247**, 252, **279**, 364, **375**,
 506, 525, **531**
 Fleming, T. R., 206, 210, **210**, 221–223,
 225, **248**
 Fligner, M. A., 370, **375**
 Flyer, P. A., 276, **279**
 Flynn, T. N., 369, **375–376**
 Follman, D. A., 235, **248**
 Forbes, B. E., 141, **163**
 Ford, I., 146, 149, 151, **163**
 Forster, B. P., 3, **67**
 Fortuin, C. M., 385, **410**
 Foulkes, M., 205, **210**
 Fox, P. N., 114, 118, **134**
 Francis, I., 369, **377**
 Freedman, B., 278, **279**
 Freedman, L. S., 204, **210**, 232, **249**
 Freeman, G. H., 111, **134**
 Freeman, H. A., 220, **247**
 Freese, J., 370, **376**
 Friedman, L. M., 168, **210**, 213, 230, **247**
 Friedman, M., 220, **247**
 Fries, A., 301, 309, **328**
 Fritjers, J. E. R., 369, **375**
 Fuller, H., 477, **499**
 Furberg, C. D., 168, 209, **210**
 Furrer, R., 392, **410**
 Fyfe, J. L., 27–28, **67**

 Gadbury, G. L., 80, **107**
 Gadsden, R. J., 511, 520, 528, **531**
 Gaeth, G., 333, **375**
 Gail, M. H., 274, **279**
 Garrett, K. A., 106, **108**
 Gehan, E. A., 220, **248**

 Geller, N. L., 234, **248**
 Genton, M. G., 392, **410**
 Geraldes, M., 263, **280**
 Gerard, K., 334, 340, **377**
 Ghosh, D. K., 41, 59–61, **67**
 Ghosh, H., 57–58, **67**
 Ghosh, M., 138, 141, 162, **163**
 Ghosh, S., 59, **67**, 289–291, 293–294,
 297–298
 Gilbert, N., 27–28, **67**
 Gilmour, A. R., 114–117, 119, 122,
 133–135
 Gilmour, S. G., 320–321, **329**, 482, **500**
 Girshick, M. A., 223, **246**
 Gleeson, A. C., 114–115, 117–118,
 133–134
 Glonek, G. F. V., 89, 91–92, **107**
 Gogel, B. J., 122, **133**
 Goldsman, D., 416, **439**
 Goos, P., 339, 368–370, **376**, **378**, 496, **499**
 Gould, S. J., 178, **210**
 Govindarajulu, Z., 267, **279**
 Gramacy, R. B., 393, **410**
 Grando, S., 116, **135**
 Grasshoff, U., 346–347, 365–366, 370, **375**
 Graves, T. S., 508, **531**
 Green, P., 114, **134**, 364, 367, **375**
 Green, P. J., 393, **410**
 Greenhouse, S. W., 278, **279**
 Greenwood, M., 189, **210**
 Grewank, A., 405, **410**
 Griffing, B., 16, 21–22, 24–25, 27, 30, 33,
 65, **67**
 Grondona, M. O., 114, 118, **134**
 Gronwald, W., 461, **468**
 Grossmann, H., 346–347, 365–366, **375**
 Gupta, B. C., 293, **298**
 Gupta, S., 32, 42, 44–46, 50–51, 54–57,
 66–68
 Gupta, V. K., 49–51, 55–56, 59–60, **69–70**
 Gurmani, A. H., 115, **132**
 Guttmann, R., 334, **376**

 Haddard, L., 312, **327**
 Haenszel, W., 193, **211**
 Haines, L. M., 146, 149, 155, **163**
 Halfon, M. S., 79, **107**
 Hall, J., 351, 355, **376**
 Hallstrom, A., 213, 230, **247**

- Halperin, M., 278, **279**
 Halpern, J., 205, **210**
 Hamada, M., 461–462, **468–469**
 Hammersley, J. M., 415, **439**
 Hancock, T. W., 113–114, **135**
 Handcock, M. S., 393, **410**
 Handoo, M. I., 63, **69**
 Handscomb, D. C., 415, **439**
 Hanley, N., 334, **376**
 Harbaugh, A. W., 393, 395, **410**
 Harper, C., 261–262, 267, **280**
 Harrison, D., 508, 511, 520, 528, **531**
 Harville, D. A., 114–115, 117, **136**
 Haynes, F. L., 114, **134**
 Hayman, B. I., 1, 12, 14–15, 17, 25, 65, **68**
 He, X., 263, 269, **280**
 Heagerty, P. J., 120, **133**
 Hedayat, A., 54, 59, **67–68**, 334, 347, **376**
 Hedayat, A. S., 301, 309, 315–317, **327**,
 421–422, **439**
 Heebner, D., 393, **410**
 Helland, I. S., 226, **247**
 Hensher, D., 340, 363, **376**
 Hernandez, A. S., 419, **439**
 Herson, J., 206, **210**
 Hess, S., 368, **374**
 Higdon, D., 405, **410**
 Hinch, J., 415, **440**
 Hinkelmann, K., xviii, **xviii**, 1, 6–7, 9, 13,
 27–28, 30–31, 33, 37–38, 44–45, 56,
 62–63, 65, **68**, **70**, 417–418, **439**
 Hirtz, D. G., 178, **211**
 Hobbs, B., 77–79, **107**
 Hoeber, F. P., 417, **439**
 Hoffmann, D., 461, **468**
 Hohm, T., 461, **468**
 Holford, T., 178, **211**
 Hollick, L. J., 421, **440**
 Holling, H., 346–347, 365–366, **375**
 Holloway, A., 77, **108**
 Holmes, M., 77, **108**
 Hossain, I., 351, **376**
 Howes, C. W., 113, **134**
 Hu, F., 252, 258–259, 262–264, 266–269,
 279–281
 Hua, J., 85, **106**
 Huang, S., 158, **164**
 Huang, X., 3, **68**
 Huber, J., 339–340, **376**
 Hughes, J., 355, **377**
 Hung, H. M. J., 245, **247**
 Hunsberger, S. A., 241, 243, **248**
 Hunter, J. S., 301, 309, **327**, 421–422, **438**
 Hunter, W. G., 301, 303, **328**, 421–422, **438**
 Hwang, J. T. G., 80, **108**
 Iman, R. L., 382, **410**
 Ingram, D., 310, **328**
 Irizarry, R. A., 76–79, **107**
 Ivanova, A., 258–259, 261–262, 267,
 279–280
 Jaffree, H. A., 77–79, **107**
 Jammalamadaka, S. R., 504, **531**
 Jannink, J., 3, **71**
 Jarrett, R. G., 80, 84, 105, **107**
 Jenkins, C., 351, **376**
 Jenkins, G. M., 117, **133**
 Jennison, C., 114, **134**, 206, **210**, 218–219,
 242, 245–246, **247–248**, 260, **280**
 Jensen, J., 3, **68**
 Jiang, Q., 173, **211**
 Jinks, J. L., 1, 3–4, 17, **69**
 JMP, 425, **439**
 John, J. A., 111, **135**, 482, **499**
 John, P. W. M., 59, **68**
 Johnson, F., 367, **376**
 Johnson, M. E., 399, 402, **410**
 Jones, B., 368, **376**, 472, 496, **499**
 Jones, S., 449, **467**
 Jorgensen, A., 369, **375**
 Jupp, P. E., 507–508, 511–514, 525–526,
 530, **532**
 Kabera, M. G., 155, **163**
 Kageyama, S., 32, 45–46, 50, 54, 56–57,
 59, **66–69**
 Kalbfleisch, J. D., 195, **210**
 Kang, M. S., 65, **68**
 Kang, S. M., 452, **469**
 Kanji, G. K., 508, 511, 520, 528, **531**
 Kannenberg, L. W., 12, 21, **69**
 Kaplan, E. L., 190, **210**
 Karlin, S., 509, **532**
 Kasprzak, E. M., 461, **468**
 Kaul, R., 455, **467**
 Kay, R., 275, **279**
 Kayo, T., 80, **107**

- Keane, M. P., 370, **375**
 Kearsey, M. J., 2, 4, 16, **68**
 Kelly, E., 327, **327**
 Kempthorne, O., xviii, **xviii**, 1, 6, 8, 12–13,
 27–30, 33, 38–39, **68–69**, 417–418, **439**
 Kempton, R., 113–114, **132–134**
 Kendall, D. G., 450, **467**
 Kendzierski, C. M., 85, **107**
 Kennedy, M., 405–407, **410**
 Kenny, P., 355, **376**
 Kenward, M. G., 98, **107**, 115, 118, 125,
 134–135
 Kerr, M. K., 80–81, 84–85, 87–88, **107**
 Kessels, R., 339, 368, **376**
 Khanin, R., 87, **108**
 Khuri, A. I., 138, 141, 162, **163**, 452, 456,
 468, 472, **499**
 Kiefer, J., 44, **68**, 111, **134**, 144–146, **163**,
 285, **297**
 Kienitz, K. H., 419, **441**
 Kim, I. F., 73, **107**
 Kim, K., 235, **248**
 Kim, S.-H., 416, **439**
 King, M., 351, 355, **376**
 Kirk, H. J., 114, **134**
 Kleijnen, J. P. C., 416, 418, 421, **439**
 Knapp, S. J., 3, **68**
 Knox, S., 369, **377**
 Kodak, C. F., 115, **132**
 Koechler, F. E., 115, **132**
 Köhne, K., 241, 245, **247**
 Koornneef, M., 3, **68**
 Kowalski, S. M., 482, 485–486, 490–491,
 493, **500**
 Krams, M., 277, **280**
 Kuhfeld, W. F., 365, 370–371, **376**
 Kuznetsova, O., 258, **280**
- Lachin, J., 205, **210**
 Lachin, J. L., 252, 255, 258, 261, 266, **280**
 Läuter, E., 287, **298**
 Lagakos, S. W., 178, **211**
 Lakatos, E., 205, **211**
 Lan, H., 85, **107**
 Lan, K. K., 182, 194, 205–206, **211**, 214,
 218, 220, 223, 225–228, 232, 234–235,
 238–239, 244, 246, **248**
 Lanesar, E., 351, **376**
 Lander, E. S., 3, **68**
- Landgrebe, J., 81, 93–94, **107**
 Langhans, I., 496, **499**
 Larntz, K., 149, **163**
 Lash, A. E., 73, **107**
 Lau, S., 323, **329**
 Law, A. M., 422, **439**
 Laycock, P. J., 509, **532**
 Ledoux, P., 73, **107**
 Lee, H. K. H., 393, **410**
 Lee, J. W., 119, **135**
 Lee, Y., 450, **468**
 Lee, Y. J., 178, **211**
 Lehman, A., 430, **440**
 Lehmann, E. L., 276, **280**
 Lentner, M., 481–482, 484, **499**
 Letsinger, J. D., 481–482, 484, **499**
 Lewis, J., 226, **249**
 Lewis, K. E., 401, **468**
 Lewis, S. M., 155, **163–164**
 Lewis, T., 503, 507, 525, **531**
 Li, P., 310, **329**
 Li, R., 390–391, **410**
 Li, W., 320, **328**, 462, **467–468**
 Li, Z., 234, **248**
 Liang, K. Y., 120, **133**
 Liebman, J., 529, **531**
 Lill, W. J., 114, **134**
 Lim, L. Y., 178, **211**
 Lin, C. S., 111, **134**
 Lin, D. K. J., 402, **410–411**
 Lin, D. M., 78, 81, **108**
 Lockshin, L., 369, **377**
 Long, J. S., 370, **376**
 Longworth, L., 355, **376**
 Lopez, G. A., 116, **133**
 Louviere, J. J., 340, 351, 355, 363, 369–370,
 374–376, **378**
 Lu, L., 461, **468**
 Luan, Y., 294, **297**
 Lucas, J. M., 453, 461, **467**
 Lucas, N. J., 423, **439**
 Lucas, T. W., 416, 418–419, 421, 425, 437,
 438–440
 Luu, P., 78, 81, **108**
 Lynch, M., 2, **68**
- Ma, C. X., 321, **328**
 Ma, F. S., 310, **328**
 MacKay, R. J., 448–449, **467**, **469**

- MacKay, T. F. C., 2, **67**
Mackay, W., 529, **531**
Maddala, T., 367, **376**
Malhotra, R. S., 116, **135**
Mallenby, D. M., 293, **298**
Mandal, S., 261, **279**
Manson, R. A., 60–61, **69**
Mansson, R. A., 59, **68**
Mantel, N., 193, **211**
Mardia, K. V., 501–504, 506–509, 511–514,
 522, 525–526, 530, **532**
Maria-Joao, P., 3, **68**
Marin, J.-M., 393, **410**
Markaryan, T., 272, **280**
Markoff, J., 417, **439**
Marley, A., 369, **376**
Marshall, A. W., 306, **328**
Marshall, K. A., 73, **107**
Martin, R., 111, 114, 117, **134**
Martin, R. J., 51, **69**, 117, **134**
Martinez, F., 437, **439**
Martinsson, P., 370, **374**
Marubini, E., 229, **248**
Mather, D. E., 113, **136**
Mather, K., 1, 3–4, 17, **69**
Mathew, T., 149, 152, **163**
Mathur, S. N., 56, **69**
Matzinger, D. F., 6, **69**
Mayo, O., 113–114, **135**
McCullagh, P., 138, **163**
McCulloch, C. E., 138, **163**
McDonald, M. G., 393, 395, **410**
McIntosh, E., 355, **377**
McIntyre, G. A., 80, **107**
McIntyre, L., 3, **71**
McKay, M. D., 384–385, 400–401, **410–411**
McLoad, R. G., 327, **328**
McNamara, J., 275, **279**
McPherson, C. K., 220, **246**
Mead, R., 114, **134**, 320, **329**
Mee, R. W., 311, 327, **327–328**
Mehnen, J., 461, **469**
Meier, P., 190, **210**, 223, **248**
Melfi, V. F., 263, **280**
Mendez, I., 114, **135**
Michelson, A. M., 79, **107**
Midha, C. K., 47–48, 51, 57, 60, **67**
Miller, A., 327, **328**
Miller, E., 278, **280**
Miller, G. A., 367, **377**
Miller, P. B., 278, **280**
Miller, R. G., 189, 192, 195, **211**
Milliken, G. A., 106, **108**
MIL-STD-105E, 220, **248**
Minkin, S., 149, 151, **163**
Miro-Quesada, G., 460, **468**
Mitchell, T. J., 382, 401, 403, 405, **411**, 416,
 440
Mitroff, I. I., 415, **439**
Monaghan, F., 1, **69**
Monroe, R. J., 114, **134**
Montgomery, D., 421–422, **439**
Montgomery, D. C., 449, 455–456, 461,
 467–468, 472, 474, 482, **499**
Moore, C. S., 113, **134**
Moore, L., 327, **327**
Moore, L. M., 385, 399, 402, **410–411**
Moore, T. J., 214, **248**
Morgan, J. P., 111, **134**
Morris, M. D., 385, 401, 403, 405, **411**
Mosteller, F., 79, **108**, 220, **247**
Mountz, J. D., 80, **107**
Mourato, S., 334, **376**
Müller, M., 516, **532**
Mueller, S., 369, **377**
Muertter, R. N., 73, **107**
Mukerjee, R., 53–54, 56, 59, **66**, **69**, 92,
 106, 287, 293, **297**, 307, **327**
Mukherjee, B., 138, 141, 162, **163**
Mulitz, D. K., 114–115, **135**
Mull, D. J., 115, **132**
Mullen, K., 369, **375**
Murphy, J., 278, **280**
Myers, R. H., 447, 449–450, 452, 455–456,
 461, 463, **467–469**, 472, 481–482, 484,
 499
Myers, W. R., 450, 455, 463–466, **467–469**
Nachtsheim, C. J., 320, **328**, 462, **468**, 472,
 499
Naderman, G. C., 114, **135**
Nair, K. R., 110, **138**
Nair, V. N., 447–448, **468**
Nannini, C. J., 423, 426, **439**
Narain, P., 27, 30, 56, **66**, **69**
Ndlovu, P., 155, **163**
Nelder, J. A., 138, **163**, 450, **468**
Nelson, B. L., 416, 421, **439**, **441**

- Nelson, L. A., 114, **135**
 Nettleton, D., 80, 85, **108**
 Ngai, J., 78, 81, **108**
 Nguyen, D. V., 74, **108**
 Nies, A. S., 213, 230, **247**
 Nigam, A. K., 27, **69**
 NIST/Sematech, 421, **439**
 Nobile, A., 87, **108**
 Noble, R. B., 523, **531**
 Nolan, K. B., 178, **211**
 Notz, W. I., 419, 422, **440**
 Nychka, D., 392, **410**
- O'Brien, P. C., 221–223, 225, **248**
 O'Brien, T. E. O., 155, **163**
 Office of the Secretary of Defense, 423,
 437, **440**
 O'Hagan, A., 406–407, **410**
 Ohnishi, T., 293, **298**
 Old, P., 355, **377**
 Olivas, J. D., 327, **327**
 Olkin, I., 306, **328**
 Osborn, K., 437, **440**
 Oshlack, A., 77, **108**
- Page, C., 263, **280**
 Page, G. P., 80, **107**
 Panda, D. K., 60, **69**
 Papadakis, J. S., 113, **134**
 Parker, P. A., 485–486, **500**
 Parsad, R., 49–51, 54, 56, 60, **67, 69–70**
 Patel, J. D., 12, 21, **69**
 Patterson, H., 110, **134**
 Paula, G., 521, **531**
 Payne, R. W., 65, **69**
 Pearce, S. C., 56, **69**, 113, **134**
 Pederson, D. G., 30, **69**
 Peduzzi, P., 178, **211**
 Peng, V., 78, 81, **108**
 Permana, P. A., 80, **107**
 Pesarin, F., 269, **280**
 Peters, T. J., 369, **375**
 Peterson, J. J., 460, **468**
 Peto, J., 192–193, **211**
 Peto, R., 192–193, 209, **211**, 226, **249**
 Petschek, A. G., 385, **410**
 Pfeiffer, W. H., 114, 118, **134**
 Phadke, M. S., 446, **468**
 Philliply, K. H., 73, **107**
- Phillips, K., 367, **376**
 Phimister, B., 74, **108**
 Phoa, F. K. H., 322, **329**
 Piantadosi, S., 168, 178, **211**, 274, **279**
 Pickle, S., 451, **468**
 Pike, D. J., 114, **134**
 Pintar, A., 461, **469**
 Plackett, R. L., 149, **162**
 Pocock, S., 168, **211**
 Pocock, S. J., 220–223, **248**, 257, **280**
 Ponnuswamy, K. N., 63, **69–70**
 Pooni, H. S., 2, 4, 16, **68**
 Posch, M., 246, **248**
 Poushinsky, G., 111, **134**
 Preece, D. A., 45, **69**
 Prentice, R. J., 270, **280**
 Prentice, R. L., 195, **210**
 Prescott, P., 59–61, **68–69**
 Prewitt, K., 451, **467**
 Probstfield, J., 209, **211**
 Prolla, T. A., 80, **107**
 Proschan, M., 213–214, 218, 220, 223,
 225–228, 230, 235, 238–239, 241,
 243–244, 246, **247–249**
 Proschan, M. A., 182, 206, **211**
 Pukelsheim, F., 87, **108**, 146, **163**, 285, **298**,
 338, **377**
- Qian, P. Z. G., 402, **411**
 Qiao, C. G., 116, **134**
 Quinlan, J. A., 277, **280**
- Raghavarao, D., 111, **133**
 Rajagopal, R., 460, **468**
 Ramberg, J. S., 421, **440**
 Rao, C. R., 40, **70**, 285, **298**
 Rao, J. S., 510, **532**
 Rao, S. B., 59, **68**
 Ratcliffe, J., 355, **376**
 Rawlings, J. O., 61–64, **70**
 Reboussin, D. M., 235, **248**
 Reddel, H. K., 351, **376**
 Richardson, M., 368, **377**
 Ricks, M., 261–262, 267, **280**
 Ripley, B. D., 115, **135**, 370, **378**
 Ritchie, M. E., 77, **108**
 Robert, C. P., 393, **410**
 Robins, J. M., 178, **211**, 218, **248**
 Robinson, H. F., 9, 12, **66**

- Robinson, T. J., 447, 451, 455, 461, 463–466, **467–469**
 Rogatko, A., 155, **164**
 Roger, J. H., 98, **107**, 118, **134**
 Roginski, J. W., 437, **439**
 Romig, H. G., 220, **247**
 Rosa, G. J. M., 106, **108**
 Rose, J. M., 368, **374**
 Roseberry, T. D., 220, **248**
 Rosenberger, W. F., 252–253, 255–259, 261–264, 266–268, 271–272, **279–281**
 Rowe, B. C., 220, **246**
 Roy, R. M., 30, **70**
 Royall, R. M., 278, **281**
 Ruck, J., 423, 430, **438**
 Rudnev, D., 73, **107**
 Ruggiero, K., 80, 84, 107, **107**
 Ruppert, D., 80, **108**
 Ruskin, J., 213, 230, **247**
 Russell, K. G., 155, **163–164**
 Ryan, M., 334, 340, 355, **377**
 Ryan, T. P., 422, **440**
 Ryne, R. D., 405, **410**
 Sacks, J., 398, **411**, 416, **440**, 452, **469**
 Saeger, K., 415, **440**
 Sall, J., 430, **440**
 Saltelli, A., 384, **411**
 Sanchez, P. J., 421, **440**
 Sanchez, S. M., 416, 418–419, 421, 437, **439–441**
 Sandor, Z., 368, 370, **377**
 Santner, T. J., 419, 422, **440**
 Sarkar, S., 60, **70**
 Sarker, A., 116, **135**
 Savage, E., 363, **378**
 Sawa, T., 287, **298**
 Scarpa, R., 332, 368, 370, **375**
 Schabenberger, O., 237–238, **249**
 Scharfstein, D. O., 218, **248**
 Scheffé, H., 284, **298**
 Schena, M., 73, 75, **108**
 Scherf, U., 77–79, **107**
 Schiably, J. H., 385, **410**
 Schiller, S. B., 398, **411**
 Schlottfeldt, C. S., 114, **133**
 Schmidt, J., 8, **70**
 Schmitz, N., 242, **249**
 Schneiderman, M., 220, **246**, **249**
 Schoenfeld, D., 204, **211**
 Schron, E., 213, 230, **247**
 Schwabe, R., 346–347, 365–366, 370, **375**
 Scott, A., 355, 357, **377**
 Scott, D. S., 382, **411**
 Searle, S. R., 138, **163**
 Seheult, A., 114, **134**
 Sellke, T., 234, **249**
 Sen, S., 3, **70**
 SenGupta, A., 504, **531**
 Sengupta, S., 510, **532**
 Seraphin, J. C., 114, **134**
 Shaffer, J. G., 51, **70**
 Shah, B. V., 62, **70**
 Shalon, D., 73, 75, **108**
 Shankar, A., 51, **70**
 Shao, J., 266, **281**
 Sharma, M. K., 51, **70**
 Sharma, V. K., 60, **69**
 Shen, H., 421, **440**
 Sherman, P. M., 73, **107**
 Shewry, M. C., 399, **411**
 Shih, J., 205, **211**
 Shih, J. H., 84–85, **107**
 Shih, W. J., 241, **249**
 Shirakura, T., 293, **297–298**
 Shuler, K. E., 385, **410**
 Sickinger, L. R., 437, **439**
 Siegmund, D., 234, **249**
 Silva, J. C. E., 116, **133**, **135**
 Silver, J., 77, **108**
 Silver, J. D., 77, **108**
 Silvey, S. D., 92, **108**, 141, 150, 152, **163**
 Simon, R., 84–85, **107**, 257, 275, 278, **280–281**
 Sinclair, E. J., 422, **440**
 Singh, M., 27, 30–31, 33, 37, 39, 44–45, 51, **56**, **70**, 116, **135**
 Singh, R. K., 2, 65, **70**
 Singhi, N. M., 59, **67**
 Sinha, B. K., 138, 141, 149, 152, 162, **163**
 Sitter, R. R., 141, 149, 155–156, **163–164**, 323, 327, **327–328**, 462, **467**
 Sleight, P., 226, **249**
 Sloane, J., 421–422, **439**
 Sloane, N. J. A., 334, 347, **376–377**
 Slud, E. V., 234, **249**
 Smith, A. B., 111, 116, 122, 127, **133**, **135**
 Smith, R. L., 255, **281**

- Smyth, G. K., 77–78, **108**
 Snappin, S., 173, **211**
 Sobol, I. M., 384–385, **411**
 Soboleva, A., 73, **107**
 Solomon, P. J., 89, 91–92, **107**
 Son, Y. N., 44, **66**
 Speed, T. P., 77–79, 81, 85, **107–108**
 Spiegelhalter, D. J., 232, **249**
 Sprague, G. F., 7, 9, **70**
 Spurr, B. D., 509, **532**
 Srinivasan, M. R., 63, **70**
 Srivastav, S. K., 51, 60, **70**
 Srivastava, J. N., 283, 286–287, 291, 293,
297–298
 Srivastava, R., 49–50, 59, **67, 69–70**
 Stallard, N., 261–262, 267, **280**
 Stam, P., 3, **70**
 Stefanova, K. T., 122, **135**
 Steibel, J. P., 106, **108**
 Stein, C., 236, **249**
 Stein, M. L., 115, **135**, 389, 393, 401,
410–411
 Steinberg, D. M., 155, **163**, 287, **297**,
 305–306, 314, 321, 325, **328**, 402, **411**
 Steiner, S. H., 448–449, **467, 469**
 Stephens, M. A., 506, 509, 519, 525, **532**
 Stern, K., 9, 27, **68**
 Stram, D. O., 119, **135**
 Strauss, H., 213, 230, **247**
 Street, A. J., 111, **135**
 Street, A. P., 365, **377**
 Street, D. J., 51, **70**, 111, **135**, 331, 335,
 339–341, 343, 345, 347, 349, 351,
 355–360, 362, 365, 369, **374, 377**
 Stroup, W. W., 106, **108**, 114–115, **135**
 Stufken, J., 138, 141, 147–149, 158, 160,
164, 334, 347, **376**, 421–422, **439**
 Studden, W. J., 509, **532**
 Subbarao, C., 27, **69**
 Sudjianto, A., 390, **410**
 Suen, C.-Y., 316, **328**
 Suetsugu, T., 293, **298**
 Stich, R., 460, **467**
 Sun, D. X., 287, **297**, 305–308, 314, 319,
 321, 325, **327–328**
 Sutton, T. W., 522, **532**
 Sverdlov, O., 256–257, **280**
 Swait, J., 340, 363, **376**
 Sword, A. M., 114, **134**
 Tabis, Z., 59, **66**
 Taguchi, G., 421, **441**, 444, 446–447,
468–469
 Takahashi, T., 293, **298**
 Tamura, R. N., 114, **135**
 Tan, W. Y., 274, **279**
 Tang, B., 287, **297**, 310, 315, 317–321,
327–328, 402, **411**
 Tatum, L. A., 7, **70**
 Tempelman, R. J., 85, 106, **108**
 Teschmacher, L., 293, **297**
 Thomas, W. T. B., 3, **67**
 Thompson, L., 370, **377**
 Thompson, R., 110, **134**
 Tian, Y., 289–290, 294, **297**
 Tighiouart, M., 155, **164**
 Tindall, J., 369, **375**
 Tobias, R. D., 285, **297**, 338, **374**
 Tomashovsky, M., 73, **107**
 Tombak, L. M., 312, **328**
 Toran, L., 393, **410**
 Torgerson, W., 369, **377**
 Torsney, B., 146, 149, 151, 155–156, **163**
 Train, K. E., 335–336, 370, **377**
 Trautmann, H., 461, **469**
 Travers, S. E., 106, **108**
 Trinca, L. A., 482, **500**
 Troup, D. B., 73, **107**
 Tsai, P.-W., 320–321, 325–327, **327, 329**
 Tsiatis, A. A., 218, 226, 234, **248–249**
 Tsuji, T., 293, **298**
 Tukey, J. W., 79, **108**, 414, **441**
 Turnbull, B. W., 206, **210**, 218–219, 242,
 245–246, **247–248**, 260, **280**
 Tyroler, H., 209, **211**
 Uddin, N., 111, **134**
 Underwood, A. J., 519, **532**
 Upton, G. J. G., 509, **532**
 Valsecchi, M. G., 229, **248**
 Vandebroek, M., 339, 368–370, **376, 378**,
 496, **499**
 Vartak, M. N., 30, **70**
 Venables, W. N., 370, **378**
 Verbyla, A. P., 114–117, 119, 122, 125,
134–135
 Verdinelli, I., 368, **375**
 Verhoeven, K. J., 3, **71**

- Vermeulen, B., 369, **378**
 Verter, J., 213, 230, **247**
 Vieira, H., 419, **441**
 Villafranca, R. R., 463, **469**
 Vining, G. G., 449, 451–452, 456, **468–469**,
 482, 485–486, 490–491, 493, **500**
 Vivacqua, C., 327, **329**
 von Mises, R., 504, 506, **532**
- Waincko, A. T., 113, **135**
 Wald, A., 219, 222, **249**
 Wallis, W. A., 220, **247**
 Walsh, B., 2, **68**
 Walters, L., 213, 230, **247**
 Walther, A., 405, **410**
 Wan, H., 421, **440–441**
 Wang, H., 310, **328**
 Wang, N., 74, **108**
 Wang, S.-J., 245, **247**
 Warren, J. A., 114, **135**
 Wasi, N., 370, **375**
 Watson, G. S., 382, **411**, 506–507, 509, 519,
 525, **532**
 Wedel, M., 368, 370, **377**
 Wehlau, L., 312, **327**
 Wehner, R., 516, **532**
 Wei, L.-J., 234, **249**, 255, 258–259, **281**
 Weijer, C., 278, **280**
 Weindruch, R., 80, **107**
 Weinstein, M. C., 278, **281**
 Weir, B. S., 3, **67**
 Welch, W. J., 398, **411**, 416, **440**, 452, **469**
 Welham, S. J., 115, 125, **135**
 Wellendorf, H., 116, **133**
 Wendler, D., 258, **281**
 Wilhite, S. E., 73, **107**
 Wilkinson, G. N., 113–114, **135**
 Willham, R. L., 6, 8, **71**
 Williams, B. J., 419, 422, **440**
 Williams, E. J., 111, **135**, 506, 509, 519,
 525, **532**
 Williams, E., R., 111, **135**
 Wilson, K. B., 453, **467**, 472, **499**
 Winker, P., 402, **410**
 Wit, E., 87, **108**
 Wittes, J., 178, 182, 194, 206, 209, **211**, 214,
 218, 220, 223, 225–228, 235, 237–239,
 244, 246, **248–249**
 Wolfowitz, J., 146, **163**, 219, **249**
- Wong, W. K., 322, **329**
 Woodroffe, M. B., 267, **279**
 Woods, D. C., 155, **163–164**
 Wright, R. E., 334, **376**
 Wu, C. F. J., 146, 149, 151, **163–164**,
 307–308, 310, 312, 315–317, 321–324,
 327–329, 462, **469**, 508, 519–521, **531**
 Wu, H., 310, 312, **329**, 510, **532**
 Wu, T., 113, **136**
 Wu, Z., 76–79, **107**
 Wulff, S. S., 464, **467**
 Wyle, H., 385, **411**
 Wynn, H. P., 111, **134**, 399, **411**, 416, **440**
- Xu, H., 317, 321–323, **329**
- Yang, M., 138, 141, 147–149, 152, 158,
 160, **164**
 Yang, Y. H., 78, 81, 85, **107**
 Yates, F., 110, **136**, 364, **378**
 Ye, K. Q., 321, **328**, 419, **441**
 Yeoh, A., 355, **376**
 Ylvisaker, D., 399, 402, 405, **410–411**
 Yu, J., 370, **378**
 Yu, T. K., 452, **469**
 Yu, X., 266, **281**
 Yuan, A., 262, **281**
 Yusuf, S., 209, **211**, 226, **249**
- Zeger, S. L., 120, **133**
 Zelen, M., 167, 180, **211**, 257, **281**
 Zellner, A., 393, **411**
 Zeng, Z.-B., 3, **67**, **71**
 Zhang, B., 158, **164**
 Zhang, L.-X., 259, 262–263, 269,
 280–281
 Zhang, L., 253, 259, 261, 263, 268, **281**
 Zhang, R. C., 310, **329**
 Zhang, W., 85, **108**
 Zhang, X. D., 293, **297**
 Zhang, Y., 85, **107**, 402, **410**
 Zhao, S. L., 310, **329**
 Zhong, B., 266, **281**
 Zhu, W., 141, 146, 149, **162**
 Zimmerman, D. L., 114–115, 117, **136**
 Zucker, D., 237–238, **249**
 Zunica, L., 463, **469**
 Zunica, R. R., 463, **469**
 Zwerina, K., 339–340, **376**

Subject Index

- Adaptive clinical trial, 277
- Adaptive decision rule, 252
- Adaptive design, 251
- Adaptive randomization, 251
- Adaptive sampling, 252
- Adaptive stopping rule, 252
- Algorithm, 384
 - search, 461
 - summarization, 79
- Alias set, 300, 325
- Alias structure, 304, 311
- Allocation
 - Neyman, 260
 - optimal, 258, 260
 - proportion, 260
 - optimal, 261, 263, 267
 - rule, random (RAR), 254
- Analysis
 - Bayesian, 398
 - directional data, 502
 - flexibility, 419
 - generation mean, 4
 - group sequential, 206
 - interim, 214, 224, 231–324, 238
 - local sensitivity, 382–383
 - microarray data, 95
 - nearest neighbor, 113
 - predictive, 392
 - randomization based, 117, 123
 - of randomized clinical trial, 177
- sequential, 206, 219, 252
- spatial, 109, 113, 117, 123
- survival, 186
- uncertainty, 383
- Analysis of variance (ANOVA), 62, 509, 526
 - Hayman’s, 14
 - multivariate, 522
 - multiway, 523, 528
 - one-way, 517, 523
 - two-way, 528
- Approach
 - analytic, 141, 146, 150, 162
 - Bayesian, 368, 392
 - predictive posterior, 458
 - dual response, 450
 - embedding, 511
 - geometric, 141, 151
 - Griffing’s, 65
 - Hayman’s, 65
 - Monte Carlo, 400
 - multistage, 141, 147
 - nominal p -value, 238
 - Pareto frontier, 461
 - response surface, dual/single models, 444
 - semi-parametric, 451
 - Taguchi-based, 444
 - tangential, 514
 - two-step, 446

Design and Analysis of Experiments: Special Designs and Applications, First Edition. Edited by Klaus Hinkelmann.

© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

- Array
 - combined, 452
 - crossed
 - design, 447
 - experiment, 445
 - fractionated, 449
- Bayesian monitoring, 246
- Bias, 175, 285
 - accidental, 252
 - estimation, 385
- Bifurcation, controlled sequential, 421
 - fractional factorial, 421
- Bonferroni correction, 208
- Bootstrap procedure, 267
- Boundary
 - monitoring, 223
 - O'Brien-Fleming, 222
 - Pocock, 221
- Brownian motion, 217, 224
- Choice set, 332, 340
 - size, 332
- Central limit theorem, 218
 - Lindeberg, 224
 - martingale, 226
- Clinical trial
 - adaptive, 277
 - monitoring, 214, 219, 223
 - two-armed, 215, 221
- Competing risk, 184–186
- Control group, 174
- Combining ability
 - estimation of, 56
 - general, 7
 - estimating linear functions of, 53
 - specific, 8
 - effect, 39
- Common base option, 355
- Comparisons, paired, 358
- Computer search, 56
- Coordinate system
 - Cartesian, 504, 522
 - spherical, 504
- Correlation
 - matrix, 117, 389, 408
 - product, 387
- Covariance
 - between relatives, 2, 5
 - genetic, 63
 - nugget, 389
 - statistical, 63
 - structure 122
- Covariate, nearest neighbor, 113
- Critical value
 - O'Brien-Fleming, 222
 - no-interim-look, 222
- Cross
 - double, 63
 - four-way, 63
 - three-way, 61
- Curve
 - Kaplan-Meier, 190
 - product-limit, 190
 - survival, 185, 204–205
- Data
 - axial, 524
 - binary, 141
 - circular, 503, 509
 - count, 141–142
 - cylindrical, 503, 509, 521–522
 - directional, 501–502, 514, 518
 - endpoint, 171
 - missing, 182
 - interim, 230
 - mining, 425
 - smoother, 385
 - spherical, 503, 524, 528
- Defining relation, 300
- Defining words, 300
 - block, 323
 - independent, 300
 - treatment, 323
- Diagnostic tool, 119
- Diallel cross, 3
 - complete (CDC), 14
 - A-optimal, 57
 - orthogonally blocked, 56
 - variance balanced, 43, 48–50
 - partial (PDC), 14, 25–29
 - A-optimal, 56
 - augmented, 30
 - D-optimal, 53
 - E-optimal, 53

- efficient, 57
- in incomplete blocks, 32, 44
- Design(s), *see also* Experiment(s); Plan(s)
 - adaptive, 241, 245, 251
 - efficient randomized, 263
 - admissible, 89, 92
 - alpha, 110
 - alternating loop, 84
 - approximate, 139, 143, 155
 - array
 - combined, 456
 - crossed, 447
 - augmented, 111, 127
 - balanced, 497
 - balanced equivalent, 482, 485
 - balanced incomplete block (BIB),
 - 364
 - nested, 45, 49, 58
 - biased coin
 - doubly-adaptive, 262
 - Efron's, 254, 270, 272
 - Smith's generalized, 255
 - binary pairwise balanced, 51, 60
 - Box-Behnken, 472
 - Cartesian product, 482
 - central composite, 420, 453–454, 472, 483
 - split-plot, 481
 - on a circle, 509
 - combinatorial, 364
 - completely randomized (CRD), 7, 80–81, 86, 508
 - connected block, 44
 - construction, 347
 - diallel cross block, 54
 - binary, 56
 - optimal, 45–51
 - dye
 - balanced, 85, 88
 - swap, 84
 - environmental, 1, 7
 - equivalent
 - balanced, 482, 485
 - unbalanced, 486
 - error-control, 508
 - exact, 139
 - experimental, 381
 - factor, 461
 - factorial, 508
 - fractional (*see* fractional factorial)
 - symmetrical, 334
 - flexibility, 419
 - forced balance, 254
 - forced choice, 357
 - fractional factorial, 288, 299, 320, 334, 382
 - admissible, 326
 - complementary, 315, 317
 - even, 316
 - foldover, 302, 455
 - inadmissible, 326
 - maximal even, 302, 316–317
 - mixed-level, 321
 - mixed resolution, 453, 461
 - model-robust, 320
 - nonmaximal resolution IV, 312
 - nonregular, 317, 319, 321
 - projection of, 302–303
 - regular, 300, 318
 - resolution III, 302, 343
 - resolution III+, 302
 - resolution IV, 302
 - resolution IV+, 302, 313
 - resolution V, 347, 363, 420
 - resolution V+, 305
 - resolution V plus one, 291
 - saturated, 302, 311, 313
 - second order saturated, 311
 - split-plot, 326
 - two-level, 299
 - Græco-Latin square, 43
 - group-sequential, 241, 246
 - equal batch size, 242
 - nonadaptive, 242
 - optimal, 242
 - hierarchical, 509
 - inadmissible, 91
 - incomplete block, 7, 80, 86, 94, 110
 - resolvable, 110
 - Latin hypercube, 400–401, 418
 - nearly orthogonal, 419
 - nested, 402
 - Latin rectangle, 86
 - Latin square, 80, 86, 88
 - latinized row-column, 111
 - lattice, 110
 - line \times tester, 21

- Design(s), *see also* Experiment(s);
 Plan(s) (*cont'd*)
 mating, 1, 7, 21
 diallel type I, 8
 factorial, 9
 incomplete diallel, 13
 partial diallel, 13
 mating-environmental (M-E), 8, 33, 41
 analysis of, 36
 matrix, 419
 maximin distance, 399
 minimum aberration, 301, 304, 306,
 308
 blocked, 323
 construction of, 315
 weakly, 301
 minimum discrepancy, 402
 near-optimal, 349
 neighbor-balanced, 111
 North Carolina
 I, 9
 II, 10
 III, 11
 optimal, 138, 149, 343, 345, 356, 368,
 418
 A-, 152–153, 285, 339
 Bayesian, 149
 D-, 151–152, 155, 285, 339, 343, 359,
 463–464, 496–497
 E-, 285, 339
 G-, 339
 L-, 87
 locally, 138–139, 145
 V-, 339
 w-, 92
 optimality, 462
 orthogonal, 482
 partially balanced incomplete block
 (PBIB), 27, 34–35
 resolvable, 44
 triangular, 48, 60
 partially replicated, 112, 127
 permuted block, 254
 phase-1, 94
 phase-2, 94, 106
 point, 147, 389, 461
 induced, 148
 randomized complete block (RCBD),
 7, 80–81, 110
 reference, 85
 region, 461
 residual, 60
 response surface, second order, 488
 robust, 421
 parameter, 443
 resolvable spatial, 111
 saturated/nearly saturated, 420
 search, 283
 factor screening, 291
 foldover, 290
 near-foldover, 290
 semi-Bayesian, 369
 shifted, 340
 single stage, 413
 size, 451
 space, 147
 filling, 419
 induced, 146
 spatial, 111
 split-block, 122
 split-plot, 80, 474
 first-order industrial, 477
 second-order industrial, 477, 487,
 496
 stated preference, 357
 strip-plot, 122
 strongly equineighbored, 51
 Taguchi, 447
 truncated binomial, 254
 two-dimensional, 111
 uniform, 402
 urn, Wei's 255
 variance spending, 241
 within-block looping, 105
 Distribution
 exponential, 192, 205
 F-, 520
 approximate, 528
 jumping, 393
 marginal, 385
 normal, 493, 504
 bivariate, 216
 multivariate, 218
 standard, 193, 218
 prior, 368
 symmetric unimodal, 504
 uniform, 504, 507
 circular, 505

- von Mises, 504, 510–512, 519
 - Fisher, 506
 - l*-modal, 509
- Watson, 507, 526, 529
 - axial, 528
 - concentrated bipolar, 507, 528
 - spherical, 526
- Effect(s)
 - additive, 4
 - epistatic, 61
 - dominance, 4
 - group, 158
 - higher order, 288
 - interaction, 351
 - two-factor, 362
 - lower order, 288
 - main, 288, 336, 343, 351, 362
 - non-negligible, 288
 - position, 361–362
 - random block, 325
 - sparsity, 382
 - subplot, 475, 477, 488
 - nominal, 491
 - treatment, interim, 241
 - whole-plot, 476–477, 488
- Efficiency
 - A-, 59
 - D-, 463
 - D_s -mean, 464
 - D_s -variance, 464
 - e-, 93
 - loss of, 61
 - reduction of, 59–60
- Empirical Bayes, 389
- End point, time-to-event, 225
- Entropy, 399
- Equivalence theorem, 146
- Error
 - conditional, 238
 - function, 243
 - cross-validation, 394
 - inflation, 223, 234
 - measurement, 388, 408
 - minimum integrated mean square, 398
 - sampling, 388
 - squared, loss, 445
 - subplot, 475
 - whole-plot, 475
- Estimate(s), *see also* Estimation; Estimator(s)
 - empirical best linear unbiased (E-BLUE), 118, 131–132
 - least squares, 523
 - maximum likelihood, 505, 519, 523
 - residual (REML), 96, 123
 - restricted (REML), 96, 123
 - OLS-GLS equivalent, 482
- Estimation, *see also* Estimate; Estimator
 - bias, 385
 - capacity, 304, 309, 324
 - full, 287
 - maximum, 304–305, 309, 314
 - index, 310–312
 - maximum likelihood, 390
- Estimator(s), *see also* Estimate(s); Estimation
 - least squares, 294
 - maximum likelihood, 139, 266
 - product-limit, 191
 - Tukey's bi-weight, 79
- Experiment(s), *see also* Design(s);
 - Plan(s)
 - best-worst, 369
 - scaling, 369
 - choice
 - binary, 355
 - discrete, 332
 - forced, 333, 351–353, 357
 - stated, 331–332
 - computer, 381, 387
 - crossed array, 445
 - desensitization, 448
 - diallel, 2
 - exploratory, 7
 - factorial, 523
 - genetic, 1–2, 65
 - line \times tester, 12
 - microarray, 73, 85
 - physical, 382, 418
 - robustness, 448
 - series of, 471
 - simulation, 416–417, 421–422
 - split-plot, 449, 471
 - second order, 492
 - stated preference, 331, 357
 - two-phase, 80
 - two-treatment two color, 85

- Factor(s)
 - adjustment, 446
 - blocking, 82
 - categorical, 473
 - control, 446–449, 452
 - easy-to-change, 471
 - efficiency, 38
 - average, 31
 - of M-E design, 37
 - environmental, 1
 - genetic, 1
 - hard-to-change, 471
 - noise, 444, 448–449, 452
 - categorical, 465
 - nuisance, 82
 - pseudo, 521
 - quantitative, 473
 - subplot, 471, 474
 - tuning, 446
 - whole-plot, 471, 474
- Fast Fourier transform, 385
- Finite differences, 406
- Function
 - Bessel, 505
 - confluent hypergeometric, 507
 - correlation, 386, 389, 392
 - spatial, 387
 - desirability, 461
 - discrepancy, 403, 407
 - star, 403
 - estimable, 284–285, 289
 - hazard, 184
 - likelihood, 337
 - log-, 337, 389–390
 - link, 140, 142
 - mean, 386
 - objective, 145, 461
 - optimality criterion, 289
 - spending, 232–233
 - O'Brien-Fleming like, 235
 - Pocock like, 234
 - survival, 184, 261
 - tapered covariance, 392
 - variance, 386
- Gene expression, 73–74
 - types of, 75
- Genetic mapping, 3
- Genetics, biometrical, 1
- Griffing's method, 21–25
- Hazard
 - function, 184
 - proportional, 192
 - model, 194
 - rate, 184
- Heritability, 20–22
 - broad sense, 2
 - estimation of, 5, 57
 - optimal design for, 57–58
 - narrow sense, 2, 28, 57
- Histogram, 425–426
- Hybridization, 75
- Inference
 - likelihood-based, 264, 276
 - randomization-based, 269, 276
- Information
 - capacity, 319
 - criterion
 - Akaike, 287
 - Bayesian, 287
 - Fisher's, 268
 - prior, 368
- Intent-to-treat, 177
- Interaction
 - control-by-control, 455
 - control-by-noise, 452–455, 464
 - effect, 351
 - two-factor, 345
 - clear, 307–309
 - eligible, 307
 - ineligible, 307
- Latin hypercube
 - design, 400–401, 418
 - nearly orthogonal, 419, 425
 - nested, 402
 - space filling, 425
 - orthogonal-array based, 402
 - sample, 400
- Least squares
 - estimate, 523
 - estimated weighted, 450
 - estimator, 294
 - generalized, 480
 - iterated reweighted, 481
 - ordinary, 480
 - regression, 427
 - smoothing, 114
 - theory of, 40

- Likelihood
 maximum, 394
 estimation, 390
 method of, 389
 penalized, 390
- Loss, to-follow-up, 181, 186
- Markov chain, Monte Carlo reversible jump, 392
- Matrix(es)
 contrast, 351, 357, 361
 correlation, 117, 389, 408
 ill-conditioned, 391
 numerically singular, 390
- covariance, 143
- design, 419
 -model, 462
- Hadamard, 319, 366
- idempotent, 493
- incidence, 36–38, 41
- information, 41, 47–49, 53, 337–338, 352, 355–357, 360–361
 Fisher, 140–143, 145–147, 153
- moment, 462
- scatter plot, 419
- variance, 117
- variance-covariance, 294, 338, 368, 452, 463
- Maximum likelihood
 estimate, 505, 519, 523
 estimation, 390
 estimator, 139, 266
 method of, 389
 residual (REML), 114, 481, 484
 restricted (REML), 481, 484
 sequential, 263
- Mendel, 1
- Method
 Bayesian, 461
 bootstrap, 514
 group-sequential, 214
 hybrid, 421
 randomization, 514
 resampling, 514
- Microarray
 data, 83
 definition of, 74
 experiment, 73, 85
 design of, 73, 80–82
 multifactor two-color, 89
- single-channel, 76–78
 slide, 75, 78–82, 85, 90, 94
 technology, 73–74
 two-color, 75–78, 85, 90, 94, 106
 data, 83
- Minimum aberration, 314, 324
 criterion, 301
 generalized, 321
 G_2 -, 320
 generalized, 320
 weak, 301
- Model
 additive-dominance, 19
 binary data, 149
 Bradley-Terry, 358–360
 building, 119
 calibration, 407
 computer, 381, 386–388, 406, 414
 deterministic, 380
 error-in-variables, 114
 first difference, 114
 first-order, 472
 Gaussian stochastic process, 386, 398, 405
 stationary, 391, 406–408
 treed, 393
- generalized linear, 138, 450
- geostatistical, 115
- linear
 general, 284
 response, 460–462
 search, 283–286
- linear mixed, 97, 116, 129–130
 spatial, 116
- local shift, 276
- logistic, 149–153, 157–160
 regression, 139, 155
- Markov, 205
- mathematical, 379, 414
- means, 451
- meta, 415
- mixed logit, 370
- multinomial logit (MNL), 331, 335, 368–370
 G -, 370
- nested, 286
- nonlinear, 138, 148
- parametric, 451
- Poisson regression, 149–150, 158

- Model (*cont'd*)
 - position effects, 360
 - probit, 149–153, 157–159
 - proportional hazards, 194
 - Cox's, 370
 - random field linear, 115
 - randomization-based, 131–132
 - regression, 429, 449, 473
 - logistic, 139, 155
 - scale heterogeneity, 370
 - simulation, 414, 422
 - single response, 456
 - smooth trend, 113
 - spatial, 113–114, 128, 131–132
 - trend, 114
 - stochastic, 380
 - time series, 117
 - urn, 258–259
 - real-valued, 262
 - validation, 407
 - variance, 450–451- Monte Carlo, 270, 389
 - test, conditional/unconditional, 271
- Markov chain reversible jump, 392
- Multiplicity, 206
- Noncompliance, 179
- Normalization, 77
- Observation(s), censored, 186. *See also* Data
- Optimality
 - A-, 44, 57, 144, 149
 - c-, 144
 - criterion, 44
 - Bayesian, 368
 - slope, 455
 - weighted, 464
 - D-, 44, 144, 149, 487, 497
 - D_s-, 463
 - E-, 44, 144, 149
 - F-, 149
 - L-, 58
 - MS-, 54–56
 - Φ-, 54
- Orthogonal array 317–320, 334, 364–365, 402, 421
 - asymmetric, 334
- Outcome
 - exploratory, 172
 - primary, 171
 - secondary, 171
 - time-to-event, 234
- Parameter, concentration, 505, 518–519
- Partition tree, 425
- Penalty, smoothly clipped absolute deviation, 391
- Pilot study, internal, 237–239
- Plan(s), *see also* Design(s); Experiment(s)
 - orthogonal main effect, 335, 350
- Plot(s)
 - box, 425
 - and-whisker, 426
 - check, 111–113, 127
 - circular, 517
 - control, 111
 - contour, 425
 - parallel, 425, 436
 - probability
 - half-normal, 521
 - normal, 494
 - residual, 119, 494
 - scatter, 425
- Population
 - reference, 168
 - study, 169–170
 - target, 168–170
- Power
 - conditional, 229–230, 241–243
 - predictive, 232
 - unconditional, 227, 243
- Precision, 175
- Principal components, 405
- Process
 - averaging, 430
 - binning, 430
 - cross-validation, 391
 - discrete Brownian bridge, 404
 - Gaussian stochastic, 386, 398, 405
 - stationary, 406–408
 - mean, 464

- random
 - autoregressive (AR(1)), 117
 - autoregressive integrated moving average (ARIMA), 114
 - lattice, 117
- stationary, 387
- variance, 464
- Profiler
 - interaction, 425
 - prediction, 425, 434
- Projection, k -dimensional, 303
- Quantitative trait locus (QTL), 3
- Randomization, 175
 - adaptive, 251, 268
 - covariate, 256–257, 266, 270, 275
 - response, 258, 261, 264–267, 271
- complete, 255, 272
- probability, 253
- restricted, 253, 257, 266
- strata, 209
- stratified, 194
- test, 253, 269, 276
- Random
 - field, stationary, 120
 - process
 - autoregressive (AR(1)), 117
 - autoregressive integrated moving average (ARIMA), 114
 - lattice, 117
- Reference set, 273
 - conditional/unconditional, 270
- Regression
 - backward elimination, 523
 - gamma, 451
 - least squares, 427
 - model, 429, 449, 473
 - modeling, 274
 - multiple linear, 523
 - polynomial, 113, 385
 - stepwise, 425
 - tree, 430
 - trend, 113
- Resampling, 514
- Response surface methodology, 448, 471
- Robustness, 59
 - of BIB designs, 59
 - of group-divisible designs, 59
 - of Youden square designs, 59
- Sample size, 168, 180–182, 203–205, 220, 224, 227, 461
 - calculation, 227, 235
 - fixed-sample, 244
 - interim, 241
 - modification, 236, 239–240
 - pretrial, 243
- Satterthwaite procedure, 490
- Sensitivity
 - first-order, 384
 - main-effect, 384
- Signal-to-noise ratio, 444
- Simulation
 - constructive, 414
 - experiment, 416–417, 422
 - live, 414
 - model, 414, 422
 - run, 388
 - virtual, 414
- Software, 502
 - CIRCSTAT, 502
 - DDSTAP, 502
 - DESIGN EXPERT, 462
 - GLIM, 371
 - group-sequential, 220
 - image processing, 77
 - JMP, 394, 425, 462
 - Lan-DeMets, 216
 - MATLAB, 142
 - MINITAB, 462
 - R, 16, 65, 371
 - SAS, 65, 142, 195, 258, 371, 458, 462
 - PROC MIXED, 98
 - PROC LIFETEST, 199
 - PROC OPTEX, 462
 - PROC PHREG, 199
 - PROC TRANSREG, 371
 - spending function, 235
 - S-PLUS, 201
 - SPSS, 142
 - STATA, 502
 - statistical, 65
- Spatial analysis, 109, 113, 117, 123
- Spatial autocorrelation, 128

- Spatial design, 111
 Spatial model, 113–114, 128, 131–132
 Spatial trend, 109
 Spatial variability, 109
 Stationarity, second-order, 120
 Statistic, *see also* Test
 - linear rank, 234
 - logrank, 225
 - Stein's *t*-, 237
 - Wald, 118
 Stratification
 - adaptive, 275
 - pre-, 274
 Study(s), *see also* Experiment(s)
 - dose-response, 139
 - growth, 139
 Summarization, 78
 - algorithm, 79
 Support points, 143, 148, 160
 Survival
 - analysis, 186
 - curve, 185, 204–205
 - function, 184
 - exponential, 261
 - method, 182
 - probability, 189, 192
 - time, 184, 190–192
 Table, life, 185–187, 191
 Test, *see also* Statistic
 - Dunnett's, 207
 - of equality of concentrations, 512
 - Fourier amplitude sensitivity, 385
 - Gehan, 196
 - likelihood ratio, 510, 526
 - linear rank, 270
 - logrank, 193–196, 214
 - weighted, 194
 - Mantel-Haenszel, 193
 - Monte Carlo, conditional/
 - unconditional, 271
 - multisample Watson-Williams, 511
 - nonparametric, 512
 - permutation, 269
 - randomization, 253, 269, 276
 - residual maximum likelihood ratio (REMLRT), 118, 124
 - sequential probability ratio (SPRT), 219
 stratified, 273
 stratum-specific, 274
 Wald, 125–126, 129, 264
 Wilcoxon, 207
 Time
 - failure, 178, 205
 - follow-up, 168
 - loss-to, 205
 - information, 217, 225, 230, 233
 - line, 173
 - recruitment, 205
 - survival, 184, 190–192
 Trait, quantitative, 1–2
 - locus, 3
 Treatment, subplot/whole-plot, 475
 Trend
 - global, 115, 124
 - local, 115, 121
 Trial
 - confirmatory, 166
 - double-masked, 168
 - equivalence, 209
 - monitoring, 223
 - noninferiority, 209
 - phases 1/2/3, 166
 - preclinical, 166
 - randomized clinical, 165–167, 213
 - analysis of, 177
 - cohort, 183, 187
 - single-/triple-masked, 168
 U.S. Food and Drug Administration (FDA), 277, 279
 Validity, external/internal, 171
 Variable, nonuniform control, 466
 Variability, reduction in, 383
 Variance, *see also* Variation
 - additive genetic, 5
 - component, 23–25, 61, 529
 - additive, 18
 - dominance, 18
 - environmental, 9–10
 - genetic, 9–12, 17, 62–64
 - statistical, 62–65
 - conditional, 383, 389
 - dominance genetic, 5
 - genetic, 5, 27, 63–65

- interblock/intrablock, 325
- minimum process, 455
- model, 450
- phenotypic, 63
- statistical, 63
- unconditional, 384
- Variation, *see also* Variance
 - environmental, 5
 - extraneous, 116
 - first-order, 383
 - genetic, 5
- noise, 457
- total, 383
- Variogram
 - Cartesian semi-, 119
 - face, 129–130
 - omni-directional semi-, 120
 - sample, 119, 125–129
 - semi-, 120
- Wordlength pattern, 303, 316, 323
 - generalized, 320

WILEY SERIES IN PROBABILITY AND STATISTICS
ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Harvey Goldstein, Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*

Editors Emeriti: *Vic Barnett, J. Stuart Hunter, Joseph B. Kadane, Jozef L. Teugels*

The *Wiley Series in Probability and Statistics* is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

- † ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data, *Second Edition*
AGRESTI · An Introduction to Categorical Data Analysis, *Second Edition*
AGRESTI · Categorical Data Analysis, *Second Edition*
ALTMAN, GILL, and McDONALD · Numerical Issues in Statistical Computing for the Social Scientist
AMARATUNGA and CABRERA · Exploration and Analysis of DNA Microarray and Protein Array Data
ANDÉL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, *Third Edition*
* ANDERSON · The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) · Advances in Biometry
ARNOLD, BALAKRISHNAN, and NAGARAJA · Records
* ARTHANARI and DODGE · Mathematical Programming in Statistics
* BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BAJORSKI · Statistics for Imaging, Optics, and Photonics
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications
BALAKRISHNAN and NG · Precedence-Type Tests and Applications
BARNETT · Comparative Statistical Inference, *Third Edition*
BARNETT · Environmental Statistics
BARNETT and LEWIS · Outliers in Statistical Data, *Third Edition*
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference
BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications
BASU and RIGDON · Statistical Methods for the Reliability of Repairable Systems
BATES and WATTS · Nonlinear Regression Analysis and Its Applications
BECHHOFER, SANTNER, and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons
BEIRLANT, GOEGEBEUR, SEGERS, TEUGELS, and DE WAAL · Statistics of Extremes: Theory and Applications

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression
- † BELSLEY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
- BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, *Fourth Edition*
- BERNARDO and SMITH · Bayesian Theory
- BERRY, CHALONER, and GEWEKE · Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner
- BHAT and MILLER · Elements of Applied Stochastic Processes, *Third Edition*
- BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications
- BIEMER, GROVES, LYBERG, MATHIOWETZ, and SUDMAN · Measurement Errors in Surveys
- BILLINGSLEY · Convergence of Probability Measures, *Second Edition*
- BILLINGSLEY · Probability and Measure, *Third Edition*
- BIRKES and DODGE · Alternative Methods of Regression
- BISGAARD and KULAHCI · Time Series Analysis and Forecasting by Example
- BISWAS, DATTA, FINE, and SEGAL · Statistical Advances in the Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics
- BLISCHKE AND MURTHY (editors) · Case Studies in Reliability and Maintenance
- BLISCHKE AND MURTHY · Reliability: Modeling, Prediction, and Optimization
- BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, *Second Edition*
- BOLLEN · Structural Equations with Latent Variables
- BOLLEN and CURRAN · Latent Curve Models: A Structural Equation Perspective
- BOROVKOV · Ergodicity and Stability of Stochastic Processes
- BOSQ and BLANKE · Inference and Prediction in Large Dimensions
- BOULEAU · Numerical Methods for Stochastic Processes
- BOX · Bayesian Inference in Statistical Analysis
- BOX · Improving Almost Anything, *Revised Edition*
- BOX · R. A. Fisher, the Life of a Scientist
- BOX and DRAPER · Empirical Model-Building and Response Surfaces
- * BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement
- BOX and DRAPER · Response Surfaces, Mixtures, and Ridge Analyses, *Second Edition*
- BOX, HUNTER, and HUNTER · Statistics for Experimenters: Design, Innovation, and Discovery, *Second Edition*
- BOX, JENKINS, and REINSEL · Time Series Analysis: Forecasting and Control, *Fourth Edition*
- BOX, LUCEÑO, and PANIAGUA-QUIÑONES · Statistical Control by Monitoring and Adjustment, *Second Edition*
- BRANDIMARTE · Numerical Methods in Finance: A MATLAB-Based Introduction
- † BROWN and HOLLANDER · Statistics: A Biomedical Introduction
- BRUNNER, DOMHOF, and LANGER · Nonparametric Analysis of Longitudinal Data in Factorial Experiments
- BUCKLEW · Large Deviation Techniques in Decision, Simulation, and Estimation
- CAIROLI and DALANG · Sequential Stochastic Optimization
- CASTILLO, HADI, BALAKRISHNAN, and SARABIA · Extreme Value and Related Models with Applications in Engineering and Science
- CHAN · Time Series: Applications to Finance with R and S-Plus®, *Second Edition*
- CHARALAMBIDES · Combinatorial Methods in Discrete Distributions
- CHATTERJEE and HADI · Regression Analysis by Example, *Fourth Edition*
- CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression
- CHERNICK · Bootstrap Methods: A Guide for Practitioners and Researchers, *Second Edition*
- CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- CHILÈS and DELFINER · Geostatistics: Modeling Spatial Uncertainty
- CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies, *Second Edition*
- CLARKE · Linear Models: The Theory and Application of Analysis of Variance
- CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, *Second Edition*
- * COCHRAN and COX · Experimental Designs, *Second Edition*
- COLLINS and LANZA · Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences
- CONGDON · Applied Bayesian Modelling
- CONGDON · Bayesian Models for Categorical Data
- CONGDON · Bayesian Statistical Modelling, *Second Edition*
- CONOVER · Practical Nonparametric Statistics, *Third Edition*
- COOK · Regression Graphics
- COOK and WEISBERG · An Introduction to Regression Graphics
- COOK and WEISBERG · Applied Regression Including Computing and Graphics
- CORNELL · A Primer on Experiments with Mixtures
- CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Third Edition*
- COVER and THOMAS · Elements of Information Theory
- COX · A Handbook of Introductory Statistical Methods
- * COX · Planning of Experiments
- CRESSIE · Statistics for Spatial Data, *Revised Edition*
- CRESSIE and WIKLE · Statistics for Spatio-Temporal Data
- CSÖRGŐ and HORVÁTH · Limit Theorems in Change Point Analysis
- DANIEL · Applications of Statistics to Industrial Experimentation
- DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Eighth Edition*
- * DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*
- DASU and JOHNSON · Exploratory Data Mining and Data Cleaning
- DAVID and NAGARAJA · Order Statistics, *Third Edition*
- * DEGROOT, FIENBERG, and KADANE · Statistics and the Law
- DEL CASTILLO · Statistical Process Adjustment for Quality Control
- DEMARIS · Regression with Social Data: Modeling Continuous and Limited Response Variables
- DEMIDENKO · Mixed Models: Theory and Applications
- DENISON, HOLMES, MALLICK and SMITH · Bayesian Methods for Nonlinear Classification and Regression
- DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis
- DEY and MUKERJEE · Fractional Factorial Plans
- DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
- DODGE · Alternative Methods of Regression
- * DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*
- * DOOB · Stochastic Processes
- DOWDY, WEARDEN, and CHILKO · Statistics for Research, *Third Edition*
- DRAPER and SMITH · Applied Regression Analysis, *Third Edition*
- DRYDEN and MARDIA · Statistical Shape Analysis
- DUDEWICZ and MISHRA · Modern Mathematical Statistics
- DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, *Third Edition*
- DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- EDLER and KITSOS · Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment
- * ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
ENDERS · Applied Econometric Time Series
- † ETHIER and KURTZ · Markov Processes: Characterization and Convergence
EVANS, HASTINGS, and PEACOCK · Statistical Distributions, *Third Edition*
EVERITT · Cluster Analysis, *Fifth Edition*
- FELLER · An Introduction to Probability Theory and Its Applications, Volume I, *Third Edition*, Revised; Volume II, *Second Edition*
- FISHER and VAN BELLE · Biostatistics: A Methodology for the Health Sciences
- FITZMAURICE, LAIRD, and WARE · Applied Longitudinal Analysis, *Second Edition*
- * FLEISS · The Design and Analysis of Clinical Experiments
FLEISS · Statistical Methods for Rates and Proportions, *Third Edition*
- † FLEMING and HARRINGTON · Counting Processes and Survival Analysis
- FUJIKOSHI, ULYANOV, and SHIMIZU · Multivariate Statistics: High-Dimensional and Large-Sample Approximations
- FULLER · Introduction to Statistical Time Series, *Second Edition*
- † FULLER · Measurement Error Models
GALLANT · Nonlinear Statistical Models
- GEISSER · Modes of Parametric Statistical Inference
- GELMAN and MENG · Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives
- GEWEKE · Contemporary Bayesian Econometrics and Statistics
- GHOSH, MUKHOPADHYAY, and SEN · Sequential Estimation
- GIESBRECHT and GUMPERTZ · Planning, Construction, and Statistical Analysis of Comparative Experiments
- GIFI · Nonlinear Multivariate Analysis
- GIVENS and HOETING · Computational Statistics
- GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems
- GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*
- GOLDSTEIN · Multilevel Statistical Models, *Fourth Edition*
- GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues
- GOLDSTEIN and WOOFF · Bayes Linear Statistics
- GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing
- GROSS, SHORTLE, THOMPSON, and HARRIS · Fundamentals of Queueing Theory, *Fourth Edition*
- GROSS, SHORTLE, THOMPSON, and HARRIS · Solutions Manual to Accompany Fundamentals of Queueing Theory, *Fourth Edition*
- * HAHN and SHAPIRO · Statistical Models in Engineering
HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners
- HALD · A History of Probability and Statistics and their Applications Before 1750
HALD · A History of Mathematical Statistics from 1750 to 1930
- † HAMPEL · Robust Statistics: The Approach Based on Influence Functions
HANNAN and DEISTLER · The Statistical Theory of Linear Systems
- HARMAN and KULKARNI · An Elementary Introduction to Statistical Learning Theory
- HARTUNG, KNAPP, and SINHA · Statistical Meta-Analysis with Applications
- HEIBERGER · Computation for the Analysis of Designed Experiments
- HEDAYAT and SINHA · Design and Inference in Finite Population Sampling
- HEDEKER and GIBBONS · Longitudinal Data Analysis
- HELLER · MACSYMA for Statisticians
- HERITIER, CANTONI, COPT, and VICTORIA-FESER · Robust Methods in Biostatistics

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 1:
 Introduction to Experimental Design, *Second Edition*
- HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 2:
 Advanced Experimental Design
- HINKELMANN (editor) · Design and Analysis of Experiments, Volume 3: Special
 Designs and Applications
- HOAGLIN, MOSTELLER, and TUKEY · Fundamentals of Exploratory Analysis
 of Variance
- * HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes
 - * HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory
 Data Analysis
- HOCHBERG and TAMHANE · Multiple Comparison Procedures
- HOCKING · Methods and Applications of Linear Models: Regression and the Analysis
 of Variance, *Second Edition*
- HOEL · Introduction to Mathematical Statistics, *Fifth Edition*
- HOGG and KLUGMAN · Loss Distributions
- HOLLANDER and WOLFE · Nonparametric Statistical Methods, *Second Edition*
- HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*
- HOSMER, LEMESHOW, and MAY · Applied Survival Analysis: Regression Modeling
 of Time-to-Event Data, *Second Edition*
- HUBER · Data Analysis: What Can Be Learned From the Past 50 Years
- HUBER · Robust Statistics
- † HUBER and RONCHETTI · Robust Statistics, *Second Edition*
- HUBERTY · Applied Discriminant Analysis, *Second Edition*
- HUBERTY and OLEJNIK · Applied MANOVA and Discriminant Analysis,
Second Edition
- HUITEMA · The Analysis of Covariance and Alternatives: Statistical Methods for
 Experiments, Quasi-Experiments, and Single-Case Studies, *Second Edition*
- HUNT and KENNEDY · Financial Derivatives in Theory and Practice, *Revised Edition*
- HURD and MIAMEE · Periodically Correlated Random Sequences: Spectral Theory
 and Practice
- HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek—
 with Commentary
- HUZURBAZAR · Flowgraph Models for Multistate Time-to-Event Data
- IMAN and CONOVER · A Modern Approach to Statistics
- JACKMAN · Bayesian Analysis for the Social Sciences
- † JACKSON · A User's Guide to Principle Components
- JOHN · Statistical Methods in Engineering and Quality Assurance
- JOHNSON · Multivariate Statistical Simulation
- JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics:
 A Volume in Honor of Samuel Kotz
- JOHNSON and BHATTACHARYYA · Statistics: Principles and Methods, *Fifth Edition*
- JOHNSON, KEMP, and KOTZ · Univariate Discrete Distributions, *Third Edition*
- JOHNSON and KOTZ · Distributions in Statistics
- JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the
 Seventeenth Century to the Present
- JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions,
 Volume 1, *Second Edition*
- JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions,
 Volume 2, *Second Edition*
- JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions
- JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE · The Theory and Practice of
 Econometrics, *Second Edition*
- JUREČ KOVÁ and SEN · Robust Statistical Procedures: Asymptotics and Interrelations

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- JUREK and MASON · Operator-Limit Distributions in Probability Theory
- KADANE · Bayesian Methods and Ethics in a Clinical Trial Design
- KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence
- KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, *Second Edition*
- KARIYA and KURATA · Generalized Least Squares
- KASS and VOS · Geometrical Foundations of Asymptotic Inference
- † KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis
- KEDEM and FOKIANOS · Regression Models for Time Series Analysis
- KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory
- KHURI · Advanced Calculus with Applications in Statistics, *Second Edition*
- KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models
- * KISH · Statistical Design for Research
- KLEIBER and KOTZ · Statistical Size Distributions in Economics and Actuarial Sciences
- KLEMELÄ · Smoothing of Multivariate Data: Density Estimation and Visualization
- KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions, *Third Edition*
- KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions, *Third Edition*
- KOSKI and NOBLE · Bayesian Networks: An Introduction
- KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions, Volume 1, *Second Edition*
- KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index
- KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Supplement Volume
- KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 1
- KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 2
- KOVALENKO, KUZNETZOV, and PEGG · Mathematical Theory of Reliability of Time-Dependent Systems with Practical Applications
- KOWALSKI and TU · Modern Applied U-Statistics
- KRISHNAMOORTHY and MATHEW · Statistical Tolerance Regions: Theory, Applications, and Computation
- KROESE, TAIMRE, and BOTEV · Handbook of Monte Carlo Methods
- KROONENBERG · Applied Multiway Data Analysis
- KULINSKAYA, MORGENTHALER, and STAUDTE · Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence
- KUROWICKA and COOKE · Uncertainty Analysis with High Dimensional Dependence Modelling
- KVAM and VIDAKOVIC · Nonparametric Statistics with Applications to Science and Engineering
- LACHIN · Biostatistical Methods: The Assessment of Relative Risks, *Second Edition*
- LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction
- LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*
- LANGE, RYAN, BILLARD, BRILLINGER, CONQUEST, and GREENHOUSE · Case Studies in Biometry
- LARSON · Introduction to Probability Theory and Statistical Inference, *Third Edition*
- LAWLESS · Statistical Models and Methods for Lifetime Data, *Second Edition*
- LAWSON · Statistical Methods in Spatial Epidemiology, *Second Edition*
- LE · Applied Categorical Data Analysis

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- LE · Applied Survival Analysis
 LEE · Structural Equation Modeling: A Bayesian Approach
 LEE and WANG · Statistical Methods for Survival Data Analysis, *Third Edition*
 LEPAGE and BILLARD · Exploring the Limits of Bootstrap
 LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics
 LIAO · Statistical Group Comparison
 LINDVALL · Lectures on the Coupling Method
 LIN · Introductory Stochastic Analysis for Finance and Insurance
 LINHART and ZUCCHINI · Model Selection
 LITTLE and RUBIN · Statistical Analysis with Missing Data, *Second Edition*
 LLOYD · The Statistical Analysis of Categorical Data
 LOWEN and TEICH · Fractal-Based Point Processes
 MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in
 Statistics and Econometrics, *Revised Edition*
 MALLER and ZHOU · Survival Analysis with Long Term Survivors
 MALLOWS · Design, Data, and Analysis by Some Friends of Cuthbert Daniel
 MANN, SCHAFER, and SINGPURWALLA · Methods for Statistical Analysis of
 Reliability and Life Data
 MANTON, WOODBURY, and TOLLEY · Statistical Applications Using Fuzzy Sets
 MARCHETTE · Random Graphs for Statistical Pattern Recognition
 MARDIA and JUPP · Directional Statistics
 MARKOVICH · Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and
 Practice
 MARONNA, MARTIN and YOHAI · Robust Statistics: Theory and Methods
 MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with
 Applications to Engineering and Science, *Second Edition*
 McCULLOCH, SEARLE, and NEUHAUS · Generalized, Linear, and Mixed Models,
 Second Edition
 McFADDEN · Management of Data in Clinical Trials, *Second Edition*
 * McLACHLAN · Discriminant Analysis and Statistical Pattern Recognition
 McLACHLAN, DO, and AMBROISE · Analyzing Microarray Gene Expression Data
 McLACHLAN and KRISHNAN · The EM Algorithm and Extensions, *Second Edition*
 McLACHLAN and PEEL · Finite Mixture Models
 McNEIL · Epidemiological Research Methods
 MEEKER and ESCOBAR · Statistical Methods for Reliability Data
 MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent
 Random Vectors: Heavy Tails in Theory and Practice
 MENGERSEN, ROBERT, and TITTERINGTON · Mixtures: Estimation and
 Applications
 MICKEY, DUNN, and CLARK · Applied Statistics: Analysis of Variance and
 Regression, *Third Edition*
 * MILLER · Survival Analysis, *Second Edition*
 MONTGOMERY, JENNINGS, and KULAHCI · Introduction to Time Series Analysis
 and Forecasting
 MONTGOMERY, PECK, and VINING · Introduction to Linear Regression Analysis,
 Fourth Edition
 MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical
 Robustness
 MUIRHEAD · Aspects of Multivariate Statistical Theory
 MULLER and STOYAN · Comparison Methods for Stochastic Models and Risks
 MURRAY · X-STAT 2.0 Statistical Experimentation, Design Data Analysis, and
 Nonlinear Optimization
 MURTHY, XIE, and JIANG · Weibull Models

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

- MYERS, MONTGOMERY, and ANDERSON-COOK · Response Surface Methodology:
 Process and Product Optimization Using Designed Experiments, *Third Edition*
- MYERS, MONTGOMERY, Vining, and ROBINSON · Generalized Linear Models.
 With Applications in Engineering and the Sciences, *Second Edition*
- † NELSON · Accelerated Testing, Statistical Models, Test Plans, and Data Analyses
- † NELSON · Applied Life Data Analysis
- NEWMAN · Biostatistical Methods in Epidemiology
- OCHI · Applied Probability and Stochastic Processes in Engineering and Physical Sciences
- OKABE, BOOTS, SUGIHARA, and CHIU · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, *Second Edition*
- OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis
- PALTA · Quantitative Methods in Population Health: Extensions of Ordinary Regressions
- PANJER · Operational Risk: Modeling and Analytics
- PANKRATZ · Forecasting with Dynamic Regression Models
- PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
- PARDOUX · Markov Processes and Applications: Algorithms, Networks, Genome and Finance
- * PARZEN · Modern Probability Theory and Its Applications
- PEÑA, TIAO, and TSAY · A Course in Time Series Analysis
- PIANTADOSI · Clinical Trials: A Methodologic Perspective
- PORT · Theoretical Probability for Applications
- POURAHMADI · Foundations of Time Series Analysis and Prediction Theory
- POWELL · Approximate Dynamic Programming: Solving the Curses of Dimensionality, *Second Edition*
- PRESS · Bayesian Statistics: Principles, Models, and Applications
- PRESS · Subjective and Objective Bayesian Statistics, *Second Edition*
- PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach
- PUKELSHEIM · Optimal Experimental Design
- PURI, VILAPLANA, and WERTZ · New Perspectives in Theoretical and Applied Statistics
- † PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming
- QIU · Image Processing and Jump Regression Analysis
- * RAO · Linear Statistical Inference and Its Applications, *Second Edition*
- RAO · Statistical Inference for Fractional Diffusion Processes
- RAUSAND and HØYLAND · System Reliability Theory: Models, Statistical Methods, and Applications, *Second Edition*
- RAYNER · Smooth Tests of Goodness of Fit: Using R, *Second Edition*
- RENCHER · Linear Models in Statistics
- RENCHER · Methods of Multivariate Analysis, *Second Edition*
- RENCHER · Multivariate Statistical Inference with Applications
- * RIPLEY · Spatial Statistics
- * RIPLEY · Stochastic Simulation
- ROBINSON · Practical Strategies for Experimenting
- ROHATGI and SALEH · An Introduction to Probability and Statistics, *Second Edition*
- ROLSKI, SCHMIDL, SCHMIDT, and TEUGELS · Stochastic Processes for Insurance and Finance
- ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice
- ROSS · Introduction to Probability and Statistics for Engineers and Scientists
- ROSSI, ALLENBY, and MCCULLOCH · Bayesian Statistics and Marketing
- † ROUSSEEUW and LEROY · Robust Regression and Outlier Detection
- ROYSTON and SAUERBREI · Multivariate Model Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modeling Continuous Variables

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- * RUBIN · Multiple Imputation for Nonresponse in Surveys
RUBINSTEIN and KROESE · Simulation and the Monte Carlo Method, *Second Edition*
RUBINSTEIN and MELAMED · Modern Simulation and Modeling
RYAN · Modern Engineering Statistics
RYAN · Modern Experimental Design
RYAN · Modern Regression Methods, *Second Edition*
RYAN · Statistical Methods for Quality Improvement, *Third Edition*
SALEH · Theory of Preliminary Test and Stein-Type Estimation with Applications
SALTELLI, CHAN, and SCOTT (editors) · Sensitivity Analysis
- * SCHEFFE · The Analysis of Variance
SCHIMEK · Smoothing and Regression: Approaches, Computation, and Application
SCHOTT · Matrix Analysis for Statistics, *Second Edition*
SCHOUTENS · Levy Processes in Finance: Pricing Financial Derivatives
SCHUSS · Theory and Applications of Stochastic Differential Equations
SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization
- * SEARLE · Linear Models
† SEARLE · Linear Models for Unbalanced Data
† SEARLE · Matrix Algebra Useful for Statistics
† SEARLE, CASELLA, and McCULLOCH · Variance Components
SEARLE and WILLETT · Matrix Algebra for Applied Economics
SEBER · A Matrix Handbook For Statisticians
† SEBER · Multivariate Observations
SEBER and LEE · Linear Regression Analysis, *Second Edition*
† SEBER and WILD · Nonlinear Regression
SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems
- * SERFLING · Approximation Theorems of Mathematical Statistics
SHAFER and VOVK · Probability and Finance: It's Only a Game!
SHERMAN · Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties
SILVAPULLE and SEN · Constrained Statistical Inference: Inequality, Order, and Shape Restrictions
SINGPURWALLA · Reliability and Risk: A Bayesian Perspective
SMALL and MCLEISH · Hilbert Space Methods in Probability and Statistical Inference
SRIVASTAVA · Methods of Multivariate Statistics
STAPLETON · Linear Statistical Models, *Second Edition*
STAPLETON · Models for Probability and Statistical Inference: Theory and Applications
STAUDTE and SHEATHER · Robust Estimation and Testing
STOYAN, KENDALL, and MECKE · Stochastic Geometry and Its Applications, *Second Edition*
STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics
STREET and BURGESS · The Construction of Optimal Stated Choice Experiments: Theory and Methods
STYAN · The Collected Papers of T. W. Anderson: 1943–1985
SUTTON, ABRAMS, JONES, SHELDON, and SONG · Methods for Meta-Analysis in Medical Research
TAKEZAWA · Introduction to Nonparametric Regression
TAMHANE · Statistical Analysis of Designed Experiments: Theory and Applications
TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
THOMPSON · Empirical Model Building: Data, Models, and Reality, *Second Edition*
THOMPSON · Sampling, *Third Edition*
THOMPSON · Simulation: A Modeler's Approach
THOMPSON and SEBER · Adaptive Sampling
THOMPSON, WILLIAMS, and FINDLAY · Models for Investors in Real World Markets

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.

- TIAO, BISGAARD, HILL, PEÑA, and STIGLER (editors) · Box on Quality and Discovery: with Design, Control, and Robustness
- TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
- TSAY · Analysis of Financial Time Series, *Third Edition*
- UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
- † VAN BELLE · Statistical Rules of Thumb, *Second Edition*
- VAN BELLE, FISHER, HEAGERTY, and LUMLEY · Biostatistics: A Methodology for the Health Sciences, *Second Edition*
- VESTRUP · The Theory of Measures and Integration
- VIDAKOVIC · Statistical Modeling by Wavelets
- VINOD and REAGLE · Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments
- WALLER and GOTWAY · Applied Spatial Statistics for Public Health Data
- WEERAHANDI · Generalized Inference in Repeated Measures: Exact Methods in MANOVA and Mixed Models
- WEISBERG · Applied Linear Regression, *Third Edition*
- WEISBERG · Bias and Causation: Models and Judgment for Valid Comparisons
- WELSH · Aspects of Statistical Inference
- WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment
- * WHITTAKER · Graphical Models in Applied Multivariate Statistics
- WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting
- WONNACOTT and WONNACOTT · Econometrics, *Second Edition*
- WOODING · Planning Pharmaceutical Clinical Trials: Basic Statistical Principles
- WOODWORTH · Biostatistics: A Bayesian Introduction
- WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data, *Second Edition*
- WU and HAMADA · Experiments: Planning, Analysis, and Parameter Design Optimization, *Second Edition*
- WU and ZHANG · Nonparametric Regression Methods for Longitudinal Data Analysis
- YANG · The Construction Theory of Denumerable Markov Processes
- YOUNG, VALERO-MORA, and FRIENDLY · Visual Statistics: Seeing Data with Dynamic Interactive Graphics
- ZACKS · Stage-Wise Adaptive Designs
- * ZELLNER · An Introduction to Bayesian Inference in Econometrics
- ZELTERMAN · Discrete Distributions—Applications in the Health Sciences
- ZHOU, OBUCHOWSKI, and MCCLISH · Statistical Methods in Diagnostic Medicine, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley-Interscience Paperback Series.