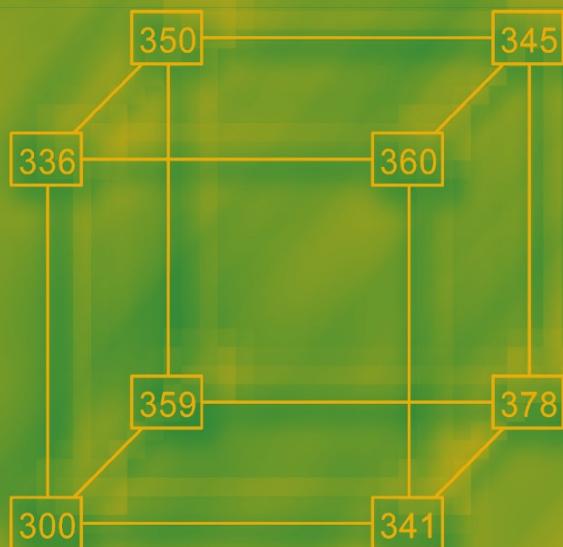


ROBERT W. MEE

A COMPREHENSIVE GUIDE TO

Factorial Two-Level Experimentation



Springer

A Comprehensive Guide to Factorial Two-Level Experimentation

This document provides a detailed guide to factorial two-level experimentation, covering the design, analysis, and interpretation of such experiments.

The guide is organized into several sections:

- Introduction to Factorial Two-Level Experiments

- Designing Factorial Two-Level Experiments

- Analysis of Factorial Two-Level Experiments

- Interpreting Results from Factorial Two-Level Experiments

- Best Practices for Conducting Factorial Two-Level Experiments

- Conclusion and Summary

The guide includes numerous examples, case studies, and exercises to help readers apply the concepts learned in the guide.

For more information, please refer to the references and further reading section at the end of the guide.

We hope you find this guide useful for your research needs.

If you have any questions or comments, please feel free to contact us.

Thank you for reading this guide to Factorial Two-Level Experimentation.

Best regards,

The Team Behind This Guide

Robert W. Mee

A Comprehensive Guide to Factorial Two-Level Experimentation



Springer

Robert W. Mee
Department of Statistics, Operations,
and Management Science
The University of Tennessee
333 Stokely Management Center
Knoxville, TN 37996-0532
USA

ISBN 978-0-387-89102-6 e-ISBN 978-0-387-89103-3
DOI 10.1007/b105081
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009927712

© Springer Science+Business Media, LLC 2009

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

ASTM® is a registered trademark of ASTM International. AT&T® is a registered trademark of AT&T in the United States and other countries. Baskin-Robbins® is a registered trademark of BR IP Holder LLC. JMP® and SAS® are registered trademarks of the SAS Institute, Inc. Kapton® is a registered trademark of DuPont. Minitab® is a registered trademark of Minitab, Inc. Design-Expert® is a registered trademark of Stat-Ease, Inc. MATLAB® is a registered trademark of The MathWorks, Inc. Web of Science® is a registered trademark of Thompson Reuters. The Isaiah quotation is from The Holy Bible, English Standard Version, copyright ©2001 by Crossway Bibles, a division of Good News Publishers. Used by permission. All rights reserved.

Dedication

To Cherol, my lovely wife and helpmate,
and our children Joseph, Elizabeth and Evangeline

To David Harville, thanks for the “license to learn”

“All these things my hand has made, and so all these things came to be,
declares the Lord. But this is the one to whom I will look: he who is humble
and contrite in spirit and trembles at my word.” — Isaiah 66.2

Is the universe simply made up of numbers?
Our eyes tell us everything is countable,
action versus reaction,
a swinging pendulum.
And yet a tiny voice in each of us
speaks to the opposite –
that every blade of grass
only masks the Uncountable,
that every ray of light
only clothes the Infinite.

— Joseph Mee

Nil sine Deo

Being a student and professor is a great privilege. For 33 years I have learned about designed experiments from teachers, colleagues, and students alike. Beginning with Russell Heikes and Doug Montgomery at Georgia Tech, and David Harville and Oscar Kempthorne at Iowa State, I gained a solid foundation in the fascinating field of design and analysis of experiments. Both Dennis Lin and I came to the University of Tennessee (UT) in 1989, and Dennis's enthusiasm for research with two-level factorial designs helped reawaken my interest in the field. UT offered a 3-week Design of Experiments Institute at that time, and I gained a new appreciation for how practitioners learn to apply these tools from Richard Sanders, Bob McLean, and Mary Leitnaker. Tony Cooper, Doug Sanders, and Mike Urie challenged me with many questions as students in a master's-level design course I taught my first semester at UT; having such inquisitive students was a great incentive to understand more deeply. After Ramón León arrived at Tennessee, he included me in opportunities to see how AT&T® presented robust parameter design methods to industry. At Ramón's initiative, we prepared and taught our own version of robust parameter design for an American Statistical Association workshop; that was a great learning experience. Other prominent statisticians extended invitations to me. Especially influential were Norman Draper's invitation to a conference at Irsee, Germany and Jeff Wu's recommendations to include me on early Spring Research Conference programs. Through the DAE (and other) conferences I have benefitted from contacts with so many colleagues that I can only name a few who have been the most influential: Hegang Chen, Ching-Shui Cheng, Shao-Wei Cheng, Sam Hedayat, Brad Jones, Max Morris, Bill Notz, Eric Schoen, David Steinberg, Boxin Tang, Dan Voss, and Hongquan Xu. Finally, many UT students have worked on projects involving two-level designs. Foremost of these have been Ph.D. students Robert Block, David Edwards, and Oksoun Yee. In addition, Rodney Bates, Anne Freeman, Dawn Heaney, Chris Krohn, Huo Li, Chris McCall, Marta Peralta, Anna Romanova, Yeak-Chong Wong, Regina Xiao, Ning Xue, Philip Yates, and Janna Young all completed M.S. independent study projects involving two-level designs that contributed to the results of this book. To this multitude of students, colleagues, and teachers, I say "Thanks."

A final thanks is due to the researchers and statisticians who provided me experiment details beyond that documented in their publications. This list includes George Bandurek, Mark Bond, Dennis Boos, Laurent Broudiscou, Hisham Choueiki and Clark Mount-Campbell, Steve Gilmour, Doug Hawkins and Joshua Yi, Eric Hoàng, Adam Kramschuster, Aik Chong Lua, Cristina Martin, Eric Schoen, Wen Yang, and Stan Young.

Preface

Two-level factorial designs fascinated me when, as a senior at Georgia Tech, I was introduced to their clever structure and utility. Courses in design of experiments and response surface methodology persuaded me to pursue a career in Statistics. One year later, the eminently successful book *Statistics for Experimenters* by Box, Hunter, and Hunter (BHH) (1978) was published. That book, more than any other, has enabled scientists and engineers to employ these useful designs. I recall loaning my copy of BHH to an engineering graduate student years ago, to introduce him to fractional factorial designs. To my surprise, it was the simpler 2^k full factorial designs that captured this student's interest. I had incorrectly assumed that he and other engineers would already be familiar with full factorial experiments. But that was not the case; the notion of experimenting with many factors simultaneously was completely new. Indeed, such an idea was truly novel in the 1920s, when Sir Ronald A. Fisher, the father of experimental design, wrote:

No aphorism is more frequently repeated in connection with field trials, than that we must ask Nature few questions, or, ideally, one question, at a time. The writer is convinced that this view is wholly mistaken. Nature, he suggests, will best respond to a logical and carefully thought out questionnaire; indeed, if we ask her a single question, she will often refuse to answer until some other topic has been discussed.
(Fisher, 1926)

Two-level factorial and fractional factorial designs, Plackett–Burman designs, and two-level orthogonal arrays are now widely used. A search on Web of Science® in December 2008 yielded over 7000 articles mentioning factorial designs and nearly 500 more mentioning Plackett–Burman designs. While many of these factorial design applications involve factors with more than two levels, two-level factorial designs are the most common and are the easiest to understand and analyze. Thus, while this book will be an introduction to 2^k full factorial designs for some, its primary objectives go beyond an introduction. First, the purpose of this book is to help practitioners design and

analyze two-level factorial designs correctly. As I reviewed published examples, I routinely found mistakes and misunderstandings, especially in the analysis. This book will help nonstatisticians plan and analyze factorial experiments correctly. The following chapters contain 50 analyses of actual data from two-level designs. By carefully studying these examples, how to properly analyze one's own data will become clear. In the past, I thought intelligent software could automatically analyze the data. While it is true that statistical software packages such as JMP[®], Design-Expert[®], and Minitab[®] have made incredible strides in the last 10 years to facilitate the analysis of these designs, there are many details that distinguish one application from the next and necessitate subtle changes to the analysis. Nothing will replace the requirement for an experienced user. The numerous analyses documented in this book are intended to help build the needed expertise.

Beyond exposure to factorial designs and the knowledge to perform an analysis correctly, this book has the further objective of making new developments accessible to practitioners. Over the last 30 years, the statistical literature regarding two-level factorial designs has exploded. General design of experiment books cannot cover such growth in the literature. My goal in writing this more focused book has been to sift through the hundreds of recent articles with new theory and methods, to decide what is most useful, and then to summarize and illustrate that useful material.

This book's comprehensiveness is unique. As a reference book, it will benefit both practitioners and statisticians. To aid the reader, the book is divided into three parts. For those with little or no exposure to factorial experimentation, Part I: Full Factorial Designs is the most relevant material. Chapter 1 introduces the reader to the advantages of factorial experiments, presents the basic regression models that become the foundation for the analysis, and concludes with a four-step strategy for planning these experiments. Chapter 2 serves as a manual for data analysis. Chapter 3 concerns further design details, to improve either the precision or convenience of the experiment. Part I concludes with Chapter 4's extended analysis of three examples. In total, 15 full factorial experiments are analyzed in Part I.

Part II is intended for readers who are familiar with factorial designs and encounter applications with a large number of factors—although Chapter 2's analysis tips and Chapter 3's explanation of blocking structures should not be skipped. The seven chapters of Part II all deal with fractional factorial designs. The simplest of these, regular fractional factorial designs, are introduced in Chapter 5. Following this introduction, Chapters 6–8 present both regular fractional factorial designs and the orthogonal array designs based on Hadamard matrices. Chapter 6 presents the most frugal designs in terms of run size, including designs popularized by Plackett and Burman (1946), where the number of factors is nearly as large as the number of runs. Section 6.5 even contemplates attempts to use designs with more factors than runs. Chapter 7 presents fractional factorial designs that are somewhat less risky in their assumptions, where the number of runs is at least twice the number

of factors. Chapter 8 discusses designs that are large enough to estimate the two-factor interaction model [defined by (1.3) in Section 1.2]. Since the fractional factorial designs of Part II require assumptions to interpret the data, Chapter 9 details how one may follow these designs with additional runs either to confirm or to clarify the results. Akin to the last two chapters of Part I, Chapter 10 describes how to run and analyze fractional factorial experiments with blocking restrictions, and Chapter 11 presents detailed analysis for four more examples.

As comprehensive as Parts I and II are for full factorial and fractional factorial designs, some details were deferred, as they were judged to be either tangential to the main thrust or because their need is more specialized. Part III contains this deferred material. It begins with Chapter 12's brief introduction to designs for fitting second-order models, complete with quadratic terms for each factor. Two-level designs do not support estimation of such models, but two-level designs can easily be augmented to do so. Such topics are the domain of a field called *response surface methodology*; for a thorough treatment, the reader is directed to other books. Finally, Chapter 13 covers specialized topics related to the design choice, and Chapter 14 discusses matters of analysis. Practical questions not covered earlier are addressed here, such as how wide to space the levels of a quantitative factor and how to sample within runs to study variation. The book concludes with numerous tables needed for design construction and analysis.

The book's final section is Section 14.7: Four Analysis Blunders to Avoid. But more common and more serious than mistakes in the analysis is the failing to experiment at all, or failing to plan experiments well. Discussing Coleman and Montgomery's (1993) "A Systematic Approach to Planning for a Designed Industrial Experiment," Bert Gunter (1993) writes:

(R)ational experimental planning leads inevitably to the recognition that ALL experiments are designed experiments; the only question is whether well or poorly. The choice is therefore not whether or not statistical methods are used but whether or not sensible planning is done. Sensible planning will almost always result in the application of statistical design. Failure to do such planning will almost always lead to wasted expenditures and poor results... (T)he absence of statistical design in industry is the consequence of sloppy or absent experimental planning, not just ignorance of the methods.

Successful experimentation is hard work. This book will not remove any of the hard work. But I hope that this record of others' successes and missteps will encourage many readers to take up the challenge. May your efforts be well rewarded.

As mentioned earlier, thoroughly understanding proper statistical analysis of data requires practice. To that end, all data for examples presented in this book are available online at <http://soms.utk.edu/mee>. Repeating or extending the analyses offered here is a certain means of developing your expertise

in gaining insight through data. If, as you work through the examples, you have questions or wish to point out a suspected error, please send e-mail to rmee@utk.edu. Tips on using JMP and R software will be added to the website in response to readers' questions. As errors or ambiguities in the book are found, clarifications and corrections will also be posted.

Robert Mee
Knoxville

Contents

Preface	ix
List of Examples Analyzed	xvii

Part I Full Factorial Designs

1 Introduction to Full Factorial Designs with Two-Level Factors	3
1.1 Advantages of Full Factorial Designs	3
1.2 Standard Regression Models for Factorial Designs with Two-Level Factors	9
1.3 Least Squares Estimation of a Regression Model	12
1.4 Presenting a Fitted Model Graphically	19
1.5 Four Steps for Planning a Successful Experiment	23
2 Analysis of Full Factorial Experiments	27
2.1 Analysis Strategy Overview	27
2.2 Analysis of Numerical Responses with Replication	29
2.3 The Inclusion of Centerpoint Replicates	31
2.4 Analysis of Numerical Responses Without Replication	35
2.5 Normal Plot of Effects and Other Analysis Tools	45
2.6 Diagnostics for a Fitted Model	48
2.7 Transformations of the Response	53
2.8 Analysis of Counts, Variances, and Other Statistics	57
2.9 Unequal Replication and Unequal Variance	67
2.10 The Impact of Missing Treatment Combinations	70
3 Common Randomization Restrictions	75
3.1 Sources of Variation and a Design's Unit Structure	75
3.2 Treatment*Unit Interactions	78

3.3	Blocking: Partitioning a Factorial into Smaller Experiments	79
3.4	Analyzing Randomized Block Factorial Designs	86
3.5	Split-Unit Designs	97
3.6	Multiway Blocking.	104
4	More Full Factorial Design Examples	115
4.1	Example 4.1: Replicated 2^3 With Subsampling Within Runs	115
4.2	Example 4.2: 2^9 Factorial for Peptide Research	122
4.3	Example 4.3: 2^5 with Centerpoint Runs for Ceramic Strength .	134

Part II Fractional Factorial Designs

5	Fractional Factorial Designs: The Basics	145
5.1	Initial Fractional Factorial Example.	145
5.2	Introduction to Regular Fractional Factorial Designs	150
5.3	Basic Analysis for Regular Fractional Factorial Designs.	161
6	Fractional Factorial Designs for Estimating Main Effects	173
6.1	Analysis of Regular Resolution III Fractional Factorial Designs	174
6.2	Some Theory Regarding Resolution III Designs.	190
6.3	Nonregular Orthogonal Designs of Strength 2	194
6.4	Optimal Nonorthogonal Saturated Main Effect Designs	226
6.5	Supersaturated Designs	231
6.6	Conclusions	244
7	Designs for Estimating Main Effects and Some Two-Factor Interactions	245
7.1	Five Examples Analyzed	246
7.2	Regular Resolution IV Designs	267
7.3	Strength-3 Orthogonal Arrays.	277
7.4	Nonorthogonal Resolution IV Designs	279
7.5	Summary Regarding Choice of a Design	282
8	Resolution V Fractional Factorial Designs	283
8.1	Regular Resolution V 2^{k-f} Fractional Factorial Designs	283
8.2	Strength-4 Orthogonal Arrays.	285
8.3	Three-Quarter Fraction of Regular Resolution V Designs	288
8.4	Smaller Nonorthogonal Resolution V Designs	293
8.5	Recommendations Regarding Design Choice	298
8.6	Analysis of Resolution V Experiments	299

9	Augmenting Fractional Factorial Designs	317
9.1	Follow-up Experimentation Choices	317
9.2	Confirmation Runs	319
9.3	Steepest Ascent Search	321
9.4	Foldover After a Resolution III Fraction	328
9.5	Foldover and Semifolding After a Resolution IV Fraction	332
9.6	Optimal Design Augmentation	338
9.7	Adding and Dropping Factors	342
10	Fractional Factorial Designs with Randomization Restrictions	343
10.1	Randomized Block Designs for Fractional Factorials	343
10.2	Split-Unit Designs for Fractional Factorials	350
10.3	Analysis of Fractional Factorials with Randomization Restrictions	354
10.4	Sequences of Fractional Factorial Designs	371
11	More Fractional Factorial Design Examples	375
11.1	A Mirror-Image Foldover with Unexpected Results	375
11.2	Steepest Ascent with Constraints	382
11.3	A Group Screening Experiment	385
11.4	Nonorthogonal Blocking for a Fractional Factorial	389

Part III Additional Topics

12	Response Surface Methods and Second-Order Designs	397
12.1	The Response Surface Methodology Strategy	397
12.2	Central Composite Designs	399
12.3	Other Composite Designs	403
12.4	Box–Behnken Designs	407
12.5	Analysis/Interpretation of the Fitted Second-Order Model	409
13	Special Topics Regarding the Design	415
13.1	Power and the Choice of Sample Size	415
13.2	Choice of Factor Levels	420
13.3	Tips for Studying Variation	422
13.4	Accommodating Factors with More Levels	424
13.5	Special Requirements for Run Order and Treatment Combinations	428
14	Special Topics Regarding the Analysis	437
14.1	Minimal Replication and Lenth’s Method	437
14.2	Alternatives to Lenth <i>t</i> -Tests for Unreplicated Designs	440
14.3	Analyzing the Variation in Structured Samples	449

14.4 Generalized Least Squares Analysis When Variances Are Unequal	454
14.5 Mixed-Model Analysis	457
14.6 Highly Multivariate Response Data	461
14.7 Four Analysis Blunders to Avoid	466

Part IV Appendices and Tables

A Upper Percentiles of t Distributions, t_α	471
B Upper Percentiles of F Distributions, F_α	473
C Upper Percentiles for Lenth t Statistics, c_α^{IER} and c_α^{EER}	477
D Computing Upper Percentiles for Maximum Studentized Residual	481
E Orthogonal Blocking for Full 2^k Factorial Designs	483
F Column Labels of Generators for Regular Fractional Factorial Designs	485
G Tables of Minimum Aberration Regular Fractional Factorial Designs	487
H Minimum Aberration Blocking Schemes for Fractional Factorial Designs	497
I Alias Matrix Derivation	511
J Distinguishing Among Fractional Factorial Designs	513
References	517
Abbreviations and Symbols	539
Index	543

List of Examples Analyzed

Section Application	Design Description
1.1 Container crush resistance	Replicated 2^3
1.1 Container crush resistance	2^3
2.3 Ceramic strength	2^5 with 7 center runs
2.4 Isatin yield	2^4
2.7 Drill advance rate	2^4
2.8 Grille blemishes	2^{9-5} projected into 2^4
3.4 Polyethylene film	2^5 in blocks
3.4 Organic chemical yield	Replicated 2^3 in blocks
3.4 Potato crop yield	Replicated 2^3 in blocks
3.4 Light bulb assembly	Replicated 2^2 in blocks
3.5 Plasma treated paper	Split unit 2^5
3.6 Milk production	Replicated 2^3 in Latin square
3.6 Meat loaf rancidity	Split-split-split-unit 2^4
4.1 Tablet weight	Replicated 2^3
4.2 Peptide research	2^9
4.3 Ceramic strength (revisited)	2^5 with 7 center runs
5.1 Electroplating	2^{5-2}
5.2 Catalytic oxidation	2^{5-2}
5.3 Polymeric coating	2^{8-3}
6.1 Chromatography	2^{5-2}
6.1 Lovastatin production	2^{7-4} and 2^{6-3}
6.1 Pulp quality	2^{13-9}
6.1 Inulinase production	2^{15-11}

Section	Application	Design Description
6.3	Thermostat testing	OA(12, 2 ¹¹ , 2)
6.3	Trace amount measurement	OA(12, 2 ⁷ , 2)
6.3	cDNA microarray processing	OA(20, 2 ¹⁹ , 2)
6.3	Credit card offers	OA(20, 2 ¹⁹ , 2)
6.5	AIDS incidence model	Supersaturated, $n = 24$, $k = 138$
7.1	Carbon membrane permeability	2 ^{4–1}
7.1	Melt index measure ruggedness	2 ^{6–2}
7.1	Semiconductor etching	2 ^{6–2}
7.1	Neural network training	2 ^{10–4}
7.1	Computer architecture	OA(88, 2 ⁴¹ , 3)
8.6	Protein chemistry (revisited)	2 ^{9–2}
8.6	Protein chemistry (revisited)	(3/4)2 ^{9–2}
8.6	Protein chemistry (revisited)	Irregular 2 ^{9–3}
8.6	Aromatic food additives	Rechtschaffner, $k = 7$
9.5	Semiconductor etching	Semifolding after 2 ^{6–2}
9.6	Injection molding	Augmenting a 2 ^{8–4}
10.3	Crankshaft quality	2 ^{9–4} in 4 blocks (or split-unit 2 ^{10–5})
10.3	Injection molding	2 ^{7–4} × 2 ^{4–1}
10.3	Cheese making	1/16 of 2 ⁹ 4 ¹ as split-split-unit
10.3	Washer/dryer settings	1/2 of 2 ^{6–3} × 2 ^{4–1} as two-way split unit
11.1	Marine bacterium growth	OA(20, 2 ¹⁷ , 2) with foldover
11.2	Chromatography	2 ^{3–1} and steepest ascent
11.3	Shipping efficiency	Group screening with $k = 29$ (2 ^{8–4} , followed by 2 ^{12–8})
11.4	Livestock feed	2 ^{14–10} and $n_0 = 2$, in 3 blocks
12.2	Paper copter flights	Central composite
12.3	Paper copter flights	Noncentral composite
14.3	Semiconductor deposition	2 ³ + 4 runs, sampling within runs
14.4	Rock fracture model	2 ³
14.6	Baguette quality	2 ^{5–1} with 15 response variables

Part I

Full Factorial Designs

Introduction to Full Factorial Designs with Two-Level Factors

Factorial experiments with two-level factors are used widely because they are easy to design, efficient to run, straightforward to analyze, and full of information. This chapter illustrates these benefits. The standard regression models for summarizing data from full factorial experiments are introduced, and an example is given to illustrate the interpretability and use of such models. Some statistical analysis is introduced here for the simplest case, although most analysis tools are deferred to the next chapter. The sections are as follows:

Section 1.1. Advantages of Full Factorial Designs

Section 1.2. Standard Regression Models for Factorial Designs with Two-Level Factors

Section 1.3. Least Squares Estimation of a Regression Model

Section 1.4. Presenting a Fitted Model Graphically

Section 1.5. Four Steps for Planning a Successful Experiment

1.1 Advantages of Full Factorial Designs

This book explains how to plan and analyze experiments with multiple *factors*. Experimental factors are inputs that are purposefully changed to study the resulting effects. Although many useful experiments involve a single variable or factor, most research questions are more complicated. Rather than asking simply, “How does increasing temperature affect strength?” we are interested in knowing how temperature, belt tension, cycle speed, and a host of other factors jointly affect the output. Initially, we consider experiments with just two to four factors. However, later tools and applications will involve experimentation with a dozen or more factors.

Experimental factors can be numerical variables, such as speed and temperature, or categorical, such as different suppliers of a material. Whether numerical (e.g., 350 degrees) or categorical (e.g., supplier A), we will refer to the values of these factors as *levels*. An experimental run involves a specified level for each factor; these combinations of levels (e.g., 350 degrees and supplier A) are commonly called *treatment combinations* (t.c.). This book focuses on experiments with two-level factors, since such experiments are widely used and relatively easy to analyze.

A full factorial experiment consists of every combination of the levels of factors in the experiment. Thus, if we have k factors, each at two levels, the full factorial consists of

$$\underbrace{2 \times 2 \times \cdots \times 2}_k = 2^k$$

treatment combinations. We use the symbol 2^k to represent this type of factorial design, not just as a calculation for the number of treatment combinations. One reason for the popularity of having only two levels per factor is that this enables the most economical investigation of many variables. For instance, with four factors, increasing the number of levels from 2 to 3 increases the size of the full factorial design from $2^4 = 16$ to $3^4 = 81$, 5 times larger.

Part I of this book (Chapters 1–4) addresses full factorial designs. If each of the 2^k treatment combinations is performed only once, then we have an unreplicated 2^k factorial design. If some or all of the treatment combinations are repeated in an experiment, we will refer to the design as partially or fully replicated. When the number of factors k is small, replication is common. However, if k is large, even performing a full 2^k factorial design may be both laborious and unnecessary. Part II of this book (Chapters 5–11) presents methods based for conducting fractional factorial designs—designs that contain only a subset of the 2^k treatment combinations. Sometimes two-level factorial designs lead to questions that can only be answered by increasing the number of levels per factor. Part III has some advice for what to do in such situations, as well as additional special topics. Since this book is intended for practitioners, Parts I and II each conclude with a chapter of case studies to reinforce the ideas and to illustrate further how to conduct an appropriate analysis.

1.1.1 An initial example

Consider now a full factorial design with three factors. Huhtamaki Americas is a specialty packaging organization offering food service products, consumer packaging, and packaging machinery. A facility in California manufactures large frozen dessert cartons for customers such as Baskin-Robbins®. An upgraded forming process for the carton was implemented, with start-up assistance from another facility. Initial settings for four process variables were as follows:

- Speed: 18 tubes per minute

- Score depth: “high,” for easy fold of sidewall, and tight fit with bottom
- Mandrel temperature: 90°F, which affects adhesive curing
- Belt tension: 65.

Six months after beginning shipping product from this new process, complaints of buckling containers were received from a single customer that fills the cartons with a very dense sherbet and vanilla yogurt product. Unlike other customers, this one does not flash-freeze cartons after filling, but instead stacks pallets with three tiers of cartons before transporting to the freezer.

Cartons from the old forming process never buckled under this handling, but cartons from the new forming process did. Rather than adopt more costly remedies, such as increasing the weight of the paperboard, the Huhtamaki facility created a team to identify new machine settings that would achieve sufficient crush resistance. The team performed two 2^3 factorial experiments, leading to adjustments in all four of the factors mentioned above.

In an initial experiment, belt tension was held constant at 65, and the other three factors were investigated using two levels for each:

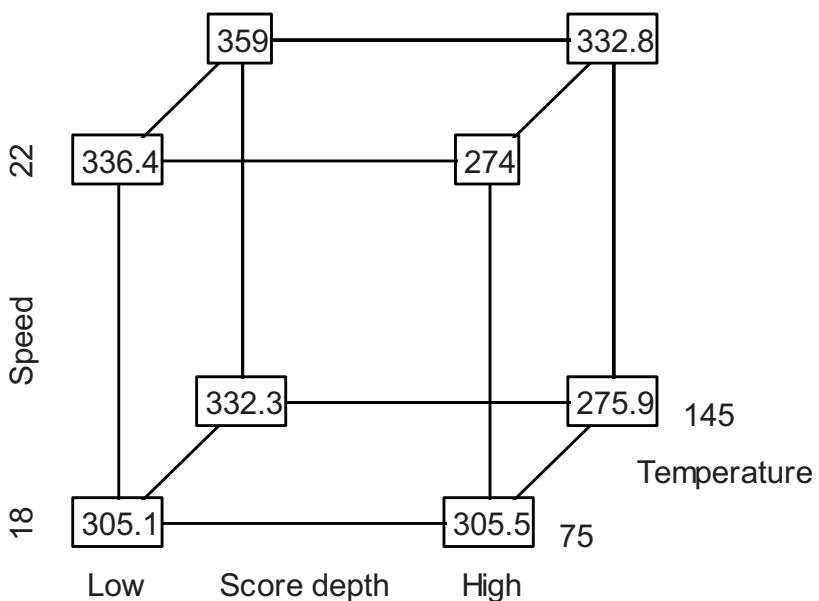
- Speed: 18 or 22 tubes per minute
- Score depth: low or high
- Temperature: 75°F or 145°F

Note that none of these changes is centered about current operating conditions, but they focus on increases in speed and temperature and a decrease in scoring depth. With three two-level factors, there are eight treatment combinations in the factorial design. Each was investigated twice; see Table 1.1. The response dry crush force is the weight in pounds required to compress a container to failure. The reported value for each run is the mean for 10 cartons.

Figure 1.1 presents a cube plot for this experiment, where each corner is labeled with the average dry crush force for two runs. The front, lower, right corner is closest to the current operating condition; the only difference is using 75°F instead of 90°F. Since the mean at this corner is as good or better than at adjacent corners, one-factor-at-a-time experiments might have led Huhtamaki to believe the current levels were near optimal for these factors. Instead, they performed a factorial experiment, which yielded best results at high speed, high temperature, and low scoring depth, all of which represent changes from the current process settings. Thus, their initial experiment identified a promising new combination of factor levels.

Table 1.1. Huhtamaki's initial dry crush experiment

Run	Score Depth	Speed	Temperature	Belt Tension	Dry Crush Mean
1	High	18	75	65	311.5
2	High	22	145	65	312.4
3	High	18	145	65	271.2
4	Low	22	145	65	365.6
5	Low	18	145	65	335.2
6	Low	18	75	65	315.1
7	Low	18	145	65	329.4
8	Low	22	75	65	353.8
9	High	22	75	65	286.4
10	Low	18	75	65	295.1
11	Low	22	145	65	352.4
12	High	18	75	65	299.5
13	High	18	145	65	280.6
14	High	22	75	65	261.6
15	High	22	145	65	353.2
16	Low	22	75	65	319.0

**Fig. 1.1.** Cube plot for predicted dry crush from Huhtamaki experiment 1

The team decided to conduct a second, briefer experiment with no replication; that is, each treatment combination was performed only once. Further, scoring depth, which required a time-consuming tooling change to switch levels, was dropped as a factor. Since promising results were obtained with low depth, a level for which they had less experience, the second experiment was conducted using only low scoring depth. Belt tension, which was held constant in the first experiment, was added as a new factor. Huhtamaki's team was interested to see if the belt tension effect depended on speed or temperature, since the belt drives the paperboard over the hot mandrel. In experiment 2, the same levels were chosen for speed and temperature, but belt tension levels were 60 and 75, straddling the current tension of 65. Since these three factors could be adjusted with the machine running, there was no need to shut down the equipment between runs. Experiment 2 is summarized in Table 1.2 and displayed in Figure 1.2. The data suggest that the belt tension effect depends on speed. At high speeds, a lower tension is better, whereas the opposite is true at low speed. Once again, the best results are achieved at the high-speed, high-temperature combination.

These experiments led Huhtamaki to change the process settings to higher speed, low scoring depth, 145°F mandrel temperature, and a belt tension of 60. Additional data were collected over several weeks at these settings to validate the changes. Using control charts to monitor the process, Huhtamaki confirmed that the process was now capable of achieving sufficiently high dry crush values.

Table 1.2. Huhtamaki's second dry crush experiment

Run	Score			Belt Tension	Dry Crush Mean
	Depth	Speed	Temperature		
17	Low	18	75	75	335.8
18	Low	22	145	60	378.1
19	Low	22	145	75	345.1
20	Low	18	75	60	299.6
21	Low	22	75	60	358.8
22	Low	22	75	75	349.9
23	Low	18	145	60	341.3
24	Low	18	145	75	359.8

1.1.2 The benefits of factorial experiments

There are two primary benefits of full factorial designs:

- Benefit 1. Full factorial designs reveal whether the effect of each factor depends on the levels of other factors in the experiment. This is the primary

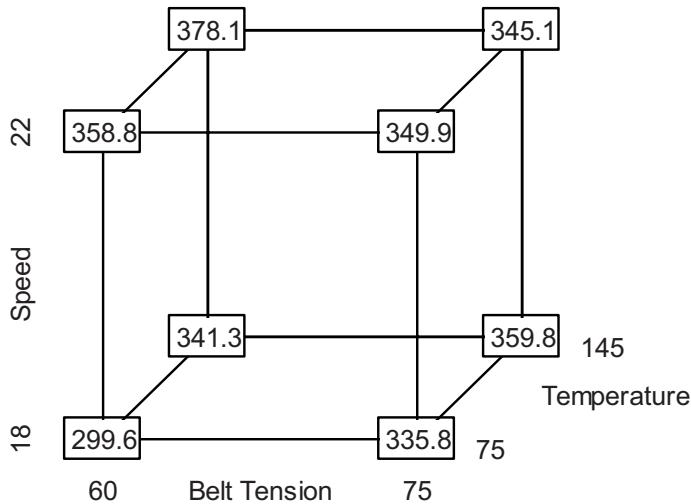


Fig. 1.2. Cube plot for dry crush from Huhtamaki experiment 2

reason for multifactor experiments. One factorial experiment can show “interaction effects” that a series of experiments each involving a single factor cannot.

- Benefit 2. Full factorial designs provide excellent precision for the regression model parameter estimates that summarize the combined effects of the factors.

Cox (1958, p. 94) explained these two benefits as follows. First, if the effect of one or more factors depends on the levels of other factors, this information will be recognized in a factorial experiment. If the actual model is simpler and the effects of the factors are additive, the factorial experiment is still better, in that all the data collected from the 2^k factorial design are used to estimate the effect for each factor. This is preferred to conducting three single-factor experiments, where each experiment tells us nothing about the effects for two of the three factors. Thus, whether the effects are additive or not, the factorial design has advantages.

Cox added a sequel to Benefit 1. Since a factorial experiment examines each factor’s effect under a range of conditions (i.e., at many combinations of levels for the other factors), we can understand the range of validity for conclusions about each factor. For example, if we held every other factor fixed while we experimented with the effect of increasing speed from 18 to 22 tubes per minute, we might conclude that increasing speed had negative consequences if the belt tension, temperature, and score depth were not at suitable levels for high speed. By experimenting with speed, temperature, and score depth simultaneously, we learn whether the speed effect on crush resistance depends

on temperature and/or score depth. Also, if such a dependence is found, we discover what temperature and score depth combinations are more conducive to high strength at the faster speed.

Full factorial designs explore the experimental region more effectively than do single-factor experiments. If each factor's effect does not depend on the levels of the other factors, then exploring the region using a 2^k design documents the additivity of the factor effects. If such dependencies do exist, using a 2^k design enables one to identify the combination(s) of factor levels that perform best.

1.2 Standard Regression Models for Factorial Designs with Two-Level Factors

In this section, we present multiple linear regression models for describing the effect of the k factors of a 2^k factorial experiment on a response. For convenience in the statistical analysis and consistency of interpretation, most models will use coded levels -1 and $+1$. For instance, we denote low score depth, 18 tubes/min and 75°F each by the coded level -1 , and high score depth, 22 tubes/min and 145°F with 1 . Using this coding, Table 1.1 is rewritten in Table 1.3, with x_1 , x_2 , and x_3 representing coded score depth, speed, and temperature, respectively.

Table 1.3. 2^3 factorial with coded levels for Huhtamaki experiment 1

Run				Crush Force
	x_1	x_2	x_3	Mean
1	1	-1	-1	311.5
2	1	1	1	312.4
3	1	-1	1	271.2
4	-1	1	1	365.6
5	-1	-1	1	335.2
6	-1	-1	-1	315.1
7	-1	-1	1	329.4
8	-1	1	-1	353.8
9	1	1	-1	286.4
10	-1	-1	-1	295.1
11	-1	1	1	352.4
12	1	-1	-1	299.5
13	1	-1	1	280.6
14	1	1	-1	261.6
15	1	1	1	353.2
16	-1	1	-1	319.0

We now present a series of regression models commonly used for analyzing factorial experiments. Using x_1, x_2, \dots, x_k to denote the coded factors is convenient for presenting these generic models.

Model 1: Simple Additive Model for 2^k Factorial

An additive model for the k factors is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \quad (1.1)$$

where the β_j 's are unknown regression coefficients to be estimated, ϵ is a random error term, and y is an observable value of the response at the treatment combination (x_1, x_2, \dots, x_k) . At this point, all we need to assume regarding ϵ is that it comes from a probability distribution that averages zero. Later, additional assumptions will be made about the variance or distribution of the random error term.

For this simple model, the effect of the j^{th} factor on the response is reflected in the coefficient β_j . If $\beta_j > 0$, we expect higher values for y at $x_j = 1$ than at $x_j = -1$; if $\beta_j < 0$, we expect lower values for y at $x_j = 1$. Furthermore, for this additive model, the effect of the j^{th} factor does not depend on the settings of the other $k - 1$ factors.

A point of clarification is needed before proceeding to the next model. Note that, on average, the difference between the response y at the two levels for x_j is $\beta_j(+1) - \beta_j(-1) = 2\beta_j$. Some books refer to $2\beta_j$ as the “effect” of factor j . However, here we emphasize the use of regression models with ± 1 coding, and so find it more natural to refer to the regression coefficient β_j as the main effect for x_j .

To represent this and subsequent models more compactly, we use summation notation. For instance, the additive model (1.1) may be written as

$$y = \beta_0 + \sum_{j=1}^k \beta_j x_j + \epsilon. \quad (1.2)$$

For many applications, the additive model provides a useful approximation for the combined effect the factors (x_1, x_2, \dots, x_k) have on a response y . However, in other cases, the effects of some factors will depend markedly on the levels of other factors. For such cases, we will use the following more complicated models.

Model 2: Two-Factor Interaction Model for 2^k Factorial

$$y = \beta_0 + \sum_{j=1}^k \beta_j x_j + \sum_{i=1}^{k-1} \sum_{j=i+1}^k \beta_{i,j} x_i x_j + \epsilon, \quad (1.3)$$

In addition to the $k + 1$ terms in model (1.2), the two-factor interaction model (1.3) contains $k(k - 1)/2$ additional terms of the form $\beta_{i,j} x_i x_j$. As earlier for β_j ,

$\beta_{i,j}$ is an unknown coefficient to be estimated, whereas $x_i x_j$ is the product of levels for factors x_i and x_j . To understand how this “interaction term” alters the model, consider the following simple numerical example for $E(y)$, the expected or average value of y . Suppose $E(y) = 50 + 10x_1 + 20x_2 - 5x_1x_2$. Then the expected values at the four treatment combinations of the 2^2 factorial are as given in Figure 1.3.

	$x_1 = -1$	$x_1 = +1$	Average across the levels of x_1
$x_2 = +1$	$50 - 10 + 20 + 5 = 65$	$50 + 10 + 20 - 5 = 75$	$50 + 20 = 70$
$x_2 = -1$	$50 - 10 - 20 - 5 = 15$	$50 + 10 - 20 + 5 = 45$	$50 - 20 = 30$
Average across the levels of x_2	$50 - 10 = 40$	$50 + 10 = 60$	

Fig. 1.3. Numerical example: $E(y) = 50 + 10x_1 + 20x_2 - 5x_1x_2$

If we average across the levels for x_2 , we obtain the equation $E(y) = 50 + 10x_1$, with $E(y)$ equal to 40 and 60 at $x_1 = -1$ and $x_1 = 1$, respectively. If we average instead across the levels for x_1 , we obtain $E(y) = 50 + 20x_2$, with $E(y)$ equal to 30 and 70 at $x_2 = -1$ and $x_2 = 1$, respectively. Thus, in an interaction model with ± 1 coding for the factor levels, the main effects represent the effects of each factor, averaging over the levels of the other factors.

Now consider the interaction term $-5x_1x_2$. At $x_2 = +1$, the equation for $E(y)$ becomes

$$E(y) = 50 + 10x_1 + 20(1) - 5x_1(1) = 70 + 5x_1,$$

whereas at $x_2 = -1$ the equation simplifies to

$$E(y) = 50 + 10x_1 + 20(-1) - 5x_1(-1) = 30 + 15x_1.$$

So the effect of factor x_1 depends on the level of x_2 . This is what we mean by *interaction*. Depending on the level of x_2 , we have a regression coefficient for x_1 of either $\beta_1 + \beta_{1.2} = 5$ or $\beta_1 - \beta_{1.2} = 15$.

Although the two-factor interaction model (1.3) is much more flexible than the additive model, there is no guarantee that either model is adequate. For cases with $k > 2$ factors, one can add higher-order interactions. A model guaranteed to fit any occasion is the saturated model.

Model 3: Saturated Model for 2^k Factorial

$$\begin{aligned} y = & \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{i=1}^{k-1} \sum_{j=i+1}^k \beta_{i,j} x_i x_j + \sum_{i=1}^{k-2} \sum_{j=i+1}^{k-1} \sum_{l=j+1}^k \beta_{i,j,l} x_i x_j x_l \\ & + \dots + \beta_{1,2,\dots,k} x_1 x_2 \cdots x_k + \epsilon, \end{aligned} \quad (1.4)$$

This model contains all possible interactions, up to and including the k -factor interaction. Since this saturated model contains 2^k β 's to be estimated, it will fit the average response perfectly at each of the 2^k treatment combinations of the full factorial.

At times we will find it useful to fit the saturated model, even though we hope that a much simpler model will be adequate. When fitting simpler models, we will follow the principle of hierarchy of effects; that is, if the k -factor interaction is included in the model, we will necessarily retain all lower-order terms. In general, a model is *hierarchical* if, for every term included, all lower-order terms involving subsets of the factors in that term are included. For instance, including the three-factor interaction $\beta_{1,2,3}$ implies the inclusion of the interactions $\beta_{1,2}$, $\beta_{1,3}$, $\beta_{2,3}$, as well as the three main effects.

Higher-order interactions can arise, and so it is useful to have models that include them. However, models with higher-order interactions are more difficult to interpret, as they preclude a simple description of the effect of the k factors on a response. Since models are intended to provide useful summaries, we will typically seek the most parsimonious (i.e., simplest) model that fits well enough to suit our purposes. George Box's astute observation is relevant: "All models are wrong, but some are useful." Often, simpler models are more useful.

1.3 Least Squares Estimation of a Regression Model

Except for the case of deterministic computer models, we never know $E(y)$. In every other situation, the response we see at a given treatment combination can change from one observation to the next. We use the random term ϵ in the models of the previous section to account for this variation. Section 1.2 assumed only that $E(\epsilon) = 0$. Now, to estimate the coefficients in these models efficiently, it is necessary to make further assumptions.

Suppose our k -factor experiment consists of 2^k treatment combinations, each observed $n \geq 1$ times. For each of these $N = n2^k$ experimental runs, we observe a response that we generically label y_i , ($i = 1, \dots, N$). Whatever model we fit, whether the additive model (1.1) or the saturated model (1.4), or something in between, it is common to assume that the errors $\epsilon_1, \dots, \epsilon_N$ are independently distributed with variance σ^2 . Given this assumption, standard

least squares provides an efficient estimator for the regression coefficients. We introduce the simple form of the least squares estimators and the necessary matrix notation and then return to the example from Section 1.1.

Let x_{ij} denote the level of factor x_j for the i^{th} run of the experiment. Then the additive model (1.1) for our N observations is

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix} \quad (1.5)$$

or, equivalently,

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

where \mathbf{Y} , \mathbf{X} , $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ denote vectors and the matrix corresponding to (1.5).

The least squares estimator for $\boldsymbol{\beta}$ of the additive model is denoted \mathbf{b} and is computed using

$$\mathbf{b} \equiv \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \quad (1.6)$$

Equation (1.6) simplifies greatly because, for a full factorial experiment with each treatment combination replicated n times,

$$\mathbf{X}'\mathbf{X} = N\mathbf{I}_{k+1}, \quad (1.7)$$

where \mathbf{I}_ν denotes an identity matrix of dimension ν . Thus, (1.6) simplifies to

$$\mathbf{b} = \mathbf{X}'\mathbf{Y}/N. \quad (1.8)$$

Since the first column of \mathbf{X} is all 1's, and the subsequent columns are half 1's and half -1 's, the solution in (1.8) is given by

$$b_0 = \sum_{i=1}^N y_i/N \equiv \bar{y} \quad (1.9a)$$

and, for $j = 1, \dots, k$,

$$b_j = \sum_{i=1}^N x_{ij}y_i/N = 0.5(\bar{y}_{j=+} - \bar{y}_{j=-}), \quad (1.9b)$$

where $\bar{y}_{j=+}$ and $\bar{y}_{j=-}$ denote means based on the $N/2$ observations where $x_{ij} = +1$ and $x_{ij} = -1$, respectively. Not only are these least squares estimators simple to compute, their variances are also simple. Because $(\mathbf{X}'\mathbf{X})^{-1}$

is a diagonal matrix with all diagonals of $1/N$, b_0, b_1, \dots, b_k are uncorrelated, each with variance

$$\text{Var}(b_j) = \sigma^2/N, \quad (1.10)$$

provided the errors $\epsilon_1, \dots, \epsilon_N$ are independently distributed with common variance σ^2 . To use (1.10), we need an estimate for the error variance σ^2 . This will be discussed momentarily. First, however, we discuss estimation of the regression coefficients for other models in addition to (1.1).

The preceding discussion in this section pertains to the least squares estimators of the simple additive model based on an equally replicated full 2^k factorial design. The same simplicity holds if we fit other models such as (1.3) and (1.4). For any equally replicated 2^k factorial design, every interaction column is orthogonal to every other column. Table 1.4 illustrates this result for a 2^3 factorial.

Table 1.4. Intercept, main effect and interaction columns for 2^3 factorial

1	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$
1	-1	-1	-1	1	1	1	-1
1	1	-1	-1	-1	-1	1	1
1	-1	1	-1	-1	1	-1	1
1	1	1	-1	1	-1	-1	-1
1	-1	-1	1	1	-1	-1	1
1	1	-1	1	-1	1	-1	-1
1	-1	1	1	-1	-1	1	-1
1	1	1	1	1	1	1	1

Table 1.4 gives the model matrix for the saturated model (1.4) for the case with $k = 3$ factors, $n = 1$, and $N = 8$. Let \mathbf{X} denote this matrix. Note that here $\mathbf{X}'\mathbf{X} = 8\mathbf{I}_8$. Since the interaction columns are orthogonal to the other columns, including interaction columns in \mathbf{X} increases the dimension of $\mathbf{X}'\mathbf{X}$ but does not change its simple form. This is true for any equally replicated 2^k . Hence, the simple estimators (1.9a)–(1.9b) are not affected by the inclusion of interaction terms. Further, the interaction coefficients have the same simple form; for example, for the two-factor interaction involving the first two factors,

$$b_{1.2} = \sum_{i=1}^N x_{i1}x_{i2}y_i/N = 0.5(\bar{y}_{1.2=+} - \bar{y}_{1.2=-}), \quad (1.11)$$

where $\bar{y}_{1.2=+}$ and $\bar{y}_{1.2=-}$ denote means based on $N/2$ observations where $x_{i1}x_{i2} = +1$ and $x_{i1}x_{i2} = -1$, respectively. The extension to other interactions should be obvious.

The fact that $\mathbf{X}'\mathbf{X}$ is a diagonal matrix adds great simplicity to model fitting. In a typical multiple regression setting, one uses stepwise regression

or all-subsets regression to choose a model. However, when $\mathbf{X}'\mathbf{X}$ is diagonal, the estimates do not change for terms in the model as other terms are added or dropped. Thus, we can fit a saturated model and use this full model to ascertain which terms to retain in a more parsimonious reduced model. To recognize which terms to retain, however, we will need an estimate for σ^2 .

Suppose one fits a saturated model for a 2^k factorial with n observations per treatment combination. Let \bar{y}_i denote the average of the n observations taken at the same treatment combination as the i^{th} observation. Partition the variation in the responses y_1, \dots, y_N as in Table 1.5.

Table 1.5. Analysis of variance partition for a saturated model

Source	Degrees of Freedom	Sum of Squares
Model (saturated)	$2^k - 1$	$\text{SS}_{\text{sat}} = \sum_{i=1}^N (\bar{y}_i - \bar{y})^2$
Pure error	$(n - 1)2^k$	$\text{SS}_{\text{pe}} = \sum_{i=1}^N (y_i - \bar{y}_i)^2$
Total (corrected)	$N - 1$	$\text{SS}_{\text{tot}} = \sum_{i=1}^N (y_i - \bar{y})^2$

Sample variances are computed by dividing sums of squares by their degrees of freedom. These sample variances in the analysis of variance are customarily called *mean squares* by statisticians, since a variance is the average (i.e., mean) of squared differences. The total variance for this sample,

$$\text{Total mean square} = \text{SS}_{\text{tot}}/(N - 1)$$

corresponding to the last row of Table 1.5, reflects the variation in the data about the overall mean \bar{y} ; its degrees of freedom are $N - 1$ since it is based on the variation of N y_i values about a single mean.

The pure error variance reflects the variation in the N y_i values about the average at each treatment combination (refer to the middle row of Table 1.5). Replication makes possible such an estimate of σ^2 . Later we will discuss estimates for σ^2 that are model dependent; that is, the validity of those estimators will depend on the unknown β 's. In contrast, the error variation in Table 1.5 depends only on having valid replication and so is called “pure error.” The degrees of freedom for pure error can be expressed as either $N - 2^k$ or as $(n - 1)2^k$. Both of these correspond to intuitive explanations of the error mean square. The first expression indicates that this is the variance of N y_i values about predicted values from a model with 2^k estimated parameters. The second expression indicates that from each of the 2^k treatment combina-

tions we have $n - 1$ degrees of freedom for estimating σ^2 and that the pure error variance pools all of these into one estimator.

Finally, the saturated model mean square shown as the first line in Table 1.5 reflects how much variation is “explained” by the model and is the variance of the predicted values about the overall mean. Since the \hat{y}_i values are computed from a model with $2^k - 1$ regression coefficients in addition to b_0 , these are the model degrees of freedom. Since the model and pure error degrees of freedom (df) sum to $N - 1$, and the model and pure error sum of squares sum to SS_{tot} , any row of Table 1.5 is easily determined given the other two. This partitioning of the sum of squares is called an analysis of variance (ANOVA) because it partitions the total variation of y into two parts: that explained by the model and that which is not explained.

We illustrate these calculations for the Huhtamaki experiment 1 data in Table 1.3. The fitted saturated model is

$$\begin{aligned}\hat{y} = & 315.125 - 18.075x_1 + 10.425x_2 + 9.875x_3 - 4.075x_1x_2 \\ & - 2.575x_1x_3 + 10.475x_2x_3 + 11.625x_1x_2x_3.\end{aligned}\quad (1.12)$$

This fitted model reproduces the mean value at each treatment combination exactly, as in Figure 1.1. Table 1.6 shows the degrees of freedom, sum of squares, and mean squares for fitting the saturated model.

Table 1.6. ANOVA for Huhtamaki experiment 1 data

Source	df	Sum of Squares	Mean Squares
Model (saturated)	7	12,816.07	1,830.87
Pure error	8	2,165.48	270.68
Total (corrected)	15	14,981.55	

The pure error sum of squares has 8 df. This is the combination of $n - 1 = 1$ df at each of the eight treatment combinations. The eight variances based on the pairs of y_i values at each treatment combination are displayed in Figure 1.4. The pure error sum of squares is the sum of the individual variances multiplied by the df $n - 1$; the pure error mean square 270.68 is the average of these eight variances. We are not surprised that the eight sample variances displayed in Figure 1.4 vary greatly, because each is based on only 1 df. The mean square pure error combines these eight variances into a single estimate, based on the assumption that $\text{Var}(\epsilon)$ is constant across the experimental region.

The validity of the pure error mean square depends on more than the assumption of a constant variance $\text{Var}(\epsilon) = \sigma^2$. It depends even more critically on the independence of the ϵ_i 's from run to run. Spacing out runs from the same treatment combination is one means of ensuring this independence. Randomizing the run order of the treatment combinations both accomplishes

such spacing and helps guard against bias from possible confounding sources of variation. For more about run order, see Section 13.5.1.

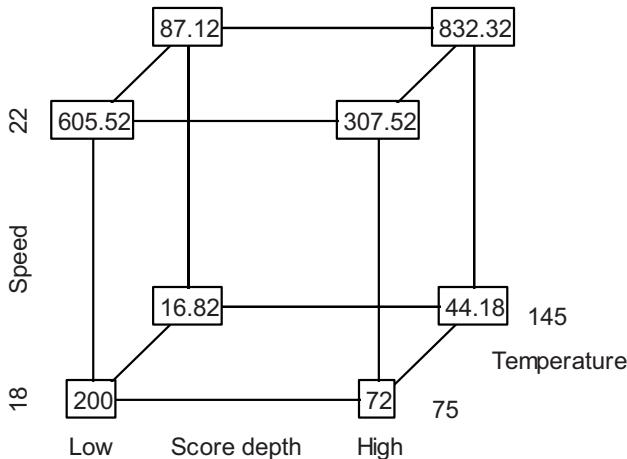


Fig. 1.4. Cube plot for variance of two y_i values at each treatment combination from Huhtamaki experiment 1

Let MS_{sat} and MS_{pe} denote mean squares calculated from the partitioning in Table 1.5. Then the ratio

$$F_{\text{sat}} = MS_{\text{sat}}/MS_{\text{pe}} \quad (1.13)$$

is used to test that all the true β coefficients besides β_0 in (1.4) are zero. Large values of F_{sat} indicate that at least some β coefficients are not zero and that their terms are needed in a model to explain variation in y . “Large” is determined relative to the percentiles of an F distribution with $\nu_1 = 2^k - 1$ df in the numerator and $\nu_2 = N - 2^k$ df for the denominator. The observed significance level (i.e., p -value) for this test is $P(F_{\nu_1, \nu_2} > F_{\text{sat}})$.

Using the mean squares in Table 1.6, we compute $F_{\text{sat}} = 1830.87/270.68 = 6.76$. Since the upper 1% point of the $F_{7,8}$ -distribution is approximately 6.2, based on interpolation in the Appendix B table, $P(F_{7,8} > 6.76) < 0.01$; the exact p -value is .0075. Such a small probability is compelling evidence that this model accounts for some systematic variation in dry crush; at least some of the β 's are not zero.

Typically, we are more interested in tests for individual coefficients. Let s denote any subset of the numbers $\{1, 2, \dots, k\}$. Then a test statistic to test the hypothesis

$$H_0 : \beta_s = 0$$

is the t -ratio

$$t = b_s / (\text{MS}_{\text{pe}}/N)^{1/2}. \quad (1.14)$$

The denominator in (1.14) is commonly named the standard error of the coefficient b_s . The usual test involves a two-sided alternative hypothesis, $H_A : \beta_s \neq 0$, in which case the p -value for the test is

$$P(|t_\nu| > t) = 2P(t_\nu > t),$$

where, in general, t_ν is a Student's t random variable and ν is the degrees of freedom for the mean square in the calculated standard error. Here $\nu = N - 2^k$, the pure error degrees of freedom.

If the errors ϵ in model (1.4) are normally distributed—in addition to being independently distributed with variance σ^2 —then the F -test and t -tests just presented are exact. These tests can also be motivated as approximations to randomization tests (see Box, Hunter and Hunter 2005, pp. 75–98). Additional details about estimation of σ^2 , as well as choice and assessment of reduced models, are deferred until Chapter 2.

Now we return to our example. Using the mean square error from Table 1.6 to estimate σ^2 , the standard error for each coefficient in (1.12) is

$$(270.68/16)^{1/2} = 4.113.$$

Standard statistical software produces the following output related to the coefficients of this fitted model:

Term	Estimate	Std Error	t	p -value
Intercept	315.125	4.113	76.61	<.0001
x_1	-18.075	4.113	-4.39	0.0023
x_2	10.425	4.113	2.53	0.0350
x_3	9.875	4.113	2.40	0.0431
$x_1 * x_2$	-4.075	4.113	-0.99	0.3508
$x_1 * x_3$	-2.575	4.113	-0.63	0.5487
$x_2 * x_3$	10.475	4.113	2.55	0.0343
$x_1 * x_2 * x_3$	11.625	4.113	2.83	0.0223

If a true coefficient $\beta_s = 0$, then the p -value for its t -test follows a uniform distribution between 0 and 1. If no factors had an effect, then we would expect to see about half of the p -values larger than .50 and half smaller than .50. On average, only one of the seven would be less than $1/7 = .14$. Instead, we see five of the seven p -values smaller than .05. This is evidence that all three main effects and the two largest interactions represent true effects for crush resistance. Since we require a hierarchical model and the highest-order interaction is statistically significant, no simplification of the model is possible.

The signs of the main effects indicate a general preference for low score depth, high speed, and high temperature for achieving higher crush resistance. However, given the two important interactions, they too must be considered. The Speed*Temperature and Depth*Speed*Temperature interaction estimates ($b_{2.3}$ and $b_{1.2.3}$) are nearly the same. Thus, when $x_1 = -1$, these

terms cancel one another, and when $x_1 = +1$, they sum. For high score depth, the fitted model for dry crush is

$$\begin{aligned}\hat{y} &= [315.125 - 18.075(1)] + [10.425 - 4.075(1)]x_2 \\ &\quad + [9.875 - 2.575(1)]x_3 + [10.475 + 11.625(1)]x_2x_3 \\ &= 297.05 + 6.35x_2 + 7.3x_3 + 22.1x_2x_3.\end{aligned}$$

Here, the interaction term dominates the main effects, so that the preferred level for temperature depends on the speed. At low score depth, the fitted model is simpler:

$$\begin{aligned}\hat{y} &= [315.125 - 18.075(-1)] + [10.425 - 4.075(-1)]x_2 \\ &\quad + [9.875 - 2.575(-1)]x_3 + [10.475 + 11.625(-1)]x_2x_3 \\ &= 333.2 + 14.5x_2 + 12.45x_3 - 1.15x_2x_3.\end{aligned}$$

Here the effects of speed and temperature are both positive and essentially additive.

1.4 Presenting a Fitted Model Graphically

Graphics are useful for displaying the results of an experiment. Interaction plots are an excellent means to assist with the interpretation. In general, an interaction plot displays the predicted y values for all combinations of two or more factors, averaging over the levels of factors not involved in the interaction. For instance, for the second Huhtamaki experiment (refer to Table 1.2), the Speed*Tension effect was believed to be important. This effect is displayed twice in Figure 1.5. In each plot, the same four means are displayed, with Speed or Tension on the horizontal axis and the response on the vertical axis. The second factor is identified by labels inside the plot. In the first plot, we see that the Speed effect on crush resistance is positive at low tension but disappears at high tension. In the second plot, we see that the Tension effect on crush resistance is positive at 18 tubes per minute but negative at 22 tubes per minute. When presenting results for a fitted model, choose the display that communicates most effectively. In this case, either is satisfactory. However, when one factor is qualitative and the other is quantitative, it is generally preferred to place the quantitative factor on the horizontal axis.

Three-factor interaction plots are needed less often but are necessary for interpretation when a three-factor interaction coefficient is large. We illustrate two options using the first Huhtamaki experiment. Figure 1.6 displays the predicted values for the eight Depth*Speed*Temperature means. From this graph we see that generally increasing Temperature improves crush resistance, with the only exception being for high Score depth at the lower Speed. When constructing this graph, it is best to use the same range for the vertical axis. Alternatively, the four combinations for two factors can be displayed in a

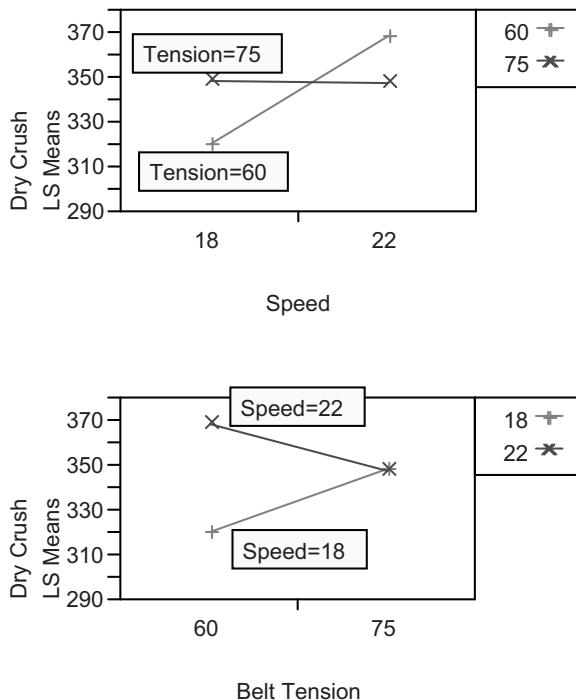


Fig. 1.5. Two versions of the Speed*Tension interaction plot from Huhtamaki experiment 2

single graph; see Figure 1.7. In either figure, the higher Temperature provides greater crush resistance—except for Speed = 18 and high Score depth, the combination that Huhtamaki previously operated this process.

Except when $k = 2$ or 3 , interaction graphs such as Figures 1.5 or Figures 1.6 and 1.7 focus on just a subset of the terms in the model. To display the model more fully we use cube plots and profiler plots. Cube plots have been used extensively throughout this chapter to display the predicted response (or variance) at each treatment combination. The profiler plot (as provided by JMP software) displays estimated $E(y)$ values for a fitted model, and confidence bounds for those estimates, at any combination of the levels of the factors. When the factors are represented as nominal categories, the profiler plot displays this predicted response and the effect of moving one factor at a time from that location. For instance, in Figure 1.8 we see the model from the perspective of (Depth = Low; Speed = 18, Temperature = 75); the predicted value here is 305.5, and no adjacent corner is superior.

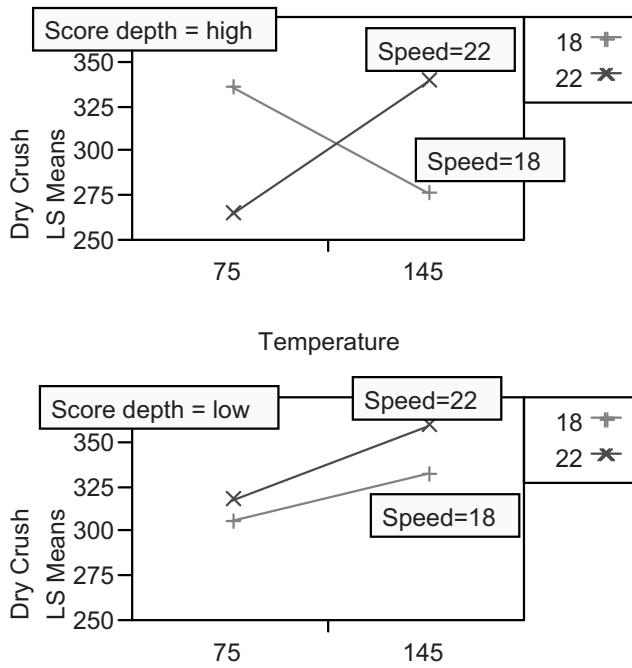


Fig. 1.6. Three-factor interaction plot for Huhtamaki experiment 1

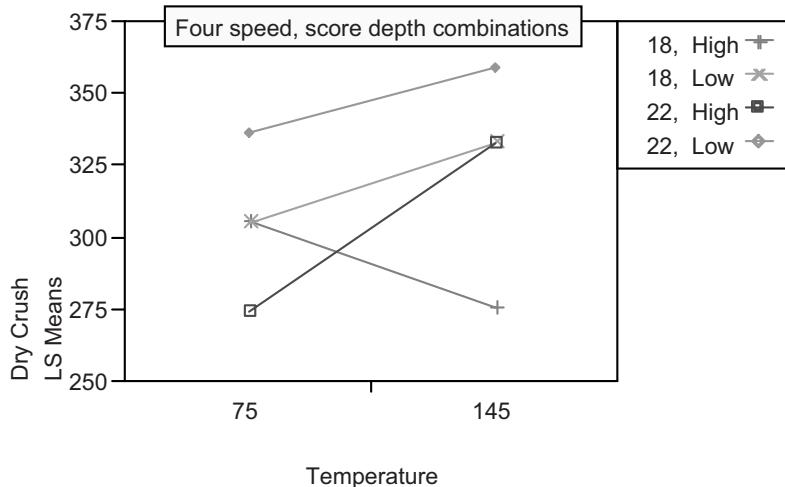


Fig. 1.7. Alternative three-factor interaction plot for Huhtamaki experiment 1

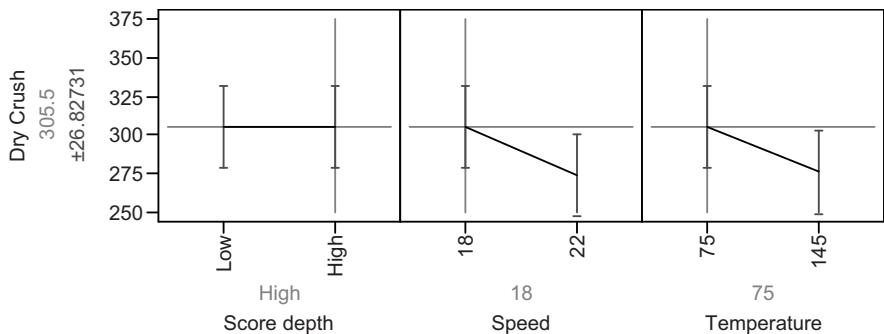


Fig. 1.8. Profiler graph for Huhtamaki experiment 1; setting at (Depth = High, Speed = 18, Temperature = 75) with factors modeled as nominal variables

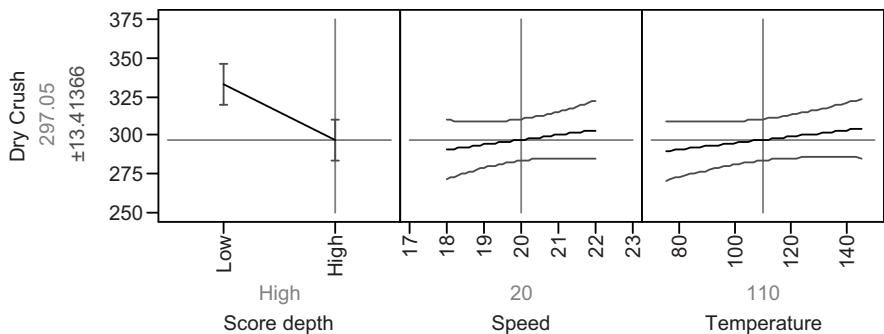


Fig. 1.9. Profiler graph for Huhtamaki experiment 1; setting at (Depth = High, Speed = 20, Temperature = 110) with Speed and Temperature as continuous variables

If the factors are continuous variables so that any setting is allowed, JMP's profiler plot is more useful. Consider Figure 1.9, which is a display for the saturated model (1.12), taken from the perspective of (Depth = Low, Speed = 18, Temperature = 75). Since both Speed and Temperature are set halfway between their low and high levels, this is the center of the right face of the cube plot (1.1). The displayed effect for Score depth in Figure 1.9 has two interpretations. Based on interpolation with the model in (1.12), this is the Score depth effect at the specified (Speed, Temperature) combination. It also represents the Score depth effect, averaging over the levels for Speed and Temperature. Comparing Figures 1.8 and 1.9, we see the impact of interactions, since the slopes of the lines have changed.

For more complicated models, surface plots and contour plots are also common for displaying fitted models, but these are of less use for models fit from two-level designs. This chapter recommended the use of regression models

with coded factors, but all graphics in this chapter were obtained using models with actual factor names and levels rather than coded levels. Coded labels and factors are more convenient for model selection and statistical tests. However, once the appropriate model is identified, it is helpful to obtain graphical output utilizing actual factor and level names, since doing so increases the readability of the figures. Gladly, some statistical software will automate such coding.

1.5 Four Steps for Planning a Successful Experiment

Successful experiments, such as Huhtamaki's two experiments, generally require hard work from many individuals. Practical experience teaches us that experiments rarely go as planned and that misunderstandings among those involved are difficult to avoid. Fortunately, imperfect experiments generally provide some useful insight. The most detailed summary of experiment planning in the industrial environment is by Coleman and Montgomery (1993). Moen, Nolan, and Provost (1998) provided a briefer strategy. The steps presented here are primarily a synthesis of these two sources. Hahn (1984) provided a series of examples that is particularly instructive for young statisticians who will be assisting others in planning experiments.

The four steps summarized in this section and listed in Figure 1.10 highlight the sequence of decisions needed to prepare for an experiment. Coleman and Montgomery (CM) (1993) provided further details regarding each issue. Both CM and Moen et al. (1998) emphasized the importance of documentation of the planning process and provided a master guide sheet as well as separate forms for various steps.

PREP: Four Steps to a Successful Experiment

Purpose: Set the objectives and identify current state of knowledge.

Responses: Select the primary responses to be measured.

Experimental factors and region: Determine factors and other variables of interest. Choose factor levels.

Plan Details: Choose the experimental design. Establish a schedule and responsibilities. Practice analysis for the chosen design.

Fig. 1.10. Summary of planning steps

Step 1: Set the objectives and identify current state of knowledge.

Coleman and Montgomery (1993) noted that “writing the objective is harder than it appears to most experimenters.” Determining the objectives and gathering background information are done in concert, because the initial description of what we hope to learn is usually modified when we discover what others know about the process being investigated. In an industrial setting, involving a team of individuals in the early planning stages is most valuable. Although doing so may appear to slow the planning process, avoiding the inefficiencies of “naive empiricism and duplication of effort” (CM, p. 4) pays off in the end. Careless, hasty experimentation is the surest means of building resistance to future experimentation.

Objectives should be stated in practical terms, emphasizing what future actions will potentially be impacted. Doing so will help gain buy-in from those whose support is needed to carry the project to conclusion. A statistical description of the objectives will be added in subsequent planning steps. For experiments pertaining to a large, multifaceted program, Barton’s (1997) goal hierarchy plot will be useful.

Step 2: Select the primary responses to be measured.

Most experiments entail a small number of primary responses that relate directly to the objectives. Ability to measure the characteristics of interest both accurately and precisely is essential; Wheeler and Lyday (1989) is a helpful reference. Invalid measurements can mislead. This author recalls George Box describing how apparent gains from indigo experiments at Imperial Chemical in the 1940s never materialized. What was the culprit? After the fact, it was discovered that the measurements for indigo were also sensitive to an impurity, so that the experiments led to conditions producing not more indigo but more impurity instead. Insufficient measurement precision is a more common problem. Imprecise measurements can render an experiment ineffective by so increasing the error variation that systematic effects of interest do not stand out above the background variation.

Each objective specified in Step 1 should be linked to one or more responses to be measured. At this step we must specify what functions of each measured responses is of interest. For quantitative responses, do our objectives relate only to the mean, or are we also concerned about variability in the response? More will be said in Section 13.3 about the within-run sampling needed to study variation. If the response is not easily quantified, is there some quasi-numerical (or ordinal scale) that can be applied, or must we resort to a simple yes/no response. Yes/no responses are less informative than quantitative and ordinal responses, so more data will be required if we do have only yes/no results. Sample size issues for yes/no data are discussed in Section 13.1.2.

Step 3: Determine factors and other variables of interest. Choose factor levels.

Depending on the objective and current state of knowledge, this step may or may not begin by brainstorming. If the experiment is a screening experiment whose purpose is to identify factors that affect the response, then it is critical to be thorough in listing the possibilities. This list will depend on the objective of the experiment; that is, candidate factors for changing the mean thickness of a deposition process will differ somewhat from candidates for decreasing the variation in thickness. Even when the experimental factors of interest are specified in the objective statement from Step 1, it is important to identify other variables that potentially affect the response. In addition to the variables that are varied as experimental factors, other possibly influential variables should be identified and held fixed during the course of the experiment—or at least held fixed within blocks. Variables that cannot be held fixed should be measured, since such measurements may prove useful for the subsequent analysis. It is important to record any changes that arise during the course of the experiment. In addition, it is useful to identify additional outputs of the process to measure that might correlate with the primary responses. Thus, the conclusion of this step is four lists:

- Variables to be varied as factors
- Variables to be held constant—or only varied between blocks
- Hard-to-control variables that can be measured
- Secondary responses that may correlate with primary responses

The choice of factor levels is essential to the success of two-level designs. Devoting an entire section to this choice (Section 13.2) emphasizes its importance. Prior knowledge or the use of trial runs is essential to the proper choice of levels for quantitative factors. A level for each held-constant variable must also be specified. For applications where there is debate about this level, consider the question, “What level is likely to render the experiment more informative?”

Step 4: Choose the experimental design. Establish a schedule and responsibilities. Practice the analysis for the chosen design.

As will be described in Chapter 3, an experimental design involves more than determining the treatment structure—that is, which treatment combinations are to be performed. It also involves the unit structure. The unit structure specifies the use of blocking and randomization. Answers to the following questions are needed to arrive at a proposed design:

- How long will each individual run take and how much experimental material is required?
- How much time is required between runs for changes in factor settings and time to reach steady state?

- Is the between-run time lengthened substantially by a small set of hard-to-change factors?
- What are the major sources of error variation associated with the responses?
- How likely are interactions to be important? What interactions are expected to be the most likely?

The efficiency of an experimental design for detecting systematic effects is largely contingent on properly anticipating the sources of random variation. For instance, if the measurements themselves are imprecise and we are interested in the mean response, then sample multiple items within each run and use the average of these measurements as the response. If batch-to-batch variation is substantial, then either treat batch as a blocking variable or blend batches together to create more homogeneous raw material.

Generally, several designs should be considered. Factor relation diagrams (see Section 4.3.3) are helpful for documenting specifically what is intended by each possible design. These visual aids have proven very helpful for contrasting alternative designs, especially when teams are involved.

The schedule should involve preparation for the experiment, including all equipment, materials, and training. The schedule should also document the timing and sequencing of the runs. For many experiments, this sequence specifies both a processing/manufacturing phase and a measurement phase. Combining likely main effects and interactions and anticipated sources of variation, a simulation model can be constructed generating pseudo-data and an analysis performed. Performing such a practice analysis before the data are collected will ensure that one recognizes the aliasing and confounding implied by the selected design.

In addition to practicing the analysis, it is generally helpful to perform one or two trial runs. From these runs, one can verify how long it takes to set up and perform a run, whether we are capable of controlling variables as specified, whether we can measure the responses as indicated, and whether the results correspond to our expectations. Ideal candidates for trial runs are the treatment combinations expected to give the worst and best (or highest and lowest) responses. Results vastly different than our expectations in the time or results may cause us to reconsider the design or choice of levels.

The next chapter focuses on the many tools available for analyzing two-level factorial designs.

Analysis of Full Factorial Experiments

This chapter details how to analyze 2^k factorial experiments and is organized as follows:

- Section 2.1. Analysis Strategy Overview
- Section 2.2. Analysis of Numerical Responses with Replication
- Section 2.3. The Inclusion of Centerpoint Replicates
- Section 2.4. Analysis of Numerical Responses Without Replication
- Section 2.5. Normal Plot of Effects and Other Analysis Tools
- Section 2.6. Diagnostics for a Fitted Model
- Section 2.7. Transformations of the Response
- Section 2.8. Analysis of Counts, Variances, and Other Statistics
- Section 2.9. Unequal Replication and Unequal Variance
- Section 2.10. The Impact of Missing Treatment Combinations

2.1 Analysis Strategy Overview

The following four-step strategy is recommended for the analysis of 2^k factorial experiments.

2.1.1 Step 1: Study the variation in y

Begin with a histogram of the response data y and observe the range and distribution of values. If the distribution is evenly spread, then fitted models will not be overly affected by just a small subset of the data. If the distribution

is severely skewed, or there are a few values far removed from the others, then the fitted models will attempt to account for this prominent variation while largely ignoring the rest.

The shape of the distribution of y can be altered by the use of a non-linear transformation. Section 2.7 explains how such transformations may be employed to find a satisfactory simpler model, to stabilize the error variance, or to emphasize the variation at the lower or upper end of the range for y . If the treatment combinations are replicated, then one should examine the within-treatment-combination variation to check for consistency.

In addition to plotting the data, one should understand how the actual y values were obtained, since this may provide insight regarding the error variation. How large is the measurement error variance for the measurement system involved? Does the variability in y increase or decrease as the mean for y increases? Is y a count, a ratio, a standard deviation, or some other statistic? Section 2.8 provides guidance for each of these cases.

2.1.2 Step 2: Fit a “full” model

Step 2 begins by fitting a “full” model. For most situations, this will be the full factorial model (1.4). Rather than fitting a simpler model from the start and assuming it to be adequate, we prefer to fit a complex model and so confirm what terms are not needed. There are exceptions [e.g., for cases of missing treatment combinations (Section 2.10) or with prior knowledge that certain interactions are not needed] where it is preferred to begin with a simpler model. However, the typical initial model for analyzing 2^k experiments will be the full factorial model (1.4).

How we proceed after fitting a complex model will depend on whether the experiment includes replication—that is, were runs repeated at some or all of the treatment combinations? Sections 2.2–2.5 will discuss methods and tools appropriate for the different cases that arise. The objective is to determine which terms are useful for explaining the variation in y and providing insight into the factor effects.

2.1.3 Step 3: Find an adequate simpler model

Now fit a reduced (i.e., simpler) model, as appears reasonable following Step 2. The purpose here is not to determine the significance of the remaining terms but rather to perform diagnostics to determine whether the reduced model adequately explains the variation in the response (see Section 2.6). If the residual analysis indicates problems, then some remedy is required, such as adding terms to the model, questioning aberrant y_i values, or transforming the response. Once a satisfactory model is obtained, one may proceed to Step 4.

2.1.4 Step 4: Interpret and utilize the final model(s)

Use graphs to summarize the results of the satisfactory model. Express the conclusions in the natural units for each factor and the response. If a transformation for y was involved in the analysis, quantitative results should also be expressed in terms of the original measurement rather than simply on the transformed scale. If two competing models seem reasonable, compare them to see in what respects they differ. For instance, do they differ regarding the preferred level for each factor? Do their predicted values differ at the treatment combination(s) of interest?

2.2 Analysis of Numerical Responses with Replication

As in Section 1.3, here we consider the simplest (although not necessarily common) case, where the 2^k treatment combinations of a full factorial are each replicated n times in a manner that yields $N = n2^k$ observations with independently distributed errors. Section 1.3 discussed t -tests for individual coefficients, as well as a test involving all the saturated model's coefficients.

Following tests for individual coefficients, one proceeds in Step 3 of the analysis strategy to fitting a reduced model with, say, r coefficients, including the intercept, b_0 , with $1 < r < 2^k$. Let \mathbf{X}_{red} denote the $N \times r$ model matrix, let \mathbf{b}_{red} denote the vector of least squares estimates for the reduced model

$$\mathbf{b}_{\text{red}} = (\mathbf{X}_{\text{red}}' \mathbf{X}_{\text{red}})^{-1} \mathbf{X}_{\text{red}}' \mathbf{Y} = \mathbf{X}_{\text{red}}' \mathbf{Y} / N,$$

and let $\hat{\mathbf{Y}} = (\hat{y}_1, \dots, \hat{y}_N)'$ denote the vector of predicted values

$$\hat{\mathbf{Y}} = \mathbf{X}_{\text{red}} \mathbf{b}_{\text{red}}.$$

The partitioning of the sum of squares corresponding to this reduced model is given in Table 2.1.

Table 2.1. Analysis of variance for a reduced model

Source	df	SS
Model (reduced)	$r - 1$	$\sum_{i=1}^N (\hat{y}_i - \bar{y})^2$
Lack-of-fit	$2^k - r$	$\sum_{i=1}^N (\bar{y}_i - \hat{y}_i)^2$
Pure error	$N - 2^k$	$\sum_{i=1}^N (y_i - \bar{y}_i)^2$
Total (corrected)	$N - 1$	$\sum_{i=1}^N (y_i - \bar{y})^2$

Table 2.1 expands Table 1.5, in that the saturated model's degrees of freedom and sum of squares are partitioned into two parts: the reduced model and lack-of-fit. The reduced model captures variation explained by the reduced model. Lack-of-fit contains variation that is explained by the saturated model but which is missed by the reduced model. In settings such as this, most statistical software will construct two F -tests:

- **Lack-of-fit test.** This is a test that the reduced model is adequate (i.e., that it explains all the systematic variation in the y_i values). The test statistic is

$$F_{\text{lof}} = \text{MS}_{\text{lof}} / \text{MS}_{\text{pe}},$$

where MS_{lof} and MS_{pe} denote the mean squares for lack-of-fit and pure error, respectively, computed from Table 2.1. The degrees of freedom for this test are $\nu_1 = 2^k - r$ and $\nu_2 = N - 2^k$, and the p -value is $P(F_{\nu_1, \nu_2} > F_{\text{lof}})$. A small p -value indicates that at least one of the β 's for terms omitted from the model is not zero; in this case, one should consider adding terms. A large p -value indicates that the reduced model is consistent with the observed data.

- **Reduced model test.** This is a test of significance for the terms in the reduced model. It is computed as

$$F_{\text{red}} = \text{MS}_{\text{red}} / \text{MSE},$$

where the denominator is the mean square error (MSE) for the reduced model obtained by pooling lack-of-fit and pure error as follows:

$$\text{MSE} = \frac{\text{SS}_{\text{lof}} + \text{SS}_{\text{pe}}}{N - r}.$$

This MSE combines MS_{pe} , an estimate for σ^2 based on replication, with MS_{lof} , an estimate for σ^2 that is dependent on the assumption that the reduced model is correct. A small p -value is an indication that the model is useful for explaining variation in the y_i 's, or, equivalently, that at least some of the β 's corresponding to terms in the model are not zero.

If both F_{lof} and F_{red} have large p -values (e.g., more than 5 or 10%), then the factors have no recognizable effect on $E(y)$.

We illustrate these F -tests for the first Huhtamaki experiment, taking the additive model (1.2) as our reduced model. The resulting lack-of-fit test is

Source	df	SS	MS	F_{lof}	p -value
Lack-of-fit	4	4289.64	1072.41	3.96	0.0463
Pure error	8	2165.48	270.69		
Total error	12	6455.12			

This test, which is statistically significant at $\alpha = .05$ indicates that this simple model does not account for all the systematic variation in dry crush resistance. Hence, one or more of the four omitted interactions is active. The corresponding F -test for the significance of the fitted reduced model is

Source	df	SS	MS	F_{red}	p-value
Model	3	8,526.43	2,842.14	5.28	0.0149
Error	12	6,455.12	537.93		
Total (corrected)	15	14,981.55			

Note that because the MS_{lof} is nearly four times the MS_{pe} , the MSE is inflated by the systematic variation in MS_{lof} , reducing the size of F_{red} as well as any t statistics computed as

$$t = b_s / (\text{MSE}/N)^{1/2}. \quad (2.1)$$

If there are sufficient degrees of freedom from replication, then it is safer to just use (1.14) rather than (2.1). Here, with both F -tests statistically significant, we would conclude that the additive model is useful but that it can be improved by the addition of interaction terms.

In summary, replication of the factorial treatment combinations serves two purposes. First, it provides information about the error variance. Replication at each treatment combination yields MS_{pe} as an estimate for σ^2 and provides some ability to check the assumption that $\text{Var}(\epsilon)$ is constant across the experimental region (something we will explore later). In addition, replication at the 2^k treatment combinations increases the precision for each estimated coefficient. When the error variance is substantial, experiments with small N may have too little power to detect effects of practical importance. The issue of sample size to achieve sufficient power is relevant for every application, and is addressed in Section 13.1.

2.3 The Inclusion of Centerpoint Replicates

Taking $n > 1$ replicates at every treatment combination, as was assumed in Section 2.2, can be quite costly, especially if there are four or more factors. One option to economize on runs is to collect replicates at only a subset of the treatment combinations (Dykstra 1959). However, such unequal replication forfeits the orthogonality of the columns of \mathbf{X} and so complicates the analysis. Section 2.9 will address how to analyze unbalanced factorial designs in general. Now consider an alternative economical approach to replication.

2.3.1 Centerpoint replication with all factors quantitative, with Example 2.1

When all of the factors are quantitative, an alternative to replicating some or all of the 2^k treatment combinations is to perform replicate runs at the center of the design. Replication at the center does not improve the precision of estimates for factorial effects, but it serves two other purposes. First, collecting data at the center provides a check on linearity of the factor effects. If the model is to be used for interpolation, this check is critical. If the centerpoint

runs indicate severe nonlinearity, then one often augments the design with additional treatment combinations to support estimation of a full second-order model. See Chapter 12 for details.

As with any true replication, centerpoint replication provides an estimate for σ^2 . Runs at the center do not affect the orthogonality of a design and so do not cause the complication that arises from partial replication of factorial treatment combinations. This method is recommended for estimating σ^2 , provided: (i) all the factors are quantitative, (ii) the constant variance assumption for ϵ is reasonable, and (iii) an unreplicated 2^k provides enough precision for estimating factorial effects.

Example 2.1: 2^5 with seven centerpoint runs

Consider now a five-factor example from Bouler et al. (1996). The experiment was conducted to improve the compressive strength of a calcium phosphate ceramic intended as a bone substitute. Biphasic calcium phosphate (BCP) is a blend of two materials denoted HA and β -TCP. BCP containing pores with diameter $\geq 100 \mu\text{m}$ promotes bone formation but generally has reduced strength. The purpose of the experiment is to create stronger BCP with such macropores. The factors and their levels are presented in Table 2.2.

Table 2.2. Factors and levels for Bouler et al.'s (1996) ceramic experiment

Factors	Levels		
	-1	0	1
x_1 HA in BCP (%)	45	60	75
x_2 Weight of naphthalene (%)	30	45	60
x_3 Diameter of macropores (μm)	100	300	500
x_4 Isostatic compaction (kPa)	1090	1630	2180
x_5 Sintering temperature ($^\circ\text{C}$)	900	1000	1100

The $2^5 = 32$ factorial treatment combinations were performed without replication; that is, $n = 1$. In addition, $n_0 = 7$ samples were made at the coded treatment combination $(0, 0, 0, 0, 0)$. Bouler et al.'s (1996) work does not mention any randomization of order in preparing or testing the samples. The observed compressive strengths ranged from 0 to 59.1 mPa. Table 2.3 presents the results for all $2^5 + n_0 = 39$ runs, sorted by compressive strength. Note that 10 of the 39 samples showed no compressive strength, including all 8 combinations with $x_2 = 1$ and $x_3 = -1$; that is, all combinations with a high weight of the smallest-diameter naphthalene. Clearly, this combination is not satisfactory.

Table 2.3. Bouler et al.'s (1996) ceramic strength data

Run	x_1	x_2	x_3	x_4	x_5	Strength (mPa)
1	-1	1	-1	-1	-1	0.0
2	1	1	-1	-1	-1	0.0
3	1	1	1	-1	-1	0.0
4	-1	1	-1	1	-1	0.0
5	1	1	-1	1	-1	0.0
6	-1	1	-1	-1	1	0.0
7	1	1	-1	-1	1	0.0
8	1	1	1	-1	1	0.0
9	-1	1	-1	1	1	0.0
10	1	1	-1	1	1	0.0
11	-1	1	1	-1	-1	2.0
12	-1	-1	-1	-1	-1	2.2
13	-1	-1	-1	1	-1	2.9
14	1	1	1	1	-1	3.3
15	-1	1	1	1	-1	4.2
16	1	-1	-1	-1	-1	5.1
17	-1	-1	1	-1	-1	6.5
18	1	-1	-1	-1	1	7.0
19	1	-1	1	-1	1	7.0
20	1	-1	1	-1	-1	8.0
21	0	0	0	0	0	10.8
22	0	0	0	0	0	11.5
23	-1	-1	1	1	-1	11.7
24	0	0	0	0	0	11.8
25	1	-1	1	1	-1	12.3
26	1	-1	-1	1	-1	12.9
27	0	0	0	0	0	13.0
28	-1	1	1	1	1	13.2
29	0	0	0	0	0	13.4
30	0	0	0	0	0	13.9
31	1	1	1	1	1	14.1
32	0	0	0	0	0	14.5
33	1	-1	1	1	1	16.7
34	-1	1	1	-1	1	17.8
35	-1	-1	-1	1	1	23.4
36	1	-1	-1	1	1	25.7
37	-1	-1	1	-1	1	46.0
38	-1	-1	-1	-1	1	48.3
39	-1	-1	1	1	1	59.1

Table 2.4. Analysis of variance for Bouler et al. (1996) data

Source	df	SS	MS
Model (full factorial)	31	7034.15	226.908
Lack-of-fit (nonlinearity)	1	18.22	18.224
Pure error	6	11.12	1.853
Total (corrected)	38	7063.49	

As in Table 2.1, we construct an ANOVA with partitions for model, lack-of-fit, and pure error (see Table 2.4). Since we have fit a model containing all (linear) main effects and interactions, the lack-of-fit term has just 1 df, and is a check for nonlinearity (or curvature) of the factor effects. The nonlinearity sum of squares is based on the difference between the average response at the n_0 centerpoint replicates and the average at the N factorial treatment combinations:

$$SS_{\text{nonlin}} = \frac{(\text{Mean@Center} - \text{Mean@Factorials})^2}{n_0^{-1} + N^{-1}}. \quad (2.2)$$

Here, the mean strength for the 7 centerpoint replicates and 32 factorial treatment combinations are 12.7 and 10.92, respectively, and (2.2) equals 18.22. This lack-of-fit is small compared to the variation explained by the model ($MS_{\text{model}} = 226.9$), but is large compared to pure error ($MS_{\text{pe}} = 1.85$). Thus, while this lack-of-fit test is statistically significant ($F_{\text{lof}} = 9.83$; $p = .02$), accounting for this minimal curvature would make little difference in our predicted values in the interior of the design region. Since the centerpoint mean exceeds the average at the factorial points, a model that ignores this curvature will give slightly pessimistic predictions near the center of the experimental region.

In addition to F -tests, software will report *R-square* (a.k.a. the *coefficient of determination*) from the ANOVA for a fitted model:

$$R^2 = SS_{\text{Model}}/SS_{\text{Total}}.$$

R^2 is the proportion of total variation that is explained by the fitted model. Here, the full factorial model's $R^2 = 7034.15/7063.49 = 0.996$, which is very high, reflecting the practical insignificance of the lack-of-fit.

Results similar to Table 2.4 are typical for processes with little error variation. If the pure error mean square is close to zero, virtually every term of a saturated model will be statistically significant. Here we might question whether the variation at the centerpoint replicates accurately reflects the true run-to-run error variation. The error term ϵ consists of errors from several sources, including the following:

- Measurement error in the testing of compressive strength

- Inhomogeneity of the materials used
- Other uncontrollable errors at each stage in the process of mixing, compressing, and heating the ceramic

For true replication, all of these sources affect each observation independently. However, if a shortcut was taken by preparing as a batch the material for all seven centerpoint specimens, then these specimens may vary less in strength than would be the case if this step were performed seven times, once for each specimen. Whether this is the case or not, the nonlinearity in compressive strength observed is not large enough to make a substantial difference to the fitted model.

Both the design and analysis for Bouler et al.'s (1996) experiment warrant further discussion. From reading their work, it appears that the materials may have been prepared in larger batches and then sampled to apply the compaction and temperature levels. If this is the case, the distribution of the ϵ_i 's is affected, which alters how one should analyze the data (see Sections 3.5 and 3.6). Further, the fact that one-fourth of the observations showed no measurable strength calls into question using a single linear model for strength based on all the data. If zero strength indicates that the ceramic powder did not bond, then perhaps the 10 observations with $y_i = 0$ should be handled differently when constructing a model for strength. We return to this example later (Section 4.3) to address these issues.

2.3.2 Centerpoint replication with one or two qualitative factors

How can we replicate economically when some of the factors are qualitative? If all factors but one are quantitative, then collect centerpoint runs for the quantitative factors at both levels of the qualitative factor. For instance, Ellekjaer, Ilseng, and Naes (1996) conducted a cheese processing experiment in which just one of the factors, melting salt, was qualitative. They included 6 center runs—3 with melting salt A and 3 with melting salt B—along with the 32 factorial runs. If there are only two qualitative factors, one might collect one or two centerpoint runs at each of the four qualitative factor level combinations.

2.4 Analysis of Numerical Responses Without Replication

2.4.1 Model-dependent estimators for σ^2 , with Example 2.2

Many two-level full factorial and fractional factorial experiments are run without any replication. In such cases, one can still produce useful estimates for the error variance σ^2 , but these estimates are model dependent; that is, some assumptions must be made about the model in order to estimate the error variance. Three general approaches have been used, which depend on slightly different assumptions:

1. **Mean Square Error From an Assumed Model:** Prior to data analysis, assume a model less than the saturated model. Use the MSE from this fitted model to estimate σ^2 . Provided all omitted terms have true coefficients of zero, this yields an unbiased estimator for σ^2 .
2. **Mean Square Error From Final Fitted Model:** Analyze the data, and arrive at a satisfactory reduced model. Use the MSE for this model to estimate σ^2 . Here, the MSE is an unbiased estimator for σ^2 only if the nonzero β 's are so large that one always selects the same model. On subsequent pages, $\text{RMSE} = \text{MSE}^{1/2}$ is the acronym for *root mean square error* and will serve as an estimator for σ .
3. **Robust Estimator for σ From Saturated Model:** Fit a saturated model and use the estimates nearest to zero to construct an estimate for σ . We will use Lenth's (1989) estimator (explained below). Here one assumes that a majority of the terms for the saturated model have true coefficients that are zero. This assumption is known as *effect sparsity*.

Approach 1 for estimating σ^2 is valid, provided the assumed model is correct. For example, with an unreplicated 2^4 factorial, we might assume that no three-factor or higher-order interactions exist and fit the model (1.3). The resulting ANOVA will have 10 df for the model and 5 df for error. Provided $\beta_{1\cdot 2\cdot 3} = \beta_{1\cdot 2\cdot 4} = \beta_{1\cdot 3\cdot 4} = \beta_{2\cdot 3\cdot 4} = \beta_{1\cdot 2\cdot 3\cdot 4} = 0$, the MSE is a valid estimator for σ^2 .

Although approach 2 is commonly used in practice to estimate σ^2 , it is the most subjective method and entails dangers that can make it unreliable. For instance, if one naively selects a model by excluding only a few of the smallest estimates (e.g., using backward elimination regression), the MSE for the reduced model will generally be much smaller than σ^2 . As a result, many inactive terms may appear statistically significant.

We now introduce Lenth's method and compare it with the first two methods for estimating σ^2 , using data from a chemistry experiment.

Example 2.2: Unreplicated 2^4 isatin experiment

Consider now the data from Davies (1954, p. 275). This 2^4 experiment involved a laboratory investigation of yield for a derivative of the chemical isatin. Table 2.5 lists the four factors of interest to the chemist and the levels used in this initial investigation. The 2^4 treatment combinations were each performed once in random order. Table 2.6 lists the yield in grams per 10 g. of base material. The range 6.04 – 6.79 for yield seems rather small. Because this is an initial investigation into the process, the chemist had no knowledge of σ^2 but believed that three-factor and higher-order interactions were unlikely.

Table 2.5. Factors and levels for isatin yield experiment (Davies 1954)

Factors	Levels	
	-1	1
x_1 Acid strength (%)	87	93
x_2 Reaction time (min)	15	30
x_3 Amount of acid (mL)	35	45
x_4 Reaction temperature ($^{\circ}$ C)	60	70

Table 2.6. Coded treatment combinations sorted by isatin yield

x_1	x_2	x_3	x_4	Yield
1	-1	-1	-1	6.04
1	1	-1	1	6.08
-1	-1	-1	-1	6.08
1	-1	1	-1	6.09
-1	1	1	-1	6.12
1	1	1	1	6.23
-1	-1	1	-1	6.31
1	1	1	-1	6.36
1	-1	1	1	6.38
1	1	-1	-1	6.43
-1	1	1	1	6.49
-1	1	-1	-1	6.53
1	-1	-1	1	6.68
-1	1	-1	1	6.73
-1	-1	1	1	6.77
-1	-1	-1	1	6.79

These data are used to illustrate the potential advantages and disadvantages of the three methods for estimating σ^2 :

- Assuming away higher-order interactions, we fit model (1.3) with four main effects and six two-factor interactions, and obtain the analysis of variance

Source	df	SS	MS	F
Model	10	0.8525	0.08525	2.216
Error	5	0.1923	0.03847	
Total (corrected)	15	1.0448		

Diagnostics (as explained in Section 2.6) for this fitted model show no outliers or systematic patterns. Although the usefulness of this model is questionable, given $F = 2.216$ (p -value = .20), the $MSE = 0.038$ provides a valid estimate for σ^2 , provided the true regression coefficients for the five higher-order interactions are zero. The resulting t -tests are as follows:

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	6.3819	0.0490	130.16	<.0001
x_1	-0.0956	0.0490	-1.95	.1086
x_2	-0.0106	0.0490	-0.22	.8370
x_3	-0.0381	0.0490	-0.78	.4720
x_4	0.1369	0.0490	2.79	.0384
x_1x_2	-0.0006	0.0490	-0.01	.9903
x_1x_3	0.0169	0.0490	0.34	.7447
x_1x_4	-0.0806	0.0490	-1.64	.1610
x_2x_3	-0.0331	0.0490	-0.68	.5293
x_2x_4	-0.1256	0.0490	-2.56	.0505
x_3x_4	-0.0131	0.0490	-0.27	.7996

With $b_4 = 0.137$, we conclude that, averaging over the levels of the other factors, increasing temperature to 70°C improves yield. However, since $b_{2.4} = -0.126$, the temperature effect may be influenced by Reaction time. At 15 min, the estimated temperature effect is $0.137 - 0.126(-1) = 0.263$, whereas at 30 min, the estimated temperature effect essentially disappears. A Time*Temperature interaction plot would display this, and would indicate a preference for the 15-min, 70°C combination.

- Using a forward selection regression procedure with $\alpha = .05$ to select a hierarchical model for yield, we include two two-factor interactions, x_1x_4 and x_2x_4 , and the three main effects x_1 , x_2 , and x_4 . The analysis of variance for this fitted hierarchical model is as follows

Source	df	SS	MS	F
Model	5	0.8044	0.16088	6.690
Error	10	0.2405	0.02405	
Total (corrected)	15	1.0448		

What has changed from the previous analysis? We have dropped five terms from model (1.3) with hardly any decrease in the model sums of squares. The smaller MSE also results in a significant F statistic for the model (p -value = .0055) and smaller standard errors and smaller p -values for the estimated coefficients:

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	6.3819	0.0388	164.62	<.0001
x_1	-0.0956	0.0388	-2.47	.0333
x_2	-0.0106	0.0388	-0.27	.7896
x_4	0.1369	0.0388	3.53	.0054
x_1x_4	-0.0806	0.0388	-2.08	.0642
x_2x_4	-0.1256	0.0388	-3.24	.0089

Now, three or four effects appear to be statistically significant. With 10 df for error, twice the error degrees of freedom for Method 1, one might

presume that $\text{MSE} = 0.024$ provides a better estimate for σ^2 . However, it is also possible that this MSE is smaller because we have overfit the model by including terms that have larger estimates just by chance.

3. Fit a saturated model to the ± 1 coded factors and use the many coefficient estimates near zero to estimate σ . A Pareto plot of the 15 estimates is given in Figure 2.1. Lenth's (1989) procedure for estimating $\sigma/N^{1/2}$, the standard error of these estimates, is as follows:

- Determine the median absolute estimate for the main effects and interactions from a saturated model and compute s_0 as 1.5 times this median. Here, $s_0 = 1.5(0.038125) = 0.0572$.
- Exclude all estimates that exceed $2.5s_0$ in magnitude and recompute the median. Here, no estimates exceed $2.5s_0 = 0.143$, so the median remains 0.038125.
- Compute $\text{PSE} = 1.5 \times \text{median}$ (of estimates less than $2.5s_0$). Here, $\text{PSE} = 0.0572$.

Lenth's pseudo-standard-error (PSE) is an estimator for the standard error of the coefficients. Note how much larger it is than the standard error of 0.0388 from Method 2 above. Lenth's method provides a reasonable estimate for $\sigma/N^{1/2}$, provided only a few coefficients differ from zero. If this assumption is not correct, then Lenth's PSE will tend to overestimate $\sigma/N^{1/2}$. Lenth's PSE = 0.0572 corresponds to an estimate for σ of $\text{PSE}(N^{1/2}) = 0.0572(16^{1/2}) = 0.2288$. [Haaland and O'Connell (1995) show that the PSE is slightly biased upward when m is small, but the bias is only about 1% for $m = 15$.]

Term	Estimate
x4	0.1368750
x2*x4	-0.1256250
x1	-0.0956250
x1*x4	-0.0806250
x1*x2*x3	0.0743750
x2*x3*x4	0.0618750
x1*x2*x4	-0.0506250
x3	-0.0381250
x2*x3	-0.0331250
x1*x3	0.0168750
x3*x4	-0.0131250
x2	-0.0106250
x1*x2*x3*x4	0.0093750
x1*x3*x4	-0.0031250
x1*x2	-0.0006250

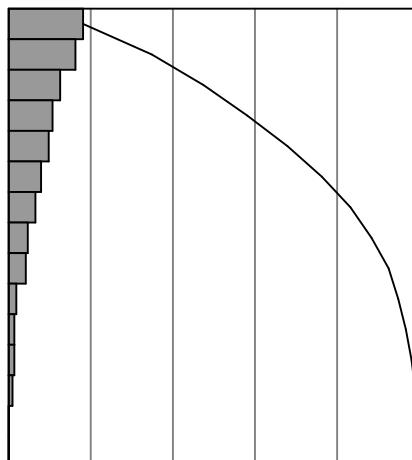
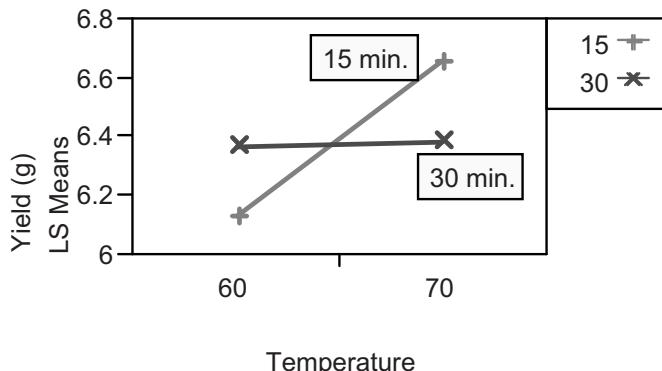


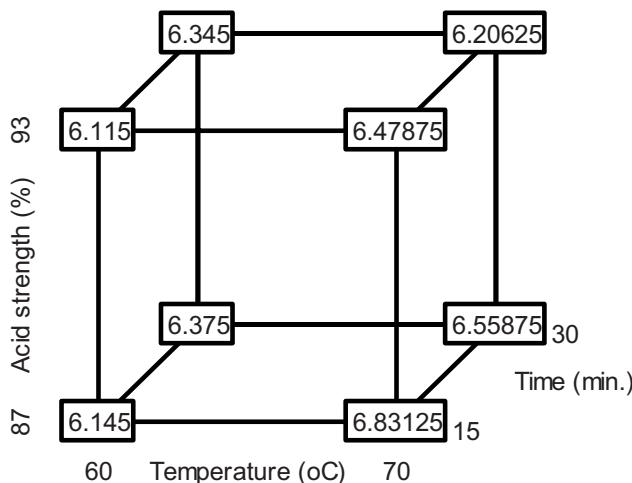
Fig. 2.1. Pareto plot of estimates from a saturated model for Davies's 2^4 experiment

These methods produced three different estimates for the error variance, ranging from Method 2's 0.024 to Method 3's $N(PSE)^2 = 0.052$. Which coefficient estimates are statistically significant also varies from method to method. Which fitted model is best and which estimate is closest to the true σ^2 are unknown. For now, we discuss the possible interaction terms and then return to the discussion about estimators for σ^2 .

With Method 1, $x_2 * x_4$ is the only statistically significant interaction. From its interaction plot



we conclude that 15 min at 70°C is preferable. With Method 2, we include an additional term or two that involve acid concentration (x_1). The model with 5 df yields the following cube plot for predicted yield:



If this model is correct, both low acid strength and shorter time are best when the process is run at 70°C. Note, however, that the predicted yield of 6.83 seems too optimistic, since no runs performed this well. If one were searching for still greater yields, then it seems reasonable to shift the experimental region in this direction and to experiment further with these factors.

It is sometimes the case that these methods differ even more in both their estimate for σ^2 and in the number of terms that are statistically significant. Choosing models without regard for statistical significance will surely lead to MSEs that underestimate σ^2 . (See blunder-to-avoid #3 in Section 14.7.) To lessen the possible downward bias estimating σ^2 using a reduced model's MSE, we adopt the following conventions:

- Restrict final models to be hierarchical. For the isatin data, the MSE for Method 2 would have been even smaller than 0.024 if x_2 had been excluded due to its large p -value.
- Always include main effect terms when analyzing full factorial designs. Daniel (1959, p. 317) offered the following advice for constraining the use of negligible terms as part of an error variance estimate:

Nominate all effects, interactions and block contrasts¹ that are thought likely to be important *before* the experiment is completed. The corresponding contrasts are then to be excluded from further arbitrary allocation to error... Only those not nominated will be studied for possible allocation to error.

Presumably prior to the isatin experiment, all four main effects were considered somewhat likely to be important. If so, then the variation explained by the x_3 term should not be pooled with error.

Fitting a model with all main effects ensures that practitioners are not misled by computer output reporting lack-of-fit tests when there is no replication. For the isatin model chosen under Method 2, some software would partition the error variation and report a lack-of-fit test with 2 df for lack-of-fit and 8 df for pure error. Since there is no replication, there can be no pure error. However, when a model with only three factors is fit to a 2^4 experiment, the software views the data as a replicated 2^3 . Such confusion is avoided if one always includes the main effects. Further, retaining all the main effects in the model documents explicitly the relative unimportance of factors with negligible coefficients. Alternatively, JMP allows the user to designate a factor as “excluded,” so that although not appearing in the model, it is recognized as a factor in the experiment. By this feature we may avoid spurious lack-of-fit tests.

With Methods 1 and 2, the MSE is used to estimate the error variance, and so the Student's t distribution is used to compute p -values for tests of individual coefficients. How to conduct tests based on Lenth's PSE will be addressed in the next subsection.

¹The term “contrast” refers to a linear combination of the observations for which the coefficients sum to zero; i.e., the sum $\sum_{i=1}^N c_i y_i$ is a contrast if $\sum_{i=1}^N c_i = 0$. All main effect and interaction columns correspond to contrasts; see Table 1.4.

2.4.2 Tests for effects using Lenth's PSE

The previous subsection introduced the use of Lenth's PSE as a means of estimating the error variance without any replication, provided a majority of the true coefficients are zero. The steps in computing PSE are as follows:

- Determine s_0 , 1.5 times the median absolute estimate from a saturated model fit to the ± 1 coded factors.
- Exclude any estimates that exceed $2.5s_0$ and recompute the median.
- Compute $\text{PSE} = 1.5 \times \text{median}$ (of estimates less than $2.5s_0$ in magnitude).

The logic behind this estimator is as follows. Suppose no effects are present so that $E(b_i) = 0$ and $\text{Var}(b_i) = E(b_i^2) = \sigma^2/N$. Then one could use the average square of the b_i 's to estimate σ^2/N . The median might also be used as an estimator that is more robust to outliers (i.e., to actual effects). Rather than using the median of the b_i^2 's in an estimate for σ^2 , Lenth (1989) proposed using the median of the $|b_i|$'s to estimate a multiple of σ . Since approximately half of a normal distribution with a mean of zero is between $-\sigma/1.5$ and $\sigma/1.5$, and the other half is further from the mean, s_0 is an initial rough estimate for the standard deviation of the b_i 's. By excluding estimates that are more than $2.5s_0$ in magnitude, we eliminate estimates that appear to represent true effects. The remaining set of estimates is thus more nearly purged of estimates corresponding to β 's that are not zero. Even if we compute the median from a list of estimates corresponding to effects, most of which are zero but with a few nonzero, the robustness of the median to outliers ensures that the PSE will not be greatly biased.

For cases with no error degrees of freedom, statistical software will often offer the option of computing Lenth t statistics as b_i/PSE . Percentiles of the sampling distribution of these statistics under the null hypothesis of no effects were estimated via simulation by Ye and Hamada (2000). The first part of Appendix C contains these IER critical values for Lenth t statistics. IER stands for “individual error rate,” since these critical values (c_{α}^{IER}) are computed to limit the probability of a Type I error for each individual test across the set of tests. Occasionally in this book, we provide p -values, computed by simulation using JMP or the code in Appendix C, when analyzing unreplicated experiments via Lenth's procedure. For those wishing to conduct tests for a specified level α , simply use the IER critical values in Appendix C. Simulation is used to obtain critical values and p -values, since attempts at approximating the distribution of Lenth t statistics with a Student's t distribution have not achieved sufficient accuracy (Edwards and Mee 2008).

Consider again the example of Davies (1954). Table 2.7 gives the estimates for the saturated model, the PSE, Lenth t statistics, and p -values obtained by simulation. For software that does not furnish these p -values, an approximation for each p -value can be obtained using Appendix C. For instance, from the IER table in Appendix C we know that b_1 , with Lenth $t = -1.672$ has p -value slightly above .10, since $c_{.10}^{\text{IER}} = 1.701 > 1.672$.

Table 2.7. Estimates and p -values based on Lenth PSE for isatin data

Term	Estimate	PSE	Lenth t	p -Value
Intercept	6.3819	.0572		
x_1	-0.0956	0.0572	-1.672	.103
x_2	-0.0106	0.0572	-0.186	.861
x_3	-0.0381	0.0572	-0.667	.500
x_4	0.1369	0.0572	2.393	.037
$x_1 * x_2$	-0.0006	0.0572	-0.011	.992
$x_1 * x_3$	0.0169	0.0572	0.295	.783
$x_2 * x_3$	-0.0331	0.0572	-0.579	.598
$x_1 * x_4$	-0.0806	0.0572	-1.410	.160
$x_2 * x_4$	-0.1256	0.0572	-2.197	.048
$x_3 * x_4$	-0.0131	0.0572	-0.230	.828
$x_1 * x_2 * x_3$	0.0744	0.0572	1.301	.190
$x_1 * x_2 * x_4$	-0.0506	0.0572	-0.885	.354
$x_1 * x_3 * x_4$	-0.0031	0.0572	-0.055	.960
$x_2 * x_3 * x_4$	0.0619	0.0572	1.082	.265
$x_1 * x_2 * x_3 * x_4$	0.0094	0.0572	0.164	.876

To understand better the performance of Lenth's t statistics versus IER critical values, consider the case for $\alpha = .05$ and 15 estimates, where the critical value is $c_{.05}^{\text{IER}} = 2.156$; that is, we test all 15 estimates and declare ones larger in magnitude than 2.156(PSE) to be statistically significant. Since $\alpha = .05$ and $15(.05) = 0.75$, we expect, on average, 0.75 effects to be declared statistically significant, if in fact all true coefficients are zero.

To illustrate this, one million sets of 15 normal random variables with zero means were simulated. From each set, the PSE was calculated and the number of "estimates" found to exceed 2.156(PSE) was determined. The resulting distribution was as follows:

No. of Significant Effects Found	Frequency	Freq./10 ⁶
0	604,881	0.6049
1	208,926	0.2089
2	94,398	0.0944
3	46,574	0.0466
4	23,907	0.0239
5	12,643	0.0126
6	6,130	0.0061
7	2,426	0.0024
8	94	0.0001
9	18	0.0000
10	3	0.0000

This distribution has a mean of 0.75, as required by using $\alpha = .05$ for 15 tests. Note that the risk of making one or more type I errors is $1 - 0.6049$

$= 0.3951$, or nearly 40%. This larger risk is called the experimentwise error rate (EER). It is informative to report both the individual error rate and the experimentwise error rate for a test procedure. Thus, for 2.156, the individual error rate is 0.05, whereas the experimentwise error rate is 0.395.

To control the experimentwise error rate, one may use the c_{α}^{EER} critical value table in Appendix C or those provided by Ye and Hamada (2000). These values were obtained by simulating $\max\{|b_1|, |b_2|, \dots, |b_m|\}/\text{PSE}$ under the null hypothesis of no effects. For instance, for 15 contrasts and $\alpha = .10$, the EER critical value from Appendix C is $c_{10}^{\text{EER}} = 3.505$. (Its individual error rate is about 0.011, and the expected number of Type I errors is $15(0.011) = 0.17$ when using the critical value 3.505.) In Table 2.7, no Lenth t statistics exceed 3.505; the largest is 2.393, which corresponds to an experimentwise error rate of 0.29 (i.e., if no true effects were present, nearly 30% of the time, one would obtain a largest Lenth t of 2.393 or greater). In experiments of this size, often controlling the individual error rate offers sufficient protection. However, when an experiment contains 2^6 or more treatment combinations, the number of eligible terms becomes so large that either controlling the experimentwise error rate or using a smaller α (e.g., .01) for IER is reasonable.

2.4.3 Alternatives to Lenth's t test

Hamada and Balakrishnan (1998) compared two dozen test procedures for unreplicated factorial designs. Most of these methods are intended to control the IER for each test. Lenth's method using IER critical values is one of the simplest, and it performs satisfactorily in terms of power. In Section 14.2, other more recent alternatives are discussed briefly.

Lenth's method, as originally proposed, is not the best for controlling the EER. A step-down version for Lenth's method proposed by Ye, Hamada, and Wu (2001) is certainly preferable. Section 14.2.1 illustrates this method and discusses some other alternatives for controlling the experimentwise error rate, including a simple step-up approach that utilizes standard F statistics for backward elimination regression. Section 14.2.2 makes the case that controlling EER is not usually of practical interest and argues for the intuitive alternative of controlling the proportion of Type I errors among all effects declared significant.

Finally, for any procedure such as Lenth (1989) based on the assumption of effect sparsity, be sure to fit a saturated model, since the method is based on the preponderance of negligible estimates. If one fits less than a saturated model, there will exist error degrees of freedom and software will use the MSE in constructing t -tests instead of the PSE, even if there is just 1 df for error. For 2^k factorial designs with no replication except at the center, most software will ignore the PSE. When the pure error degrees of freedom are very small and the sparsity of effects assumption is reasonable, then it is prudent to combine the MSE with the estimate for σ^2 that comes from Lenth's PSE. Section 14.1 presents two means for doing so.

2.5 Normal Plot of Effects and Other Analysis Tools

2.5.1 Normal and half-normal plot of effect estimates

Long before Lenth (1989) promoted the testing for effects in unreplicated experiments based on the sparsity of effects principle, Daniel (1959, 1976) and others urged that the effect estimates be plotted. If the sparsity of effects assumption is true, then for a 2^k factorial design, the majority of estimators for the coefficients in the saturated model (1.4) follow a normal distribution with mean 0 and variance $\sigma^2/2^k$. The $m = 2^k - 1$ estimates (excluding the intercept) are ordered from smallest to largest and plotted versus the standard normal quantiles Z_{P_i} ($i = 1, \dots, m$), where we use Blom's (1958) recommended proportions

$$P_i = (i - 0.375)/(m + 0.25). \quad (2.3)$$

For example, with $m = 15$, the largest estimate is plotted versus $Z_{14.625/15.25} = 1.739$ and the smallest estimate versus -1.739 . In the normal plot of estimates, most are expected to fall along a line with an intercept of zero and (unknown) slope of $\sigma/N^{1/2}$, where, here, $N = 2^k$. For instance, Figure 2.2 shows the plot of the 15 estimated factorial effects from Table 2.7. The fitted line was constrained to have an intercept of zero and a slope equal to Lenth's PSE = 0.0572. The fact that a few of the 15 estimates fall below the line on the left and above the line on the right is weak evidence that these estimates correspond to effects that are present (i.e., $\beta_s \neq 0$). The closer the estimates fall along the line, the more consistent the data are with an assumption of no true effects.

Since the statistical significance of an estimate is generally based on its size $|b_s|$, a half-normal plot is seen as more useful than a normal plot by some (see Daniel 1959). For a half-normal plot of effects, we sort the absolute values of the estimates from smallest to largest and plot these versus the standard normal quantiles Z_{Q_i} ($i = 1, \dots, m$), where

$$Q_i = 0.5 + (i - 0.055)/(2m + 1.2). \quad (2.4)$$

Use of the proportions (2.4) was determined empirically and appears to be more accurate than Daniel's choice of $0.5 + (i - 0.5)/2m$. For a more accurate closed-form approximation of half-normal order statistic expected values, see Olguin and Fearn (1997, p. 460). A half-normal plot for the estimates in Table 2.7 is given in Figure 2.3. This plot reveals even more prominently the possibility of two or more active effects.

Statistical software such as JMP and Minitab automates the plotting of effects as in Figures 2.2 and 2.3, labeling the larger estimates. Such software may use different formulas than (2.3) and (2.4), resulting in slight differences in the appearance of the plots. For example, JMP 7.0 uses $P_i = i/(m+1)$, which results in less extreme Z_{P_i} , whereas MINITAB 14 provides several options, with $P_i = (i - 0.3)/(m + 0.4)$ as the default.

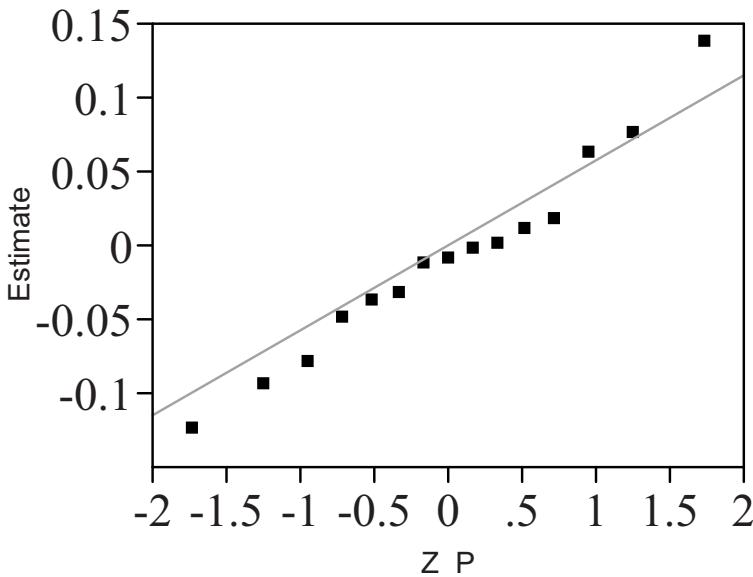


Fig. 2.2. Normal plot of effect estimates in isatin experiment

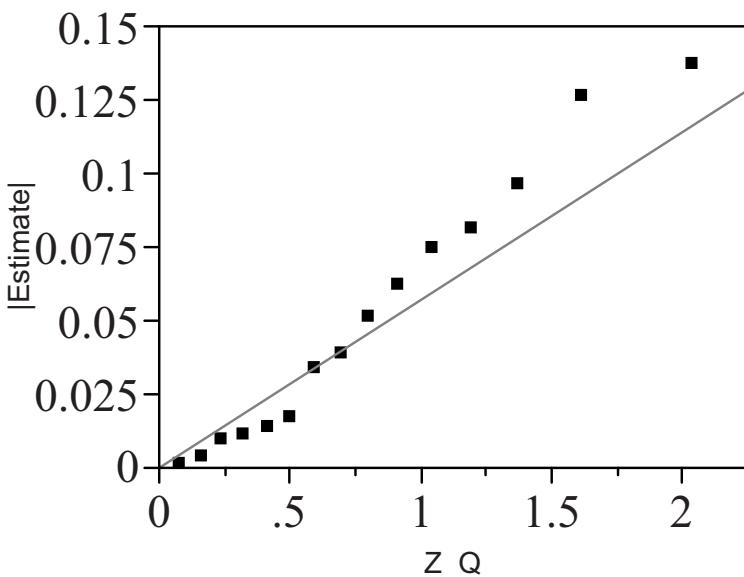
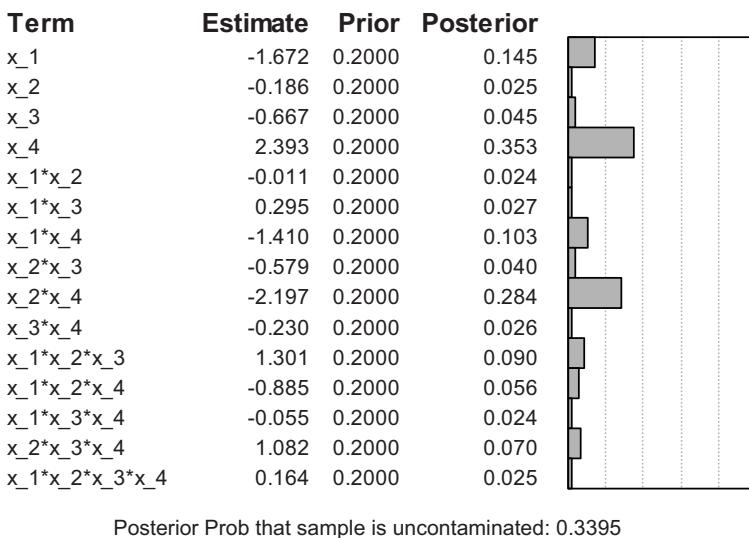


Fig. 2.3. Half-normal plot of effect estimates in isatin experiment

2.5.2 Bayesian analysis of 2^k factorial designs

Box and Meyer (1986) proposed an intuitive method for analyzing unrelicated 2^k experiments that does not involve tests of significance. Rather, it supposes that some fraction $1 - \alpha$ of the effects are zero and that the remaining proportion α come from a normal distribution with a variance that is substantially larger than σ^2/N , the error variance for the estimates. A Bayesian framework combines these prior assumptions with the data and produces a “posterior” probability for each effect that it comes from the subset of active effects. For the isatin 2^4 experiment, the posterior probabilities as computed by JMP 7.0 are displayed in Figure 2.4.



Posterior Prob that sample is uncontaminated: 0.3395

Fig. 2.4. Posterior probabilities for effects in isatin experiment; $K = 10$, $\alpha = .20$

Note that the estimates reported by JMP in Figure 2.4 are actually Lenth t statistics (refer to Table 2.7). For 13 of the 15 estimates, the posterior probability is lower than the prior probability of .2. The larger the magnitude of the estimate, the larger the posterior probability. However, even the largest estimate ($b_4 = 0.13$ with Lenth $t = 2.393$) only has a posterior probability of .35. Thus, although there is some evidence for one or two effects, that evidence is not compelling.

Recall that using an individual error rate of 0.05, the two largest estimates are statistically significant, whereas controlling the EER at 0.25 or smaller, no effects are statistically significant. The posterior probabilities in Figure 2.4 were obtained assuming 20% of effects are active and that they have a variance 10 times σ^2/N (since JMP’s $K = 10$). Since the largest estimates

were not much larger than the others, even the largest estimates are deemed more likely to correspond to null effects. However, if we lower K to 5, the posterior probabilities for b_4 and $b_{2,4}$ increase to 0.59 and 0.51, respectively. The conclusion is still that the evidence for these effects being active is rather weak, given a prior expectation that $3/15 = 20\%$ of the effects would be active. For more details on the computations, see Box and Meyer (1986, 1993). For a comparison of the Bayesian approach with Lenth's method, see Haaland and O'Connell (1995).

2.6 Diagnostics for a Fitted Model

The residual e_i is the difference between the i^{th} observed response y_i and the corresponding predicted value \hat{y}_i from a fitted model—that is, $e_i = y_i - \hat{y}_i$ ($i = 1, \dots, N$). Residuals indicate the extent of disagreement between an assumed model and the actual data, and so provide a means of checking both the tentative model for $E(y)$ and the assumptions regarding ϵ .

2.6.1 Plotting residuals versus predicted y

Plotting e_i versus \hat{y}_i is particularly helpful for assessing model adequacy, provided there are enough error degrees of freedom. (The error degrees of freedom indicate the amount of information in the residuals.) To illustrate this point, consider several residual plots for the Bouler et al. (1996) data discussed in Section 2.3. Figure 2.5a displays residuals versus predicted values for the full factorial model as summarized in the Table 2.4 ANOVA. Here, we have only 7 df for error: 1 df for lack-of-fit and 6 df for pure error. This residual plot is not very useful, since it simply shows the pure error variation among the centerpoint runs and the statistically significant lack-of-fit due to the centerpoint residuals being predominantly positive. (If one were to fit a saturated model by adding the term x_1^2 , then the residuals for all the factorial points would be zero, and the residuals for the center runs would average zero. Plotting these residuals would have no value.)

Consider a second residual versus predicted plot based on a reduced model that eliminates all interactions involving x_3 or x_4 . With 22 interactions removed, this model has 29 df for error, and its residuals are displayed in Figure 2.5b. This residual plot is more useful. First, it shows that the pure error variation at the center is small compared with the variation in other residuals, so that the lack-of-fit for this reduced model must be statistically significant. In addition, there is more variation in compressive strength when the expected compressive strength is above 30. Finally, residuals corresponding to the 10 observations with zero compressive strength have predicted values ranging from -3.4 to 6.7 . Overall, this residual plot reflects an unsatisfactory model. In Section 2.7 we will discuss how using a transformation for y can improve the model fit in such occasions. For instance, if we fit the same model (with

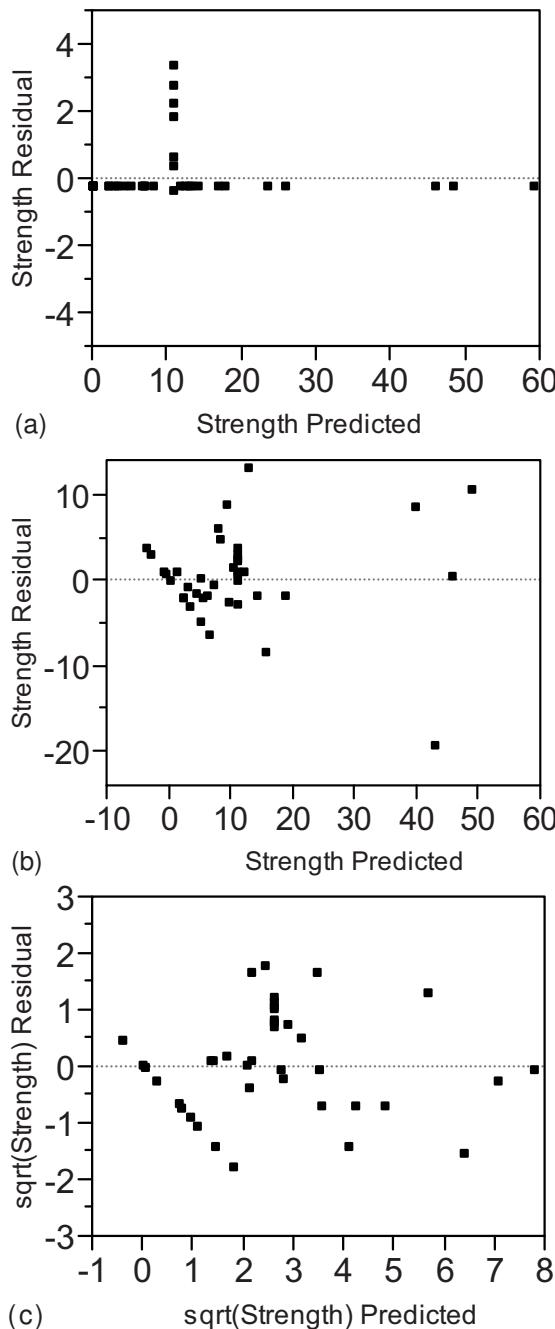


Fig. 2.5. Residual plots for (a) full factorial model for $y = \text{strength}$; (b) reduced model for $y = \text{strength}$; (c) reduced model for $y = (\text{strength})^{1/2}$

29 df for error) to the square root of compressive strength, as seen in Figure 2.5c, the variation in the residuals appears more consistent across the range of predicted values and only 1 of the 10 observations with $y_i = 0$ has a negative predicted value (-0.4).

Residual versus predicted plots are particularly helpful for spotting violations of the assumption of constant $\text{Var}(\epsilon)$. When there is replication at several or all of the treatment combinations, there exist tests for checking the equality of $\text{Var}(\epsilon)$. Common tests available in software include Bartlett's test for equality of variances and the more robust tests by O'Brien (1979, 1981).

2.6.2 Plotting residuals versus run order

Plotting residuals versus \hat{y}_i is only one of several useful means for examining the residuals. When the data are time (or spatially) ordered, it is important to plot the residuals versus that order. Such a plot is displayed in Figure 11.2, where a possible shift in the measurement process is revealed. Autocorrelation of the errors is another possibility related to time-ordered experimental runs. Use the Durbin–Watson statistic to check for first-order autocorrelation. When applying this test using any software, be sure to have the data sorted by run order. The Durbin–Watson test is more important when one is experimenting with a highly variable process where such correlation is deemed likely. Although randomization of run order does not eliminate trend or autocorrelation for ϵ , it does offer protection against the effects of such problems in most situations (Box, Hunter, and Hunter 2005, pp. 404f). For further discussion, see Section 13.5.

2.6.3 Plotting residuals in normal quantile plot

When there is a large amount of data and R^2 is low, then the distribution of ϵ becomes important. In such cases, one may construct a normal quantile plot of the residuals. For instance, see Figure 4.9. By contrast, when R^2 is above 90%, the distribution of ϵ has minimal importance, since the distribution of the residuals will reflect lack-of-fit more than it will the actual distribution of ϵ . For this reason, we do not routinely construct a normal plot of residuals for examples in this book.

2.6.4 Identifying outliers using Studentized residuals

Spurious y_i values are a serious concern, especially for small, unreplicated experiments, because of their influence on the fitted model. However, an observation that appears to be an outlier under one model may appear reasonable under a different model. For instance, the large negative residual displayed in Figure 2.5b is problematic if the error variance is constant. However, if the error variation increases as strength increases, then the same observation no

longer appears extreme. The less data one has, the more ambiguity exists regarding how to interpret such runs. A simple, practical approach to handling suspected outliers is to fit models both with and without the runs, to see their impact on the conclusions. Daniel (1959, pp. 331f) pointed out that for two-level factorials, a single outlier will bias every effect estimate by (\pm) the same amount and that this will alter the half-normal plot of effect estimates to have no clump of estimates at zero. The case study in Section 4.2 will illustrate how to address the problem of more outliers.

Literature about outliers in regression is extensive. Beckman and Cook (1983, Section 4.2) provided an excellent overview; see also Gray and Woodall (1994). The Studentized residual is defined as

$$r_i = e_i / [(1 - h_{ii}) \text{MSE}]^{1/2}, \quad (2.5)$$

where h_{ii} is the $(i, i)^{\text{th}}$ element of the “hat” matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. For a 2^k factorial (with $n \geq 1$ observations at each treatment combination and no centerpoint runs), $h_{ii} = r/N$, where r is the number of columns in X .

The distribution of the maximum Studentized residual can be simulated for any model matrix \mathbf{X} . Appendix D provides a simple simulation program that may be used to determine upper (10th and 5th) percentiles for the maximum (in absolute value) Studentized residual, and the probability of getting a maximum residual as large as that observed for a particular model. This provides a quick reference regarding whether any observations may be considered outliers. Gray and Woodall (1994) showed that the maximum value for (2.5) is $(N - r)^{1/2}$. When the degrees of freedom for error are 4 or less, there is no point checking for extreme outliers.

If the i^{th} observation is deleted and the same model fit to the data, the error sum of squares will decrease by $e_i^2/(1 - h_{ii})$ and the i^{th} “deleted residual,” the difference between y_i and the predicted value for the omitted observation, is

$$d_i = e_i / (1 - h_{ii}). \quad (2.6)$$

The Studentized version of (2.6) is the same as (2.5), except that the estimate for σ^2 is based on $\text{MSE}_{(-i)}$, the mean square error with the i^{th} observation excluded, which is

$$\text{MSE}_{(-i)} = [\text{SSE} - e_i^2 / (1 - h_{ii})] / [N - r - 1], \quad (2.7)$$

where SSE is the error sum of squares, $\sum_j^N e_j^2$. The most convenient means for computing (2.6) is to add a dummy variable column to the \mathbf{X} matrix with the value 1 for the i^{th} row and 0 otherwise. The regression coefficient for this column will equal the deleted residual (2.6), the MSE will equal (2.7) and the t statistic for the coefficient of the dummy column will equal the Studentized residual for the i^{th} observation. Typically, an observation attracted attention simply because it had one of the largest residuals. Based on the Bonferroni inequality, one may multiply the p -value by N to get the approximate probability that the biggest residual would be larger than this just by chance.

We now illustrate these computations with a small example to reinforce the concepts. Suppose we fit the isatin yield data from Table 2.6 with a model containing all four main effects and the x_2*x_4 interaction and consider whether any residual is unusually large. (Perhaps that is the reason we saw few significant terms.) The largest residual is for the fifth observation, with $y_5 = 6.12$ and $\hat{y}_5 = 6.4175$. With $MSE = 0.3212/10$,

$$e_5 = y_5 - \hat{y}_5 = 6.12 - 6.4175 = -0.2975,$$

$$r_5 = e_5 / [(1 - h_{55})MSE]^{1/2} = -0.2975 / [(1 - 6/16)0.03212]^{1/2} = -2.0997.$$

This is not unusually large. Using the simulation program in Appendix D, we determine that there is a 39.5% chance of getting a Studentized residual this far from zero.

If one were to delete the fifth observation and refit the model, the predicted value for this observation is 6.596 and

$$d_5 = 6.12 - 6.596 = -0.2975 / (1 - 6/16) = -0.476.$$

If instead of deleting this observation, one adds a dummy variable for the fifth observation, the estimated model becomes

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	0.412	0.0370	11.113	.000
x_1	-0.125	0.0370	-3.385	.008
x_2	0.019	0.0370	0.516	.618
x_3	-0.008	0.0370	-0.226	.826
x_4	0.107	0.0370	2.892	.018
x_2*x_4	-0.155	0.0370	-4.195	.002
Dummy ₅	-0.476	0.1787	-2.664	.026

with mean square error

$$\begin{aligned} MSE_{(-5)} &= [0.3212 - (-0.2975)^2 / (1 - 6/16)] / [16 - 6 - 1] \\ &= 0.1796 / 9 = 0.0200. \end{aligned}$$

The standardized deleted residual for y_5 is -2.664 (p -value = .026). However, a p -value as small as .026 is typical for the most extreme outlier. Multiplying by $N = 16$, we obtain $16(.026) = 0.414$; this Bonferroni upper bound is only slightly larger than the exact probability of .395 found using the Appendix D simulation. Assuming that this model is correct, there is no indication of any outliers among these data.

In Chapter 4, we analyze case studies in which many outliers will be evident.

2.7 Transformations of the Response

Example 2.3. Drill Advance Rate for 2⁴

Daniel (1976) introduced the use of transformations for y in a section titled “Looking for Simple Models.” His 2⁴ example involving the advance rate of a stone drill illustrates clearly the potential advantages. The data are displayed in Figure 2.6.

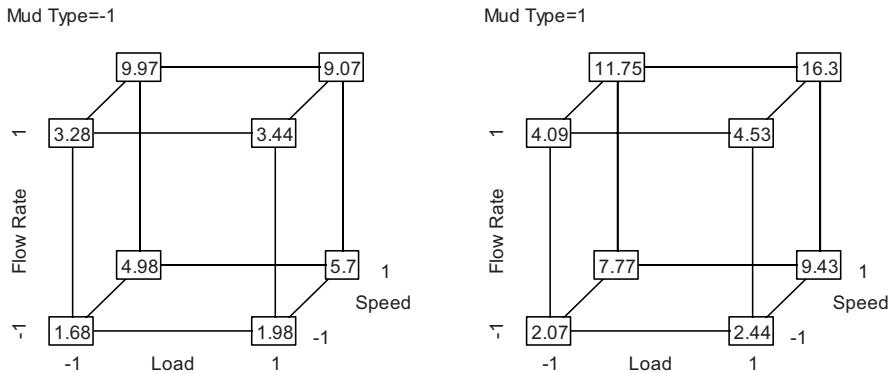


Fig. 2.6. Cube plot of Example 2.3 advance rate data from Daniel (1976)

Fitting a saturated model to these data, we obtain a half-normal plot of the effects (see Figure 2.7). This plot is pleasing, in that three of the main effects have statistically significant estimates, based on their Lenth t statistics. A reduced model would certainly contain these three terms and possibly the $x_{\text{Speed}} * x_{\text{Flow}}$ and $x_{\text{Speed}} * x_{\text{Mud}}$ interactions, since their estimates also stand off the line. The resulting model,

$$\hat{y} = 6.15 + 1.65x_{\text{Flow}} + 3.22x_{\text{Speed}} + 1.14x_{\text{Mud}} + 0.75x_{\text{Flow}} * x_{\text{Speed}} + 0.80x_{\text{Flow}} * x_{\text{Mud}}, \quad (2.8)$$

explains 95% of the variation in advance rate. However, the normal plot of effects for the saturated model (Figure 2.8) looks peculiar in that all 15 estimates are positive, so that the estimates are far from the line through the origin; there is no clump of estimates centered about 0. In addition, the residuals from the reduced model (2.8) are more scattered at large predicted advance rates (see Figure 2.9). All of these plots indicate that we are missing some systematic variation with our model, even though $R^2 = .95$.

Daniel (1976) fitted models for nine different transformations of $y = \text{advance rate}$, including different powers of y , and the log transformations

$$\ln(y + c) \quad (2.9)$$

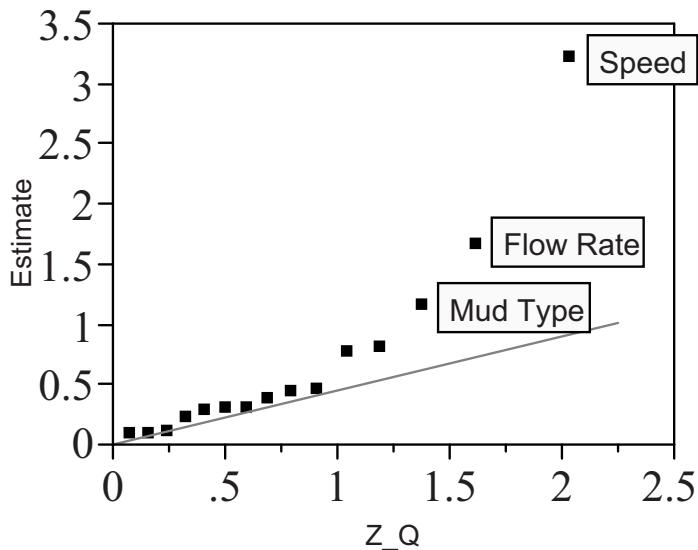


Fig. 2.7. Half-normal plot of effects for Daniel's drill data

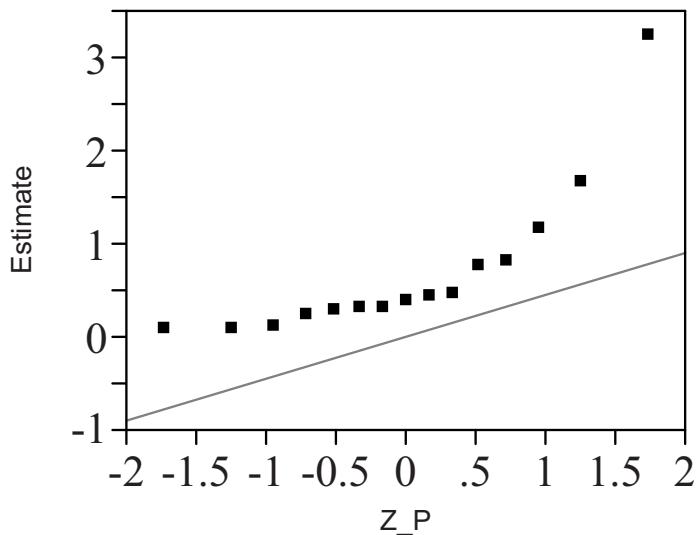


Fig. 2.8. Normal plot of effects for Daniel's drill 2^4

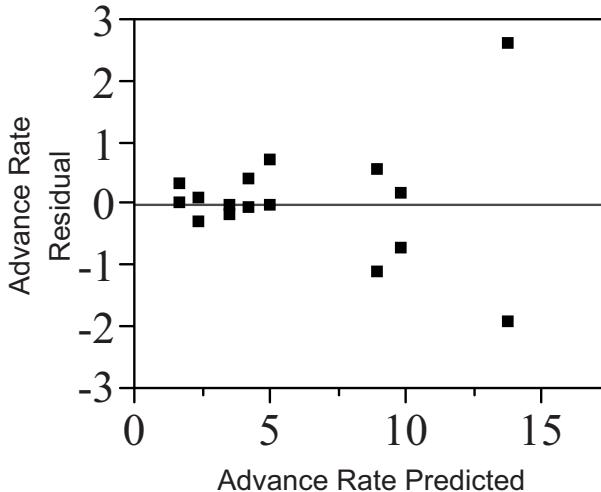


Fig. 2.9. Residuals versus predicted advance rate for reduced model (2.5) for Daniel's drill 2⁴

with different constant shifts c . The family of transformations (2.9) is valid, provided $c > -\min\{y_1, \dots, y_N\}$. The most popular set of transformations today is the family proposed by Box and Cox (1964):

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/(\lambda \dot{y}^{\lambda-1}), & \text{if } \lambda \neq 0 \\ \ln(y) \dot{y}, & \text{if } \lambda = 0 \end{cases} \quad (2.10)$$

where \dot{y} is the geometric mean $\dot{y} = \prod_{i=1}^N y_i^{1/N}$. By normalizing the power transformation as in (2.10), the value of λ for which the error sum of squares is minimized is the maximum likelihood estimator for λ . Since the normalized transformation nearly makes the total sum of squares for $y^{(\lambda)}$ invariant to λ , the λ that minimizes the SSE essentially maximizes R^2 .

Suppose we fit an additive model in the four factors for advance rate. A plot of the error sum of squares for different transformations $-2 \leq \lambda \leq 2$ as produced by JMP is shown in Figure 2.10. JMP searches the grid $\{-2, -1.8, -1.6, \dots, 2\}$ and determines that transformed values corresponding to $\lambda = 0$ have the smallest error sum of squares for the additive model.

The actual maximum likelihood estimator here is $\hat{\lambda} = -0.05$, but taking $\lambda = 0$ is simpler and produces essentially the same result. A 95% confidence interval for λ is the interval of values that produce an error sum of squares below the horizontal line in Figure 2.10. For details on the computation, see Box and Cox (1964) or Montgomery and Peck (1992). In Figure 2.10 this confidence interval is narrow for two reasons. First, the ratio $\max\{y_i\}/\min\{y_i\} = 9.7$. When this ratio is less than 2, nonlinear transformations will have lit-

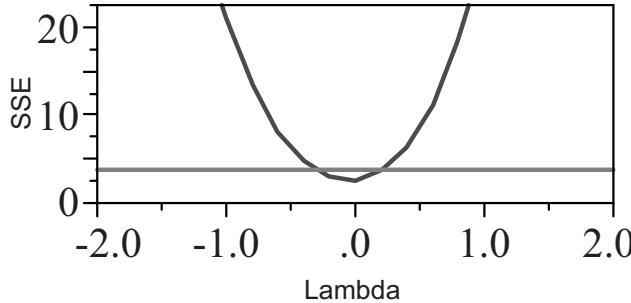


Fig. 2.10. Error sum of squares from additive model for different transformations of advance rate, $-2 \leq \lambda \leq 2$

tle effect on the result, and we will be indifferent to models for a wide range of λ . By contrast, when the maximum is an order of magnitude larger than the minimum, nonlinear transformations have a pronounced effect, and so some transformations are clearly better than others. Second, we considered different transformations for the additive model, which leaves much of the variation unexplained. If one were to choose a model with more terms, then many different λ 's may explain most all the variation, and so again the choice for the best λ will not be so clearly indicated. For instance, if the Box–Cox transformation is applied fitting the two-factor interaction model to Daniel's data, the confidence interval for λ is $(-1.23, 0.05)$; from this fit, either the log or reciprocal transformation is acceptable. We prefer the log transformation because the histogram for $\ln(y_i)$ is less skewed than the histogram for $1/y_i$. In addition, the resulting model matches the engineering expectation that the factor effects might be multiplicative.

The fitted additive model for predicted $\ln(\text{advance rate})$ is:

$$\widehat{\ln(y)} = 1.600 + 0.065x_{\text{Load}} + 0.290x_{\text{Flow}} + 0.577x_{\text{Speed}} + 0.163x_{\text{Mud}}. \quad (2.11)$$

Taking the exponential of (2.11) produces an estimate for the median (not the mean) advance rate:

$$e^{\widehat{\ln(y)}} = 4.953(1.067)^{x_{\text{Load}}} (1.336)^{x_{\text{Flow}}} (1.781)^{x_{\text{Speed}}} (1.177)^{x_{\text{Mud}}}$$

since $e^{1.6} = 4.953$, $e^{0.065} = 1.067$, etc. The predicted median rates range from 1.66 to 14.8.

In Daniel's drill example, the simple additive model in $\ln(\text{advance rate})$ accounted for 98.5% of the variation, whereas modeling advance rate directly would have required a model with many terms to achieve an R^2 so large. There are additional reasons for considering transformations. First, if there is substantial error variation and the variance is not constant, then ordinary least squares estimation loses efficiency. When the error variation depends on

$E(y)$, then choosing a suitable function f and modeling $f(y)$ rather than y directly can resolve the unequal variance problem and keep the estimation simple. This is the case for the Bouler et al. (1996) data; recall the improved residual plot in Figure 2.5c for $y = (\text{strength})^{1/2}$.

In this section, we have addressed applications in which the choice of a transformation $f(y)$ is determined empirically. Sometimes the nature of the response y suggests what transformation is appropriate. For example, when y is a count, following a Binomial or Poisson distribution, known transformations will stabilize the variance (see Sections 2.8.1 and 2.8.2). Another common response is the standard deviation. Section 2.8.3 details why the log transformation is appropriate for variances and standard deviations.

2.8 Analysis of Counts, Variances, and Other Statistics

For some experiments, responses are counts. For example,

- number of flaws in a door panel
- number of respondents to an email solicitation
- number of defective parts in a sample of 20

Counting the number of good (or bad) parts is not as informative as collecting quantitative data on each part. For example, it is better to measure the breaking strength on each of a sample of parts than it is to simply know how many failed at a certain stress. However, in some applications, quantitative data are either too expensive or impossible to collect and count data are all that are available. Count data routinely violate the assumption of constant variance for ϵ , and so specialized methods are required. The simplest of these methods is to use least squares for a transformation of the response. When the sample sizes at each treatment combination are large, use of least squares is often justified. For cases where sample sizes are smaller, other methods are recommended. After discussing and illustrating the options for count data, we discuss the common case of modeling a variance and then briefly mention analyzing correlations, ratios, lifetimes, directions, and functional responses.

2.8.1 Modeling Binomial proportions

When the measured outcome at a treatment combination is the proportion of n trials having a characteristic of interest, the Binomial distribution is generally appropriate. Let c denote the number of cases having the characteristic of interest and let $\hat{p} = c/n$ denote the observed proportion. If the outcomes of the individual trials are independently distributed and the number of trials is fixed, then c has a Binomial distribution with parameters n and p , where $p = E[\hat{p}]$. There are two problems associated with modeling \hat{p} . First, since the variance of \hat{p} depends on p , any factor that affects the mean also affects the

variance. Thus, the typical constant variance assumption will be violated. Second, since $0 \leq p \leq 1$, a fitted model for p may result in predicted proportions outside the feasible range.

$\text{Var}(\hat{p}) = p(1 - p)/n$ is maximum at $p = .5$ and is relatively stable over the interval $.3 \leq p \leq .7$. However, for problems where the proportions are not confined to this range and where least squares estimation is to be used, it is best to model some function of \hat{p} that stabilizes the variance. One option is

$$f_a(\hat{p}) = \arcsin(\sqrt{\hat{p}}).$$

Figure 2.11 shows how this function is essentially linear over the range $.3 \leq \hat{p} \leq .7$, but it accentuates differences among more extreme values for \hat{p} , where \hat{p} is less variable.

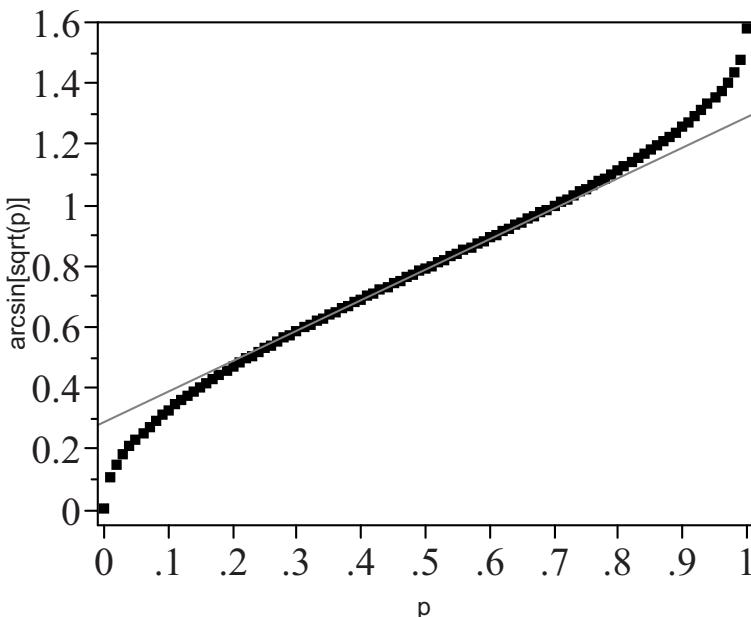


Fig. 2.11. $f_a(p) = \arcsin(\sqrt{p})$ transformation; slope ≈ 1 for $.3 \leq p \leq .7$

Freeman and Tukey (1950) recommended a modification to the transformation $f_a(\hat{p})$. The Freeman–Tukey transformation for Binomial proportions is

$$\begin{aligned} f_{\text{FT}}(\hat{p}) &= f_a[\hat{p}n/(n+1)] + f_a[(\hat{p}n+1)/(n+1)] \\ &= \arcsin[\sqrt{c/(n+1)}] + \arcsin[\sqrt{(c+1)/(n+1)}], \end{aligned} \quad (2.12)$$

where c is the number of cases out of n with the characteristic of interest. Note that this transformation depends on both \hat{p} (or c) and n .

Figure 2.12 shows the variance of $f_a(\hat{p})$ and $f_{FT}(\hat{p})$ for $n = 10$ and Figure 2.13 shows the same for $n = 40$. Reference lines are drawn at $1/n$ and $1 - 1/n$ in each figure. In Figure 2.12, $\text{Var}(\hat{p}) = p(1 - p)/n$ is also displayed. For $n = 10$, over the interval $.1 \leq p \leq .9$, $\text{Var}(\hat{p})$ ranges from .009 to .025, a max/min ratio of 2.78. Both variance-stabilizing transformations do much better. For $f_a(\hat{p})$, the max/min ratio is $= 0.04202/0.02857 = 1.45$, whereas the Freeman–Tukey transformation has the max/min ratio of $0.099998/0.0913 = 1.095$. Freeman and Tukey (1950) stated that (2.12) produces variances within $\pm 6\%$ of $1/(n + 0.5)$ for almost all cases where the expected proportion p is between $1/n$ and $1 - 1/n$. This corresponds to data where the expected count is at least 1 and not more than $n - 1$. For $n = 40$, the max/min ratio for the variance over the interval $.025 \leq p \leq .975$ is 1.61 for $f_a(\hat{p})$ and 1.12 for $f_{FT}(\hat{p})$. Use of (2.12) as the response is recommended provided one has few sample proportions of 0 or 1. Given a fitted model for (2.12), the inverse transformation (Miller 1978) is

$$\hat{p}(f) = 0.5\{1 - \text{sgn}(\cos f)[1 - (\sin f + (\sin f - 1/\sin f)/n)^2]^{1/2}\},$$

where f is the predicted value for f_{FT} and $\text{sgn}(\cos f)$ denotes the sign of $\cos f$.

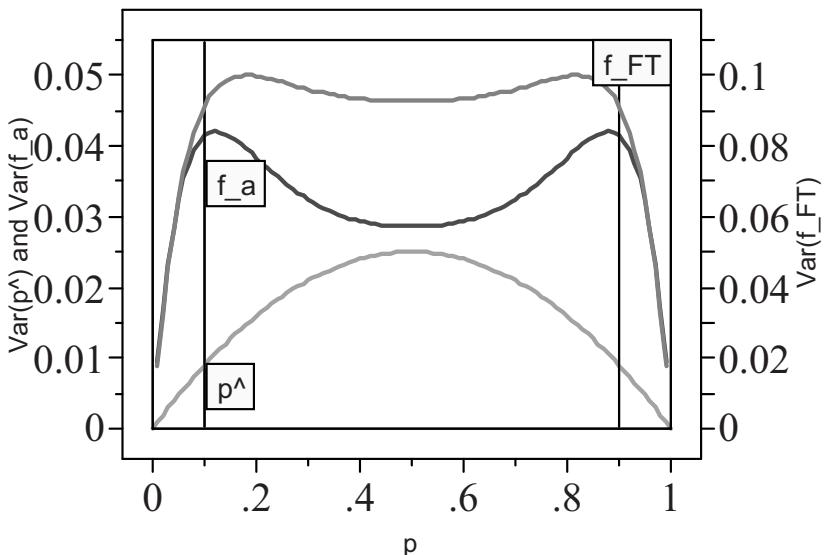


Fig. 2.12. Variance for \hat{p} , $f_a(\hat{p})$, and $f_{FT}(\hat{p})$ for $n = 10$

Arber et al. (2006) conducted a 2^4 factorial experiment to see how gender, age, race, and social class affected physicians' diagnoses and follow-up recommendations for simulated coronary heart disease patients. $N = 256$ physicians

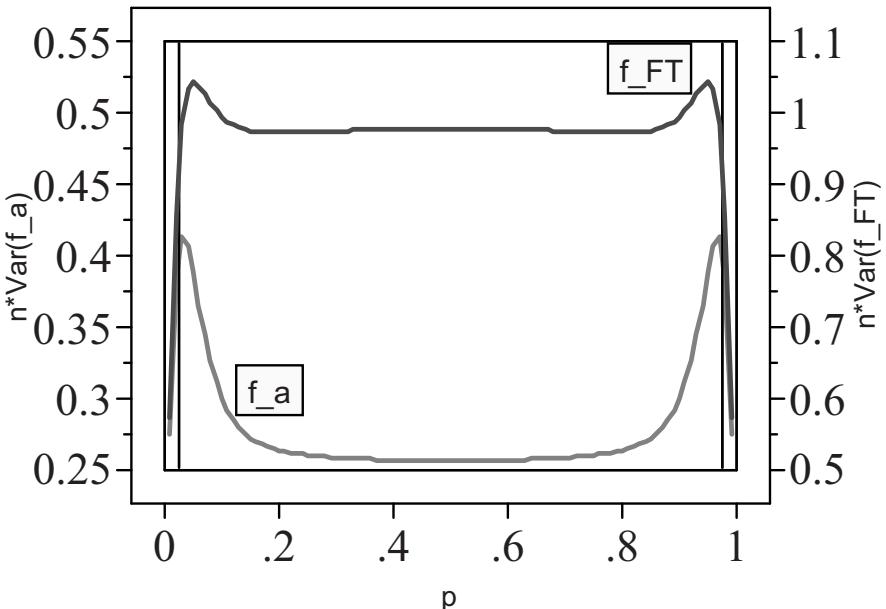


Fig. 2.13. n -Variance for $f_a(\hat{p})$ and $f_{FT}(\hat{p})$ for $n = 40$

took part, with 16 assigned to each treatment combination, 8 from the United States and 8 from the United Kingdom. Thus, for analyses that ignore country, $n = 16$ per treatment combination, and the Freeman–Tukey transformation (2.9) will effectively stabilize the variance at $1/16.5$ for p in the interval (.067, .933). For analyses of individual country data, $n = 8$, and so the range of p for which the variance of the Freeman–Tukey transformed proportions is near $1/8.5$ is constrained to (0.125, 0.875). Since several of the characteristics of interest occurred for over 90% or less than 10% of the doctors, the country-specific data cannot be effectively analyzed using least squares and the response $f_{FT}(\hat{p})$.

When some expected counts are close to zero or n , a linear model for p or $f_{FT}(p)$ may not be suitable. An alternative is to model the log-odds, $\ln[p/(1 - p)]$, with estimation via maximum likelihood rather than least squares. Two advantages of modeling the log-odds are (i) any predicted value for the log-odds corresponds to a value for p within the interval (0,1) and (ii) the models are meaningful to interpret. For instance, the additive model (1.3) translates into a model of independence, whereas models with some interactions are interpretable in terms of conditional independence. Although maximum likelihood estimation may require iteration, the required software is widely available. For a useful reference, see Collett (2002).

2.8.2 Modeling Poisson count data

Example 2.4: 2^4 factorial with $y = \text{number of blemishes}$

Hsieh and Goodwin (1986) described an experiment to reduce the number imperfections in a grille used by a Chrysler assembly plant. Porosity problems caused the blemishes, and a 16-run experiment was performed in search of a remedy. Four factors were Mold pressure (x_1), Priming method (x_2), Thickening process (x_3), and Viscosity (x_4). Pressure and Viscosity are quantitative factors, although the actual levels used were not reported; the other two factors are qualitative. The 16 treatment combinations in the order listed by the authors are shown in Table 2.8, along with the total number of “pop” defects observed for each. The observed counts range from 3 to 99 pops. We are not told whether a single part or multiple parts were inspected at each treatment combination.

Table 2.8. Hsieh and Goodwin (1986) experiment

x_1	x_2	x_3	x_4	Total No. Pops, c
-1	-1	1	-1	66
-1	-1	-1	-1	19
-1	1	1	1	3
-1	1	-1	1	7
-1	1	-1	-1	4
-1	1	1	-1	17
-1	-1	-1	1	99
-1	-1	1	1	5
1	1	1	-1	4
1	1	-1	-1	3
1	-1	1	1	5
1	-1	-1	1	14
1	-1	-1	-1	7
1	-1	1	-1	14
1	1	-1	1	5
1	1	1	1	8

If flaws arise individually and independently, then the data will follow a Poisson distribution; refer to any probability book for details (e.g., Ross 1998). Modeling Poisson counts c using ordinary least squares is not recommended, since any factor that affects the mean also affects the variance. For the Poisson distribution, the mean and variance are equal. Thus, if the factors do affect the mean, then the assumption of constant variance will be violated. Several alternative approaches are more appropriate:

- Use the simple, variance-stabilizing transformation \sqrt{c} . Think of this transformation as taking each observation c and dividing by an estimate of its

standard deviation, \sqrt{c} . If we divided c by its true standard deviation, $\sqrt{E(c)}$, the resulting standardized variable $c/\sqrt{E(c)}$ would have a variance of 1, whatever $E(c)$ is. Because the numerator and denominator of c/\sqrt{c} are correlated, \sqrt{c} has a variance smaller than 1 but one that is insensitive to $E(c)$.

- Use the Freeman and Tukey (1950) transformation for Poisson counts, $FT(c) = (\sqrt{c} + \sqrt{c+1})/2$.
- Model c directly, using an estimation method other than least squares—for example, weighted least squares or maximum likelihood of a generalized linear model (GLM) (see Wu and Hamada 2000, p. 568). However, modeling c directly causes the coefficient estimators to be correlated, due to the nonconstant variance.

Figure 2.14 displays the variance of \sqrt{c} , $(\sqrt{c} + \sqrt{c+1})/2$, and c on the same plot as a function of $E(c)$. The right axis labels values for $Var(c)$, and the straight line $y = x$ indicates the equality of $E(c)$ and $Var(c)$. The left axis denotes the variance for both transformations of c . The curve with the smaller peak near 0.4 is $Var(\sqrt{c})$ and the curve with the peak of 0.5 for $E(c) = 1$ is for the Freeman–Tukey transformation. The Freeman–Tukey transformation is essentially perfect for stabilizing the variance if $E(c) \geq 5$, but it is slightly worse than \sqrt{c} if some expected counts are below 2.5.

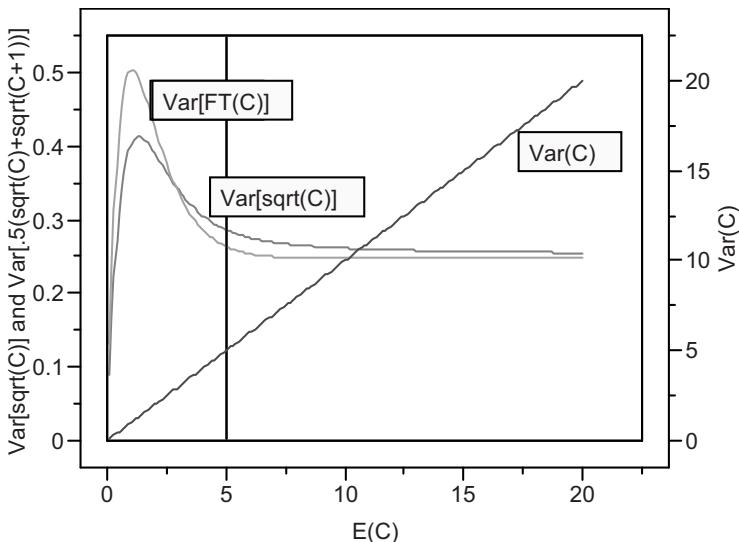


Fig. 2.14. Sqrt(Poisson count) as a variance-stabilizing transformation when expected count ≥ 5

For the Hsieh and Goodwin data, with one-fourth of the data being 3 or 4, we opt for the simpler square root transformation, since the Freeman–Tukey transformation is not better at stabilizing the variance for expected counts of 3 or less. Fitting a full factorial model for $y = (\text{number of pops})^{1/2}$, we obtain the 15 least squares estimates and display these in a half-normal plot (see Figure 2.15). In addition to drawing a line through the origin with a slope equal to Lenth's PSE ($= 0.517$), we draw a second line with a slope of $[0.3/16]^{1/2} = 0.137$, which would be the approximate standard error of the least squares estimates if in fact the data followed a Poisson distribution. (The variance of 0.3 in this calculation is taken from Figure 2.14.) The discrepancy between these two lines indicates that either the sparsity of effects assumption is violated, making Lenth's PSE too large, or the actual standard error is much larger than 0.137 because the “pop” defects do not follow a Poisson distribution. We suspect the latter, since otherwise a saturated model would be required to account for the observed data. The largest two estimates (in magnitude) are $b_2 = -1.11$ and $b_{3.4} = -1.05$, both with Lenth t statistics exceeding 2. This is evidence that the priming method coded “+1” is preferred, and that which thickening process is better depends on the viscosity factor’s level. The third largest effect is for mold pressure ($b_1 = -0.88$); although not statistically significant, it suggested to Hsieh and Goodwin that the higher pressure is better.

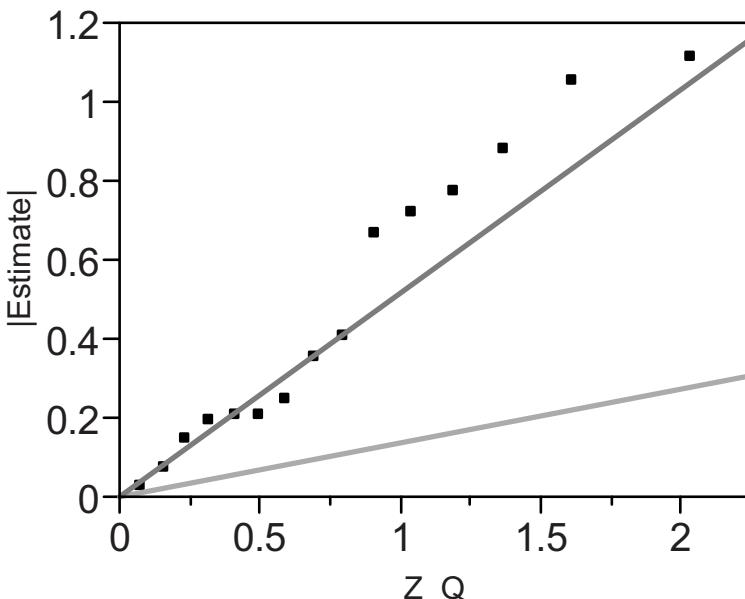


Fig. 2.15. Half-normal plot of effects for Hsieh and Goodwin (1986) data with $y = (\text{pops})^{1/2}$

2.8.3 Modeling variances

In Section 2.6.1, we discussed plotting residuals versus \hat{y} to verify that the assumption of constant variance is reasonable. For non-negative response variables where the ratio $\max(y_i)/\min(y_i)$ is large, we often find it necessary to use a transformation to satisfy the constant variance assumption. In that earlier discussion, however, the mean was primary; checking for equality of variance was a secondary concern. We now consider a different context, in which modeling variability is of primary interest.

Many process improvement applications involve sampling multiple items within each run in order to determine whether within-run variability is smaller at certain treatment combinations. When looking for differences in variability, taking only one or two observations per treatment combination renders an experiment useless. Instead, with primary interest on within-run variability, samples of $m = 10$ or more observations are recommended. For the analysis, one computes the standard deviation s_i or variance s_i^2 for each sample and then proceeds to model this measure of dispersion. If the m values from a sample are independent, normally distributed observations with some mean μ_i and variance σ_i^2 , then the sample variance s_i^2 is distributed as a multiple of a chi-square random variable; in particular,

$$s_i^2 \sim [\sigma_i^2 / (m - 1)] \chi_{m-1}^2$$

and $\text{Var}(s_i^2) = 2\sigma_i^4 / (m - 1)$. Thus, if we fit a regression model for s_i^2 and have any effects for $E(s_i^2)$, then the constant $\text{Var}(\epsilon)$ assumption will not hold. For this reason, the logarithm is the default variance-stabilizing transformation for standard deviations and variances, since

$$\text{Var}[\ln(s_i^2)] = \text{Var}[\ln(\chi_{m-1}^2)]$$

does not depend on σ_i^2 . Bartlett and Kendall (1946) is an early reference regarding the logarithm as a variance-stabilizing transformation for sample variances and standard deviations.

It was mentioned previously that a random sample of size $m \geq 10$ is recommended when studying variation. This is because the precision of a sample variance is poor when the degrees of freedom are few. Given the above result for the chi-squared distribution, the coefficient of variation (CV) for a sample variance of m independent normally distributed observations is $[2/(m-1)]^{1/2}$. Thus, for $m = 10$ observations, the CV is 47%; that is, the standard error for the sample variance is still nearly half as large as the variance we are estimating.

When our response is $\ln(s_i^2)$, the degrees of freedom in s_i^2 determines the variance of the error term in our model. In particular, suppose s_i^2 is the variance of m independent identically distributed observations from a normal distribution; then

$$\text{Var}[\ln(s_i^2)] \approx 2/(m - 2).$$

This approximation is excellent for large degrees of freedom and is adequate for m as small as 4. Thus, for a sample of size $m = 10$, we anticipate a RMSE near $[2/(10 - 2)]^{1/2} = 0.5$ for $\ln(s_i^2)$, or 0.25 for $\ln(s_i)$, provided the data are normally distributed.

Kramschuster et al. (2005) reported two 32-run experiments involving injection molding. For each run, they achieved a steady state and then selected a sample of 10 parts. After aging the parts, $m = 5$ parts per run were carefully measured for shrinkage and warpage. The means of these five observations were effective for finding several active effects for each dimension. However, if one attempts to fit a model to the standard deviations they report, no effects are found. In their case, analysis of the standard deviations is secondary, and measuring five parts per run carefully was quite time-consuming. However, for experiments for which variability is of primary concern, larger samples are generally necessary.

This book does not give any attention to methods for detecting differences in variability from unreplicated designs with no subsampling within runs—even though statisticians have proposed methods for attempting such an analysis. The basic strategy has been to fit a model for the mean, compute residuals, and then use the residuals to discover dispersion effects (i.e., factors that change the variability). For those interested in such methods, see the assessment by Brenneman and Nair (2001). Their concluding remark explains why these methods are not discussed here. “(T)he analysis of location and dispersion effects is intrinsically a difficult problem. In unreplicated experiments, it is really a minefield, one that needs to be maneuvered very carefully. George Box once compared this to trying to squeeze every last bit of water out of a wet towel. If you squeeze too hard, things start breaking down and you can end up making erroneous conclusions” (p. 403).

The first case study in Chapter 4 analyzes a 2^3 factorial with both true replication of runs ($n = 6$) and within-run sampling ($m = 25$), where important differences in within-run variability are found. The samples within each run are unstructured. In some studies of within-run variability, the physical layout suggests likely patterned differences. Section 13.3 discusses advantages of structured samples rather than random samples for variability experiments, and Section 14.3 illustrates the analysis of such data.

2.8.4 Modeling other statistics

Just as count data and variances have default transformations that facilitate the analysis, so do other statistics. Sample correlations r are bounded by the interval $[-1, 1]$, and have more variability when $E(r)$ is near the middle of that range. The default variance-stabilizing transformation for sample correlations, as devised by Sir Ronald Fisher, is

$$f(r) = 0.5 \ln[(1 + r)/(1 - r)].$$

Recent work by Fujisawa (2000) reinforces this transformation’s usefulness.

Ratios (y) constrained to the interval $[0, 1]$ may be transformed using the beta transformation advocated by Rocke (1993):

$$f_B(y; \lambda) = \int_0^y t^{\lambda-1} (1-t)^{\lambda-1} dt. \quad (2.13)$$

Examples include yield of refining and chemical processes, compositional data, and shrinkage measurements. This beta transformation family includes as special cases the $\arcsin(\sqrt{p})$ and $\ln[p/(1-p)]$ transformations mentioned earlier in Section 2.8.1. Rocke also suggested a generalization of (2.13) where the exponents for t and $1-t$ are allowed to differ.

The logarithm is a useful transformation for lifetime data, t . If the original distribution can be assumed to be lognormal, then $y = \ln(t)$ is normally distributed, and if t follows a two-parameter Weibull distribution, then $\ln(t)$ has an extreme value distribution. In both of these cases, the distribution for $\ln(t)$ is summarized by a location parameter and a scale parameter. Thus, we typically fit a model for $E[\ln(t)]$, with the hope that the variability of the residuals is nearly constant. The interpretability of the fitted model is facilitated by connecting parameters on the $\ln(t)$ scale to parameters of the distribution for t . The log transformation also applies when the response is an order statistic from a lifetime distribution; see, for example, Example 6.5.

Directional response data are often analyzed assuming the von Mises distribution (for responses on a circle) or the von Mises–Fisher distribution (for higher dimensions). Anderson and Wu (1995, 1996) fitted models for both location and dispersion for replicated angular data from a 2^4 factorial design. Anderson-Cook (2001) showed how to model the correlation between an angular response and a continuous response. These methods are relevant for any cyclic response, including time of day (or week or year).

Sometimes the response is a profile or function rather than a scalar. Walker and Wright (2002) analyzed density profiles for fiberboard products. Nair, Taam, and Ye (2002) analyze a compression strength profile for plastic foam. Nair et al. also analyze the audible noise and current of alternators as a function of speed. For each of these examples, the response for the i^{th} experimental run is a sequence of (y_{ij}, x_{ij}) , where the x_{ij} 's are univariate and fixed. Assuming the sequence of x_{ij} 's is the same for all runs, one approach is to fit a model for each $j = 1, \dots, J$. Nair et al. (2002) took this approach to analyze both the compression strength profiles and the noise output for alternators, in part because no simple functions was adequate to describe the observed data. Shen and Faraway (2004) showed how to conduct inferences for the fitted profiles, whereas Shen and Xu (2007) described diagnostic procedures. A second approach is to fit a curve to the data for each run and then to model some summary measure of each fitted curve. Nair et al. (2002) fitted a three-parameter nonlinear model for each run of the alternator current experiment and then modeled the logarithm of different functions of these parameter estimates. For similar analyses for repeated measures (i.e., longitudinal) data, see Yang, Shen, Xu, and Shoptaw (2007) and Engel (2008).

2.9 Unequal Replication and Unequal Variance

Sometimes a design has unequal replication that was planned. For instance, Snee (1985) replicated 4 of the 16 distinct treatment combinations in an experiment that involved several qualitative factors (which precluded the use of replicated center runs). For such planned imbalance, $\mathbf{X}'\mathbf{X}$ is not diagonal, but it may be block diagonal or have some other structure that may be exploited in the analysis (Dykstra 1959, Liao and Chai 2009).

In other cases, some intended runs fail to produce data, or we discard outlier observations, and end up with unequal replication that is unplanned. Let n_i denote the number of observations at each of the $i = 1, \dots, 2^k$ treatment combinations. Here we consider the case where $n_i \geq 1$ for all i ($i = 1, \dots, 2^k$); that is, we have a full factorial with unequal replication. In the next section we consider applications where $n_i = 0$ for some i .

With unequal replication of a full factorial, one can estimate the saturated model (1.4) but due to the lack of balance some regression coefficient estimates change when other terms are dropped from the model. There is some disagreement about which tests are most appropriate (Nelder and Lane 1995, Langsrud 2001). We illustrate the issues using data similar to Dykstra's (1959) 2^3 example. Table 2.9 reports the 12 responses for this experiment that contained replication at half of the treatment combinations.

Table 2.9. Partially replicated 2^3 factorial

x_1	x_2	x_3	Observations
-1	-1	-1	18.4, 20.6
1	-1	-1	25.1
-1	1	-1	24.3
1	1	-1	24.4, 26.2
-1	-1	1	20.4
1	-1	1	25.8, 27.0
-1	1	1	23.6, 24.6
1	1	1	27.9

Fitting a saturated model, we obtain a $MSE = 5.26/4 = 1.315$. The fitted model and t statistics are listed in Table 2.10. The standard error for each coefficient in the saturated model is $\sigma/(10.6)^{1/2}$, rather than $\sigma/(12)^{1/2}$, due to correlations among pairs of estimates.

Table 2.10. Saturated model for partially replicated 2^3 factorial

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	24.125	0.351	68.71	<.0001
x_1	2.050	0.351	5.84	.0043
x_2	1.275	0.351	3.63	.0221
x_3	0.575	0.351	1.64	.1768
$x_1 * x_2$	-0.850	0.351	-2.42	.0727
$x_1 * x_3$	0.400	0.351	1.14	.3182
$x_2 * x_3$	0.025	0.351	0.07	.9467
$x_1 * x_2 * x_3$	0.300	0.351	0.85	.4410

If the design were balanced, a reduced model could be selected in a single step. One might choose a reduced model with two, three, or four terms depending on whether one retains the $x_1 * x_2$ interaction and whether one follows the practice of retaining all main effects for full factorial designs (as was mentioned in Section 2.4). Regardless of which model is chosen, for balanced designs estimates for the terms in the model are unaffected, as are their t statistics if they are based on the pure error mean square.

Lack of balance complicates the choice of a model. Four possible fitted models are displayed in Figure 2.16. The columns for x_3 and $x_1 * x_2$ are correlated with a correlation of $-1/3$. If both terms are included in the model, as is the case in the reduced model with four terms, the estimates are different than when only one of these terms is included. This causes some ambiguity, since each estimate is larger when the other is omitted. Here, the other columns are orthogonal because of the careful choice of which four runs are replicated. In other nonorthogonal situations, all estimates may be correlated.

So how should one approach model selection? For a full 2^k with unequal replication, stepwise regression procedures are useful. First, fit the saturated model and use backward elimination for models restricted to be hierarchical. Then apply forward selection, again requiring hierarchical models, to see if the same model is obtained. For the data in Table 2.9, using $\alpha = .05$, both procedures lead to reduced model 3 in Figure 2.16.

Unequal replication is particularly common when the responses are from voluntary participants. If the assignment to treatment combinations is made before one knows which participants will respond, then the number of participants contacted needs to be large enough (i) to avoid empty cells and (ii) to avoid large correlations among the columns of the model matrix. Let n denote the number of participants invited per treatment combination (so that, in total, $N = 2^k n$ are invited) and let π denote a (conservative) guess for the proportion of participants who will agree to participate. Then having $n \geq 5/\pi$ is sufficient to avoid empty cells. However, when the realized 2^k sample sizes are random, the distribution for the correlation between two

Parameter Estimates for Reduced Model 1

<u>Term</u>	<u>Estimate</u>	<u>Std Error</u>	<u>t Ratio</u>	<u>Prob> t </u>
Intercept	24.0250	0.476	50.46	<.0001
X1	2.0417	0.476	4.29	0.0020
X2	1.1417	0.476	2.40	0.0400

Parameter Estimates for Reduced Model 2

<u>Term</u>	<u>Estimate</u>	<u>Std Error</u>	<u>t Ratio</u>	<u>Prob> t </u>
Intercept	24.0250	0.404	59.52	<.0001
X1	2.0417	0.404	5.06	0.0010
X2	1.1417	0.404	2.83	0.0222
X3	0.8583	0.404	2.13	0.0661

Parameter Estimates for Reduced Model 3

<u>Term</u>	<u>Estimate</u>	<u>Std Error</u>	<u>t Ratio</u>	<u>Prob> t </u>
Intercept	24.0250	0.346	69.54	<.0001
X1	2.0417	0.346	5.91	0.0004
X2	1.1417	0.346	3.30	0.0108
X1*X2	-1.0417	0.346	-3.01	0.0167

Parameter Estimates for Reduced Model 4

<u>Term</u>	<u>Estimate</u>	<u>Std Error</u>	<u>t Ratio</u>	<u>Prob> t </u>
Intercept	24.0250	0.307	78.18	<.0001
X1	2.0417	0.307	6.64	0.0003
X2	1.1417	0.307	3.71	0.0075
X3	0.5750	0.326	1.76	0.1211
X1*X2	-0.8500	0.326	-2.61	0.0350

Fig. 2.16. Four reduced models for partially replicated 2^3

columns of \mathbf{X} is symmetric about 0, with a standard deviation of approximately $[(1-\pi)/(2^k n\pi)]^{1/2}$. For a 2^3 , $n\pi = 8$ expected responses per treatment combination may seem sufficient, but if $\pi = 0.1$, $[(1-\pi)/(2^k n\pi)]^{1/2} = 0.12$, so about 5% of the correlations will exceed .24 in magnitude. The smaller the expected proportion π , the larger the expected number responding is required to avoid large correlations.

We now turn to the second topic of this section: unequal variance. If the error variance, σ^2 , is not constant, then the least squares estimators will be correlated, even if $\mathbf{X}'\mathbf{X}$ is diagonal. These correlations do not bias the ordinary least squares (OLS) estimators, but they do make OLS inefficient. When the variance of the response is a function of the mean $E(y)$, using a variance-stabilizing transformation resolves this difficulty by changing the model to one where OLS is appropriate. If replication is sufficient to estimate precisely the error variance for each run, then weighted least squares may be applied, weighting by the reciprocal of the estimated variances (see Section 14.4). This standard modification to least squares is discussed in most linear regression books. Its use is unnecessary when the unexplained variation is negligible.

2.10 The Impact of Missing Treatment Combinations

When all 2^k treatment combinations have at least one observation, one can fit the full factorial model (1.4) or any reduced model. In such cases, unequal replication results in correlated estimates of the coefficients but does not alter which models can be fit. Suppose instead that there are $m > 0$ factorial treatment combinations with no data. Then one must omit at least m coefficients from the full factorial model. Because the likelihood of missing treatment combinations is greatest for unreplicated 2^k factorials, we focus on that case.

Our approach to analyzing 2^k factorials with missing observations will be first to fit a saturated hierarchical model. If only one observation is missing, the saturated model is the full factorial model with the k -factor interaction omitted. If two or more observations are missing, there are several options. The details will be shown later.

For this section we use the following notation: $N = 2^k$ is the intended number of runs, of which m are missing, and r is the number of columns for the model matrix \mathbf{X} . If $r = N - m$, the model is saturated.

2.10.1 One missing treatment combination

If any single observation is lost from an unreplicated 2^k factorial, $(\mathbf{X}'\mathbf{X})^{-1}$ has a simple structure. Diagonal elements equal $(N - r + 1)/[N(N - r)]$ and off-diagonal elements equal $\pm 1/[N(N - r)]$. For the saturated model ($r = N - 1$), this implies that $\text{Var}(b_i) = 2\sigma^2/N$, double what it would have been for the complete 2^k , and all estimates are correlated with a correlation of $\pm .5$. If fewer terms are included in the model, these correlations $\pm 1/(N - r + 1)$ decrease in magnitude and the variances are reduced. Even then, the loss of orthogonality has a much greater impact on the analysis than does the reduction of the sample size.

Draper and Stoneman (1964) present the following example. The full data appear in the Figure 2.17 cube plot, and a half-normal plot of effects is shown in Figure 2.18a. No simple model will account for these data, primarily due to the observation $y = 44$. Upon investigation, it was learned that at the high level for all three factors, “the experimental material changed its form.” If this observation is treated as missing in the analysis, the two-factor interaction model can be estimated. Under this saturated model, the predicted value for the $(+1, +1, +1)$ treatment combination is $\hat{y} = 28$, 16 less than 44.

The half-normal plot in Figure 2.18b of the six (correlated) coefficients indicates no effect for x_3 . The reduced model

$$\hat{y} = 12.75 + 7.25x_1 - 0.75x_2 + 2.75x_1 * x_2$$

fits the data very well, except near the high level for all factors.

Contrast the two half-normal plots in Figure 2.18. In the second plot, the clump of estimates near zero for the model fitting only the seven treatment

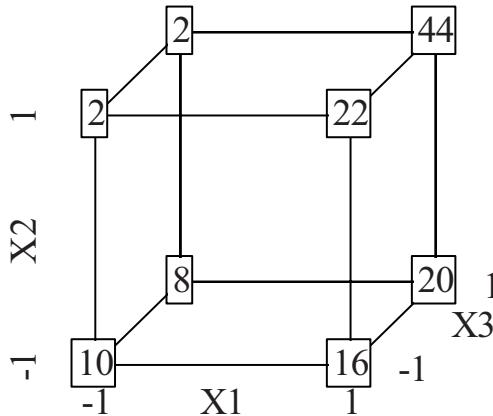


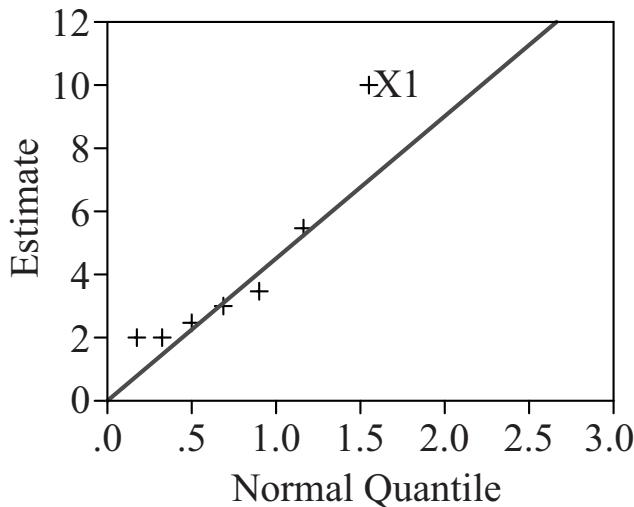
Fig. 2.17. Draper and Stoneman 2^3

combinations indicates that a model with only a few terms will fit very well. In contrast, fitting a model to all eight observations produces a clump of estimates in the range 2–3. Daniel (1959) observed that a single outlier with large error E would affect all the estimated coefficients by $\pm E/N$, pushing a majority of estimates for negligible effects away from zero. Thus, half-normal plots like those in Figures 2.18a and 2.18b are indicative of a simple model accounting for all but one of the observed y values.

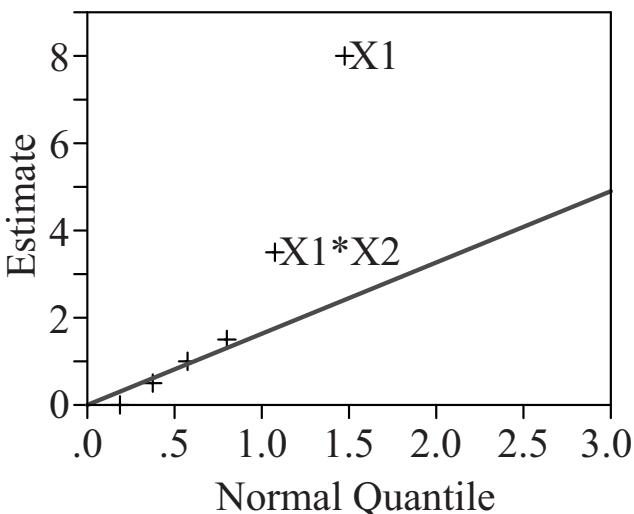
Note that the fitted model corresponding to Figure 18b assumes that the three-factor interaction coefficient is zero. If this assumption were not correct, then all the regression coefficients would be biased by $\pm \beta_{1.2.3}$. In general, with one observation missing, the bias for each coefficient from assuming away the highest-order interaction is $\pm \beta_{1.2\dots k}/(N-r)$. Thus, a clump of estimates close to zero (as in Figure 2.18b) adds credence to the assumption that the highest order interaction is zero.

2.10.2 Two or more missing treatment combinations

To fit a hierarchical model with $m > 1$ missing observations, there may be several hierarchical saturated models that can be estimated from the data. Using software to fit a model with all terms except the highest-order interaction will result in $m-1$ linear dependencies. Use these “singularity details,” as they are labeled in some software, to determine which choices one has for removing $m-1$ additional terms. For each possible saturated model, view the half-normal plot for (correlated) estimated effects. Finding a clump of estimates close to zero is consistent with the assumption that the omitted effects are negligible and that a further simplification of the model is possible. We now illustrate such an analysis for Daniel’s drill data (Figure 2.6) by omitting the observed values with the two lowest advance rates (1.68 and 1.98) and the



(a) Half-normal plot for 7 estimates from full factorial



(b) Half-normal plot for 6 correlated estimates

Fig. 2.18. Half-normal plots for Draper and Stoneman estimates

highest advance rate (16.3). Fitting a model with $r = 15$ to the $N - m = 13$ observations produces the following singularities:

$$\begin{aligned}\text{Intercept} &= x_{\text{Flow}} * x_{\text{Speed}} * x_{\text{Mud}} + \dots \\ &= -x_{\text{Load}} * x_{\text{Flow}} * x_{\text{Speed}} - x_{\text{Load}} * x_{\text{Flow}} * x_{\text{Mud}} - x_{\text{Load}} * x_{\text{Speed}} * x_{\text{Mud}} + \dots\end{aligned}$$

(which we have simplified by skipping main effects and two-factor interactions). There is no choice regarding the $x_{\text{Flow}} * x_{\text{Speed}} * x_{\text{Mud}}$ interaction; since we have no data at the $(-1, -1, -1)$ combinations for these factors, this term must be omitted. However, because the second singularity involves the other 3 three-factor interactions, this linear dependency may be removed by omitting any one of these interactions. So there are three possible hierarchical models with $r = 13$ that we may estimate (see Table 2.11). Each one results in a model with three significant main effects and b_{Load} never stands out above the clump of estimates near zero.

Table 2.11. Coefficients for three saturated models for $\ln(y)$, treating three observations from Daniel's drill data as “missing”

Term	Model 1	Model 2	Model 3	Std Error
Intercept	1.535	1.535	1.556	$\sigma/4^{1/2}$
Load	0.038	0.039	0.060	$\sigma/8^{1/2}$
Flow	0.307	0.307	0.307	$\sigma/8^{1/2}$
Speed	0.594	0.594	0.594	$\sigma/8^{1/2}$
Mud	0.181	0.181	0.181	$\sigma/8^{1/2}$
Load*Flow	-0.036	-0.036	-0.036	$\sigma/8^{1/2}$
Load*Speed	-0.014	-0.014	-0.014	$\sigma/8^{1/2}$
Load*Mud	0.014	0.014	0.014	$\sigma/8^{1/2}$
Flow*Speed	-0.088	-0.088	-0.067	$\sigma/4^{1/2}$
Flow*Mud	-0.071	-0.070	-0.049	$\sigma/4^{1/2}$
Speed*Mud	-0.014	-0.014	0.007	$\sigma/4^{1/2}$
Load*Flow*Speed	-0.021	-0.021		$\sigma/8^{1/2}$
Load*Flow*Mud	-0.001		0.021	$\sigma/8^{1/2}$
Load*Speed*Mud		0.001	0.021	$\sigma/8^{1/2}$
Flow*Speed*Mud				
Load*Flow*Speed*Mud				

To test for statistical significance requires a modification to Lenth's procedure, since the estimates are correlated. For details, see Edwards and Mee (2008). The success of finding three significant main effects for this example should not diminish the serious loss of information here. If all 16 observations are available, the standard errors are $\sigma/(16)^{1/2}$. The loss of three observations

doubles the standard error for three estimated two-factor interactions in the saturated model, causing a severe loss of power for detecting these effects.

Because of the correlations that result when we fit a saturated model to a factorial with missing observations, it is important to estimate the coefficients using a reduced model. Here, the fitted reduced model is

$$\widehat{\ln(y)} = 1.583 + 0.279x_{\text{Flow}} + 0.566x_{\text{Speed}} + 0.152x_{\text{Mud}}. \quad (2.14)$$

[Compare with (2.11).] The advantage of the estimates in (2.14) is that their standard errors are $\sigma/(11)^{1/2}$, smaller than the standard errors in Table 2.11. However, this benefit comes at a risk of bias to the estimated coefficients, if in fact omitted terms are active.

Because a saturated model can have highly inflated standard errors when treatment combinations are missing, a further step to model selection is to use some form of forward selection regression, adding interaction terms to a main effects model. For our example, fitting models with only one of the two-factor interactions with the large standard errors in Table 2.11 eliminates the largest correlations and enables one to better assess the presence of these terms. Here, no additional useful terms are found.

A final comment is in order. Because the loss of observations is so detrimental to an unreplicated 2^k factorial, such a design is not recommended unless the experimentation and measurement processes are very dependable. If such a design is run and several observations are lost, one may consider a subsequent set of runs to repair the original design. In such cases, it is advisable to run not only the missing observations but also some duplicate treatment combinations that were satisfactory, to account for a possible shift in the process since the initial 2^k was attempted (see Section 9.6).

Common Randomization Restrictions

The examples presented in Chapters 1 and 2 were either unreplicated full factorial designs with random assignment of treatment combinations to runs or they were replicated experiments obtained without any restriction to run order. Some experiments are too large to be run effectively in this manner. Hence, they are divided into several smaller experiments, commonly called blocks. Having some difficult-to-change factors is a second reason for restricting the assigned order for treatment combinations. In such cases, the assignment is restricted to make the experiment more convenient to conduct. This chapter presents the details of how to construct such designs, and how to analyze the resulting data. The sections are as follows:

Section 3.1. Sources of Variation and a Design's Unit Structure

Section 3.2. Treatment*Unit Interactions

Section 3.3. Blocking: Partitioning a Factorial into Smaller Experiments

Section 3.4. Analyzing Randomized Block Factorial Designs

Section 3.5. Split-Unit Designs

Section 3.6. Analyzing Split-Unit Designs

Section 3.7. Multiway Blocking

3.1 Sources of Variation and a Design's Unit Structure

Before a physical experiment is conducted, one should consider not only the factors to be studied but also the likely sources of extraneous variation. The more that one understands the sources of variation and their magnitude, the more efficient an experiment may be planned. There are four basic approaches for dealing with extraneous variation:

1. **Eliminate sources of variation by holding them constant.** For instance, an entire experiment might be performed using a single homogeneous batch of raw material. Doing so eliminates batch-to-batch variation from the experiment. In other situations, we exercise more careful control to reduce the variation of process inputs such as temperature. Although not entirely eliminating a source of variation, reducing it achieves the primary benefit.
2. **Isolate sources of variation by partitioning the experiment into smaller sets of runs performed under homogeneous conditions.** For instance, if the full factorial experiment will require several batches of raw material, we may systematically partition the treatment combinations into subsets and utilize one batch per subset. This chapter will describe how best to partition the treatment combinations of a 2^k factorial. The subsequent analysis will isolate any between-batch variation so that the error variance consists of the smaller, within-batch variation.
3. **Measure the sources of variation and incorporate these in the data analysis.** Suppose that each run of an experiment requires a complete raw material batch. In this situation, we cannot eliminate batch-to-batch variation or isolate it. However, if we measure batch characteristics that impact the response variable, then the variation that batch differences produce in the response may be accounted for in the data analysis. This is done using (analysis of covariance) regression models that incorporate our supplementary measurements as explanatory variables. For details, see Silknitter, Wisnowski, and Montgomery (1999).
4. **Use sufficient replication to overwhelm the unexplained variation.** The larger the extraneous variation, the more replications are required to make any systematic factor effects evident. So if the options for controlling, isolating, or explaining extraneous variation are not practical, increasing the amount of replication can achieve the desired precision.

Experiments discussed in Chapters 1 and 2 were based on Approach 1 or 4; that is, either recognized sources of variation were controlled or the replication was sufficient to average away the background variation. In either case, nothing is known about the error associated with individual runs. Some might think Approach 1 is the ideal, since by eliminating sources of variation, small experiments will reveal systematic differences resulting from the treatment combinations. However, the potential downside for such experiments is that the results may have limited validity. If a single batch of raw material is used, we will not know from the data how similar the factor effects will be for other batches. Sir Ronald Fisher articulated this potential disadvantage:

The exact standardization of experimental conditions, which is often thoughtlessly advocated as a panacea, always carries with it the real disadvantage that a highly standardized experiment supplies direct information only in respect of the narrow range of conditions achieved by standardization. (Fisher 1950)

Thus, although Approach 1 is successful for identifying factor effects by maintaining a small error variance, it is poorly suited for generalizing those results. Approach 2 is intended to achieve the high precision without the disadvantage related to high uniformity. Consider again the example of an experiment that utilizes batches of a raw material. Using several batches over the course of the experiment avoids the limitation cited above by Fisher. The design technique of orthogonal blocking described in this chapter enables the experimenter to make the variation from different batches separable from the differences due to main effects and most interactions. Other advantages of blocking include the following:

- Blocking enables increasing the size of an experiment without increasing the (unexplained) error variation.
- A blocked design is better in situations in which extenuating circumstances produce a series of failed runs. When such problems are confined to a single block, the simplest remedy is often to discard data for that block and (if necessary) repeat that entire block of treatment combinations.
- For industrial experiments, a sequence of smaller experiments may be more manageable than one very large design. For example, access to a production line for experimentation may be limited to brief time periods between scheduled production periods.

Sources of variation that may serve as blocking factors include the following:

- material: batches of raw material
- equipment: multiple machines that perform the same function
- people: different operators
- location: different environments
- time-related: different shifts or days

The purpose of a typical experiment is not to study blocking factors. Thus, we may control several of these sources of variation simultaneously. If we conduct an experiment over several days, we would commonly declare each day (or even smaller time period) as a block. Within each block we would plan to avoid changes of batches, operators, etc., and such changes would be allowed or even encouraged between blocks.

To speak further about block designs, we introduce the term *experimental unit*. In general, an experimental unit refers to the entity to which the treatment combinations are assigned and to the entity being measured. For an agricultural field experiment, the experimental unit is typically a plot of ground. Each treatment combination represents a set of conditions applied to a plot, and each response value is the measured outcome that results from a particular experimental unit and its assigned treatment combination. For experiments involving factors applied to people, individuals or groups of people become the experimental unit. For a manufacturing process, the experimental unit refers both to what is produced (and subsequently measured) and to the

material that was utilized to make the item. For Huhtamaki's experiments described in Chapter 1, the experimental unit is the carton.

As mentioned earlier, each y value we obtain reflects the influence of the experimental unit and its assigned treatment combination. If the same treatment combination were applied to every experimental unit in an experiment, all the observed variation would be attributable to differences in the experimental units, plus random measurement error. (In agriculture, such variability studies are called uniformity trials.) Any prior understanding about differences in the experimental units can be used to design a more efficient factorial experiment. For instance, if one process step involves use of a kiln or furnace, knowing about the inherent variation assignable to different locations within the kiln could be used to design an experiment with regions having homogeneous conditions as blocks. Always we seek to minimize the variation within blocks, but we allow or even promote differences between blocks.

For the experiments in Chapters 1 and 2, nothing was discussed about unit variation. That is because important sources of extraneous variation recognized in advance are held fixed throughout the experiment for a completely randomized design, and one does not utilize any knowledge about the remaining underlying variation in the units to create the design. In contrast, designs with blocking are arranged by grouping experimental units into homogeneous subsets. Thus, understanding unit-to-unit variation plays a more prominent role for randomized block designs and the closely related split-unit designs. For further reading, Bisgaard (1994) provides a very nice discussion about the usefulness of blocking for two-level factorial designs. See Cox and Reid (2000, Chapters 3-4) for a more general discussion about blocking.

3.2 Treatment*Unit Interactions

Every observed y reflects the influence of its treatment combination and its experimental unit. A common assumption is that these effects are additive; for example, that

$$y = \text{unit value} + \text{treatment effect} + \text{measurement error}.$$

When this additivity assumption holds, the true factor effects do not depend on which experimental units are included in the experiment.

Often a replicated factorial experiment is conducted in blocks such that each block is a single 2^k . This is called a complete block design, because each block has the complete set of treatment combinations. Since it is possible to estimate all of the factorial effects from each block, a complete block design makes it easy to check for consistency of factorial effects across the different blocks. Sanders, Leitnaker, and McLean (2001) described the benefits of checking for consistency of factorial effects from one block to the next. Rather than simply assuming that block and treatment effects are additive, we prefer to verify that this is the case by either fitting a model for each block

separately or analyzing the data together utilizing an initial full model with Block*Factor interactions. When the differences between the blocks are very large, the likelihood for Block*Factor interactions is greater. The next section describes the use of incomplete blocks; that is, blocks where only a subset of the treatment combinations appear in each block. Leitnaker and Mee (2001) explain how to inspect for block-by-factor interactions in such cases.

3.3 Blocking: Partitioning a Factorial into Smaller Experiments

This section addresses situations in which 2^k treatment combinations of a full factorial design are divided into subsets, and the subsets are performed as separate experiments. Although these subsets could be of many sizes, attractive statistical properties are achieved if the subsets are of equal size. Suppose we divide the 2^k treatment combinations into 2^b subsets of size 2^{k-b} . For instance, if $k = 5$ and $b = 2$, then we have $2^b = 4$ subsets of size $2^{k-b} = 8$. Rather than randomly allocating the 32 treatment combinations into 4 subsets, the treatment combinations are split into subsets systematically using carefully chosen interaction columns. Consider the following example.

3.3.1 Example 3.1 constructed: A 2^5 factorial with four blocks

We illustrate running a factorial design in blocks using the five-factor experiment of Hoà̂ng et al. (2004). This experiment was conducted to identify interaction effects for different additives in linear low-density polyethylene (LLDPE) film. Here we introduce how the experiment was conducted, and later, in Section 3.4, we analyze the data. Table 3.1 identifies the five additives by type. Note that the factors are labeled with uppercase letters rather than as x_1, x_2, \dots . Using letters avoids the need for subscripts and will later provide a means of compactly labeling the treatment combinations.

Table 3.1. Factors and levels for Hoà̂ng et al. (2004) 2^5 design

Factors	Levels	
	-1	1
A Antioxidant A (ppm)	0	400
B Antioxidant B (ppm)	0	1000
C Acid scavenger (ppm)	0	1000
D Antiblock agent (ppm)	0	2000
E Slip additive (ppm)	0	800

The full 2^5 experiment would have taken at least 3 days to complete. Thus, as a precaution, it was decided to divide the full experiment into four

subsets. Table 3.2 lists the 32 treatment combinations in the order they were performed. Rather than randomizing the run order from 1 to 32, the design was constructed as follows. The **ADE** and **ABC** interaction columns were used to partition the treatment combinations into four subsets. Note that each of these columns is the product of three factors; for example, **ABC** denotes the product $\mathbf{A} \times \mathbf{B} \times \mathbf{C}$. **ADE** alone splits the 32 treatment combinations into 2 sets of 16, while **ADE** and **ABC** together split the treatment combinations into 4 sets of 8. Why these two interactions were used will be discussed momentarily. Once the full design is split into subsets, the order of the blocks, as well as the order of the runs within each subset or block, is randomized.

What are the benefits of restricting the run order in this manner? First, suppose the experiment requires four days to complete and that day-to-day differences cause the response to be systematically higher or lower some days. With four different days, there are three degrees of freedom for “Between Days.” **ADE** and **ABC** represent two of these degrees of freedom. The third is the “generalized interaction” of these two:

$$\mathbf{ADE} \times \mathbf{ABC} = \mathbf{A}^2\mathbf{BCDE} = \mathbf{BCDE}$$

(since \mathbf{A}^2 , or any column squared, becomes a column of 1’s). Thus, if the **ABC** and **ADE** columns are constant (+1 or -1) within a day, then so is **BCDE**. Hence, the interaction contrasts **ABC**, **ADE**, and **BCDE** together capture all of the day-to-day differences. Any day-to-day differences that might arise will bias our estimates for these three interactions but not the other effects. We say that the interactions **ABC**, **ADE**, and **BCDE** are *confounded* with blocks. This means that any effect due to blocks is mixed up with these interaction effects, and we cannot separate them.

The confounding of three interactions with blocks ensures that the main effects and other interaction contrasts are all orthogonal to blocks. For instance, note that for each block in Table 3.2, there are four +1’s and four -1’s for each factor. This is true not only for the five main effects but also for all interactions except for **ABC**, **ADE**, and **BCDE**. Thus, by arranging the treatment combinations in this manner, we sacrifice information about three higher-order interactions while shielding the remaining $31 - 3 = 28$ estimates from day-to-day differences. Since the shortest interaction confounded with blocks involves three factors, we say that this blocking scheme has estimability of 2. Sun, Wu, and Chen (1997) defined a block design to have estimability e if all factorial effects up to order e are estimable, clear of block effects, but one or more interactions of length $e + 1$ is confounded with blocks. For a 2^5 in four blocks, the maximum estimability is 2.

Table 3.2. Hoàng et al. (2004) 2^5 design in four blocks

Block	A	B	C	D	E	ADE	ABC
1	-1	1	1	1	1	-1	-1
1	-1	-1	-1	-1	-1	-1	-1
1	-1	-1	-1	1	1	-1	-1
1	1	1	-1	-1	1	-1	-1
1	1	-1	1	-1	1	-1	-1
1	-1	1	1	-1	-1	-1	-1
1	1	1	-1	1	-1	-1	-1
1	1	-1	1	1	-1	-1	-1
2	-1	-1	1	1	1	-1	1
2	1	-1	-1	-1	1	-1	1
2	1	-1	-1	1	-1	-1	1
2	-1	1	-1	-1	-1	-1	1
2	-1	-1	1	-1	-1	-1	1
2	1	1	1	-1	1	-1	1
2	1	1	1	1	-1	-1	1
2	-1	1	-1	1	1	-1	1
3	-1	-1	-1	1	-1	1	-1
3	1	1	-1	-1	-1	1	-1
3	1	-1	1	-1	-1	1	-1
3	1	-1	1	1	1	1	-1
3	1	1	-1	1	1	1	-1
3	-1	1	1	1	-1	1	-1
3	-1	-1	-1	-1	1	1	-1
3	-1	1	1	-1	1	1	-1
4	-1	-1	1	-1	1	1	1
4	1	-1	-1	-1	-1	1	1
4	1	-1	-1	1	1	1	1
4	-1	1	-1	-1	1	1	1
4	-1	1	-1	1	-1	1	1
4	1	1	1	1	1	1	1
4	-1	-1	1	1	-1	1	1
4	1	1	1	-1	-1	1	1

The cube plot in Figure 3.1 provides a visualization of this partitioning of runs into blocks. If we were to randomly assign the numbers 1–4 to the 32 treatment combinations, the results would not be as balanced as we see here. Each block of eight runs is uniformly spread over the experimental region. For any pair of factors, each block produces a replicated 2^2 . This is guaranteed here because the estimability is 2; that is, no main effect or two-factor interaction is confounded with blocks. If we were to randomly assign the treatment

combinations to blocks, most factorial effect contrasts would be correlated with block-to-block differences and the analysis would be difficult.

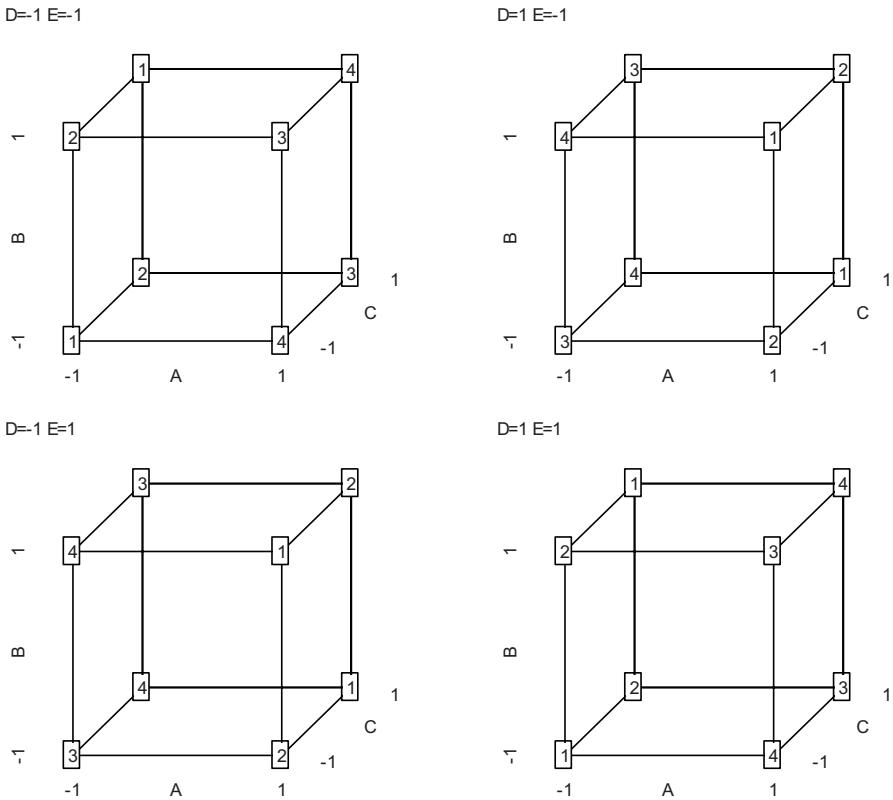


Fig. 3.1. Block number for treatment combinations in Hoàng et al.'s 2^5

3.3.2 General recipe for partitioning a 2^k into 2^b blocks

The case of two blocks is simplest. For a full factorial, one should create two blocks using the highest-order interaction to make the estimability $e = k - 1$. All main effects and interactions besides the k -factor interaction are orthogonal to blocks. This is equivalent to separating the runs with an even number of factors at the high level from those with an odd number of factors at the high level. Consider the case of a 2^4 factorial, displayed in Table 3.3. The eight treatment combinations with zero, two, or four factors at the high level form the even block, and the eight treatment combinations with either one or three factors at the high level form the odd block. Here treatment

combinations are denoted by strings of lowercase letters, where the presence (absence) of a letter indicates that factor is at the high (low) level. Yates (1935) introduced this notation. The treatment combination with all factors low is denoted by (1). As a brief aside, the block with this treatment combination is called the principal block, and its elements satisfy the mathematical properties of a subgroup; that is, the product of any two treatment combinations in this block is also in this block. For more detail, see John (1998, p. 136).

Table 3.3. 2^4 in two blocks

Even Block	A	B	C	D	Odd Block
(1)	-1	-1	-1	-1	
	1	-1	-1	-1	a
	-1	1	-1	-1	b
ab	1	1	-1	-1	
	-1	-1	1	-1	c
ac	1	-1	1	-1	
bc	-1	1	1	-1	
	1	1	1	-1	abc
	-1	-1	-1	1	d
ad	1	-1	-1	1	
bd	-1	1	-1	1	
	1	1	-1	1	abd
cd	-1	-1	1	1	
	1	-1	1	1	acd
	-1	1	1	1	bcd
abcd	1	1	1	1	

The case of four blocks is nearly as simple. To minimize the number of lower-order interactions that are confounded with blocks, each factor must appear in two of the three interactions confounded with blocks. Thus, on average, these interactions are $2k/3$ long. If k is a multiple of three, then the three interactions confounded with blocks can be of equal length. Otherwise, there must be one interaction of odd length and two interactions of even length closest to $2k/3$. For instance, for $k = 3, \dots, 9$, the three confounded effects can be as follows:

- Three factors: **AB, AC, BC**
- Four factors: **ABC, ABD, CD**
- Five factors: **ABC, ADE, BCDE**
- Six factors: **ABCE, ABDF, CDEF**
- Seven factors: **ABCDF, ABCEG, DEFG**
- Eight factors: **ABCDG, ABEFH, CDEFGH**
- Nine factors: **ABCDEH, ABCFGJ, DEFGHJ**

There is one other case that permits a general solution, the case for blocks of size 2 (i.e., $b = k - 1$). By confounding all even-length interactions with blocks, all main effects and odd-length interactions are orthogonal to blocks, and so we have estimability of 1. The case of 2^3 in four blocks illustrates this choice for blocking. The eight treatment combinations are grouped into four blocks as follows:

	$\mathbf{AB} = \mathbf{-1}$		$\mathbf{AB} = \mathbf{1}$	
$\mathbf{AC} = \mathbf{-1}$	a	bc	c	ab
$\mathbf{AC} = \mathbf{1}$	b	ac	(1)	abc

Note that each pair of treatment combinations forming a block has one high level and one low level for each factor. This makes each main effect contrast orthogonal to blocks, no matter the number of factors. For instance, for four factors, the blocking is given in Table 3.4, where seven interactions are confounded with blocks: \mathbf{AB} , \mathbf{AC} , \mathbf{BC} , \mathbf{AD} , \mathbf{BD} , \mathbf{CD} , and \mathbf{ABCD} .

Table 3.4. 2^4 in eight blocks

Block	A	B	C	D	t.c.*
1	-1	-1	-1	-1	(1)
	1	1	1	1	abcd
2	1	-1	-1	-1	a
	-1	1	1	1	bcd
3	-1	1	-1	-1	b
	1	-1	1	1	acd
4	1	1	-1	-1	ab
	-1	-1	1	1	cd
5	-1	-1	1	-1	c
	1	1	-1	1	abd
6	1	-1	1	-1	ac
	-1	1	-1	1	bd
7	-1	1	1	-1	bc
	1	-1	-1	1	ad
8	1	1	1	-1	abc
	-1	-1	-1	1	d

*t.c. = treatment combination

Thus, for either two or four blocks, or for blocks of size 2, it is straightforward to construct orthogonal blocks that maximize the order of estimability e and confound with blocks the minimum number of interactions of length $e + 1$. For other cases (i.e., for $3 \leq b \leq k - 2$), there is no general characterization of the confounded effects for optimal blocking. We know that the $2^b - 1$ confounded interactions should contain each factor 2^{b-1} times, but the best

choice of interactions must be found by computer search. Sun, Wu, and Chen (1997) found the best blocking schemes for up to $k = 8$ factors. Appendix E presents these optimal designs, or their equivalent.

In practice, the maximum block size is often limited by the physical constraints under which the experiment is run. For instance, if the experiment must be conducted over several days, with large day-to-day variation possible, then the block size should not exceed the number of runs that can be performed in a day. Suppose further that several batches of raw material are used each day. Then several batches of raw material may be blended in order to perform all of the runs each day under homogeneous conditions. If this is not feasible, then smaller blocks may be chosen corresponding to the number of runs that can be completed from a single batch; in this case, experimentation each day would consist of several blocks. Thus, although physical constraints influence the number of runs that can be accomplished under uniform conditions, some choice of block size often remains. We illustrate the practical use of Appendix E for such an example.

Suppose we have five factors and wish to conduct an experiment in small blocks, with size yet to be determined. The full array of orthogonal blocking choices are as follows:

- Two blocks of size 16: $e = 4$
- Four blocks of size 8: $e = 2$
- Eight blocks of size 4: $e = 1$
- Sixteen blocks of size 2: $e = 1$

If we are interested in most two-factor interactions, blocks of size 2 would not be considered further (unless the variation in larger blocks is excessive and it is feasible to perform several replicates of the 2^5). For blocks of size 4, only two two-factor interactions, **BD** and **CE**, are confounded with blocks, and the remaining eight two-factor interaction contrasts are orthogonal to blocks. If, in advance, we suspect that there are two interactions (involving different factors) that we do not suspect as being active, we might opt for an eight-block design, assigning factors to the letters **A–E** so that **BD** and **CE** correspond to interactions deemed unlikely. Of course, if blocks of size 8 could be performed with the same within-block consistency as blocks of size 4, then one could use four blocks of size 8 and have a design with estimability 2.

If experimental units within the same block are very similar, so that the error variation within blocks is small, then blocking increases the efficiency of estimates for effects orthogonal to blocks. Generally, the smaller the block size, the smaller the error variance. However, the smaller the block, the more effects are confounded with blocks, so this improvement of efficiency for some estimates comes at the loss of information for other interactions. In the extreme, we have blocks of size 2 created by confounding all two-factor interactions (and other even length interactions) with blocks.

When the block sizes are very small, it is common to perform more than one replicate of the 2^k in order to gain some information about additional

effects. There are two means of doing this. One is to change the confounding scheme from replicate to replicate, so that no interaction of possible interest is confounded with blocks in every replicate. Such blocking is named *partial confounding*. Yates's (1937) partial confounding example is analyzed in Section 3.4.3. Quenouille and John (1971) presented replicated designs with partial confounding for blocks of size 2 for up to eight factors. Yang and Draper (2003) considered every partial confounding option for blocks of size 2 for $k = 2, 3, 4$, or 5 factors and several replicates. See also Butler (2006) for optimal partial confounding for blocks of size 2 and 4. Kerr (2006) discussed how microarray applications may call for 2^k factorial designs in blocks of size 2.

The alternative approach is to keep the confounding the same from replicate to replicate but to estimate interactions confounded with blocks using interblock information. Section 3.4.4's example by Sheesley (1985) illustrates such a design and analysis.

On rare occasions, a choice other than orthogonal blocking should be considered. The most likely situation is when blocks of size 3 make practical sense but size 4 is not possible, or blocks of size 6 is possible, but size 8 is not. In these cases, use of nonorthogonal blocks of size 3 (6) will likely provide more information than having orthogonal blocks of size 2 (or 4). Optimal design algorithms may be used for constructing such irregular-sized blocks (see, e.g., Cook and Nachtsheim 1989, Nguyen 2001). See Section 11.4 for analysis of such a design.

Finally, we consider the use of centerpoint replicates for 2^k designs run in blocks. The common approach is to place the same number of centerpoint replicates in each block. This was done by Hoàng et al. (2004), who added one centerpoint run to each block, producing blocks of size 9; see Table 3.5. This maintains the orthogonality. However, since their four centerpoint replicates appear in different blocks, no pure error estimate of the within block variance is available. To achieve such an estimate, one would need multiple centerpoint runs in some blocks.

3.4 Analyzing Randomized Block Factorial Designs

We now illustrate the analysis of four factorial experiments that were each conducted as randomized block designs, with blocking determined by confounding interaction effects. The first example is from Hoàng et al. (2004), where the block effects appear negligible. The second example is from Davies (1954), for which block effects are substantial. The third example illustrates partial confounding, since a different interaction is confounded with blocks in each replicate. The fourth example, taken from Sheesley (1985), illustrates an analysis that uses between-block information to estimate an interaction confounded with blocks in each replicate. Figure 3.2 provides a general summary of these examples.

Example	Treatment combinations	No. of replicates	No. of blocks	Block size	Interactions confounded with blocks
3.1	2^5	1	4	8	ABC, ADE, BCDE
3.2	2^3	2	4	4	ABC in each replicate
3.3	2^3	4	8	4	Different interaction in each replicate
3.4	2^2	4	8	2	AB in each replicate

Fig. 3.2. List of randomized incomplete block examples

3.4.1 Example 3.1 analyzed: Randomized block experiment with negligible block effects

The run order and response data for Hoang et al.'s (2004) 2^5 in four blocks, each with one centerpoint run, appear in Table 3.5. Each block is a separate day, and the nine runs were performed sequentially. The responses are two stability measures (melt flow rate after the first and third extruder passes; S_1 and S_3 , respectively), a yellowness index after the third pass (YI), and a measure of long-term oxidation at 100°C (T); specifically, T is the number of hours until a carbonyl index reaches a specified level. We analyze the stability measure S_3 .

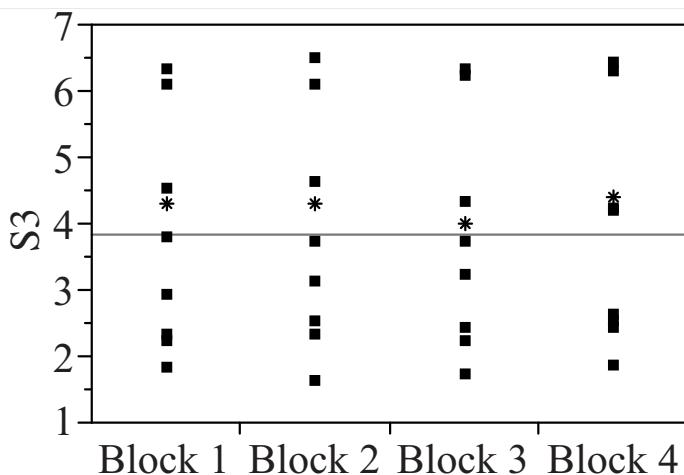


Fig. 3.3. S_3 versus blocks for Hoang et al.'s 2^5

Table 3.5. Hoàng et al. (2004) 2^5 design in four blocks

Block	A	B	C	D	E	S1	S3	YI	T
1	-1	1	1	1	1	1.02	2.90	-12.8	365
1	-1	-1	-1	-1	-1	0.67	1.80	-13.4	240
1	-1	-1	-1	1	1	0.37	2.29	-11.9	265
1	1	1	-1	-1	1	1.28	6.30	-13.0	1900
1	1	-1	1	-1	1	0.91	4.49	-11.5	1780
1	-1	1	1	-1	-1	0.79	2.20	.	365
1	1	1	-1	1	-1	1.29	6.07	-12.3	1310
1	1	-1	1	1	-1	0.49	3.78	-6.2	1220
1	0	0	0	0	0	1.06	4.29	-11.1	920
2	-1	-1	1	1	1	0.70	2.30	-14.2	335
2	1	-1	-1	-1	1	1.06	4.59	-10.5	1480
2	1	-1	-1	1	-1	0.95	3.69	-12.0	1305
2	-1	1	-1	-1	-1	0.68	2.50	-11.9	310
2	0	0	0	0	0	1.05	4.30	-10.2	1020
2	-1	-1	1	-1	-1	0.39	1.59	-9.6	290
2	1	1	1	-1	1	1.25	6.47	-8.8	1740
2	1	1	1	1	-1	1.30	6.08	-7.2	1290
2	-1	1	-1	1	1	1.00	3.09	-12.7	330
3	-1	-1	-1	1	-1	0.37	1.69	-12.2	275
3	1	1	-1	-1	-1	1.28	6.20	-9.0	1830
3	1	-1	1	-1	-1	0.90	3.70	-7.5	1615
3	1	-1	1	1	1	1.14	4.30	-8.2	1130
3	1	1	-1	1	1	1.29	6.30	-11.3	1380
3	-1	1	1	1	-1	0.47	2.39	-13.7	345
3	-1	-1	-1	-1	1	0.45	2.19	-14.1	400
3	-1	1	1	-1	1	0.96	3.20	-16.0	410
3	0	0	0	0	0	1.04	3.99	-9.5	900
4	-1	-1	1	-1	1	0.67	2.49	-15.2	310
4	1	-1	-1	-1	-1	1.19	4.19	-7.3	1640
4	1	-1	-1	1	1	0.84	4.18	-7.3	1400
4	-1	1	-1	-1	1	0.61	2.59	-13.5	275
4	-1	1	-1	1	-1	0.86	2.39	-15.5	335
4	1	1	1	1	1	1.29	6.40	-8.7	1515
4	0	0	0	0	0	1.10	4.41	-10.5	810
4	-1	-1	1	1	-1	0.46	1.85	-15.2	300
4	1	1	1	-1	-1	1.22	6.28	-8.4	1800

This design is a single replicate of a 2^5 factorial, with four centerpoint replicates. Figure 3.3 plots the S_3 data by block, using asterisks to mark the center runs. For this response there is little or no difference due to blocks.

Additionally, some curvature may be present, since the center runs are slightly above average.

It is simplest to analyze the data first, ignoring the blocking and without the centerpoint runs. Thus, we fit a saturated model to the 2^5 and examine a normal plot of the 31 orthogonal estimates; see Figure 3.4. Lenth's PSE = 0.0375, and the four effects with prominently large estimates form a hierarchical model and have Lenth t statistics ranging from 6.4 to 38.0. The next largest Lenth t statistic is -1.77 (for **CDE**). Thus, the choice of a reduced model with three main effects and one interaction is obvious. The three interaction contrasts confounded with blocks (**ABC**, **ADE**, and **BCDE**) all have t statistics of .73 or less, so there is no evidence that the blocks differ from one another for this response. While there could have been differences from one day (block) to the next, no effect on S_3 is apparent. The simple reduced model, ignoring blocks,

$$\widehat{S}_3 = 3.765 + 1.424\mathbf{A} + 0.695\mathbf{B} + 0.379\mathbf{AB} + 0.240\mathbf{E} \quad (3.1)$$

explains $R^2 = 99\%$ of the variation, and the residual plot shows no outliers. This reduced model has a mean square error of 0.0321 and standard errors for the estimated coefficients of $(0.0321/32)^{1/2} = 0.032$, only slightly smaller than Lenth's PSE.

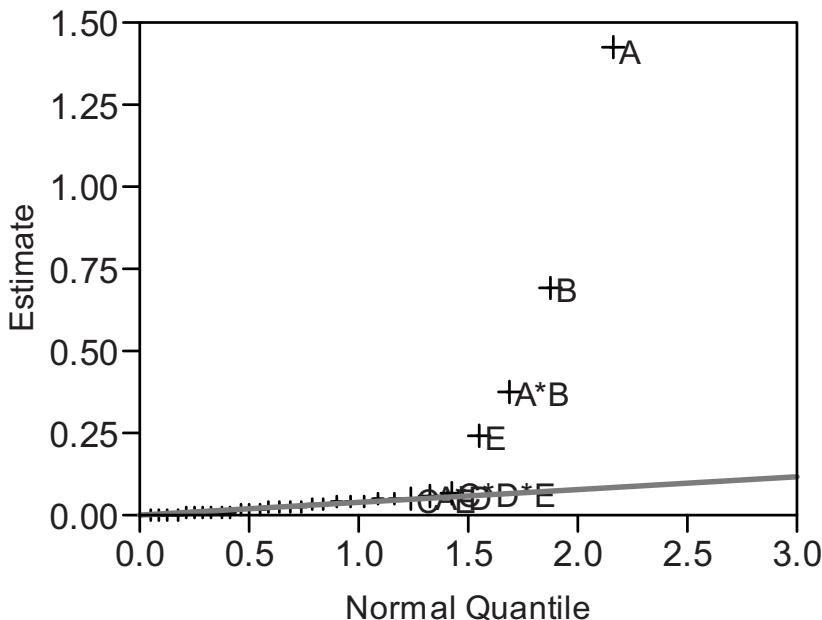


Fig. 3.4. Half-normal plot of effects from saturated model for S_3 from 2^5

Now we must consider the evidence for curvature due to the centerpoint replicates. The centerpoint values for $S3$ from blocks 1–4 were 4.29, 4.30, 3.99, and 4.41, all greater than the intercept in (3.1). With only one centerpoint run per block we have no pure error degrees of freedom. In this case, the best check for curvature is to add a quadratic term to our model. If we refit the model (3.1) to all 36 observations and include 1 quadratic term (e.g., \mathbf{A}^2), its estimated coefficient is -0.48 and its t statistic is $-0.48/0.097 = -5.08$, which is clearly significant. Note that this coefficient has a different standard error than the factorial effects. We conclude that the surface bows upward slightly and that the fitted model (3.1) will underestimate $S3$ in the center of the region. However, the pure quadratic term only explains 1% of the variation in $S3$. Thus, the linear effects of \mathbf{A} and \mathbf{B} dominate.

3.4.2 Example 3.2: Analysis of a randomized block experiment with large block effects

The previous example is not necessarily typical, since block effects can be very prominent. Davies (1954, pp. 372ff) presented a replicated 2^3 factorial experiment in four blocks of size 4, where block effects cannot be ignored. The experiment involved taking a batch of raw material and adding ammonium chloride. The chemical plant had two units for processing the conversion to produce the desired organic chemical. The output of these units was sampled and measured in a lab, bypassing the subsequent refining steps since there the identification of separate batches is impossible. The raw material is produced in batches large enough for two runs in one unit or one run in each unit. Because batch-to-batch variation is substantial, it was decided to block on batches. However, rather than using blocks of size 2, they decided to combine and blend two batches of raw material into a single homogeneous lot. Thus, each lot becomes a block of size 4. The primary factor of interest was the coarseness of the ammonium chloride. A coarse grind is used at present, but a finer grind would be worth the extra effort if it produced more than a 2.5% increase in yield. In addition to quality of ammonium chloride, the amount was varied by comparing the current level with a 10% increase. Factor levels are shown in Table 3.6.

Table 3.6. Factors and levels for chemical experiment

Factors		Levels	
		-1	1
A	Amount of NH_4Cl	Normal	+10%
B	Quality of NH_4Cl	Course	Fine
C	Processing unit	1	2

The results of the experiment are reported in Table 3.7, with the response being yield in pounds, followed by a graph of Yield versus Lot. Lot-to-lot variation accounts for 59% of the total variation in yield (see Figure 3.5).

Table 3.7. Treatment combinations and Yield for chemical experiment

Lot	A	B	C	Yield
1	-1	-1	-1	155
1	1	-1	1	152
1	-1	1	1	150
1	1	1	-1	157
2	-1	1	-1	162
2	-1	-1	1	156
2	1	1	1	161
2	1	-1	-1	168
3	-1	-1	1	161
3	1	1	1	173
3	-1	1	-1	171
3	1	-1	-1	175
4	1	1	-1	171
4	1	-1	1	162
4	-1	1	1	153
4	-1	-1	-1	164

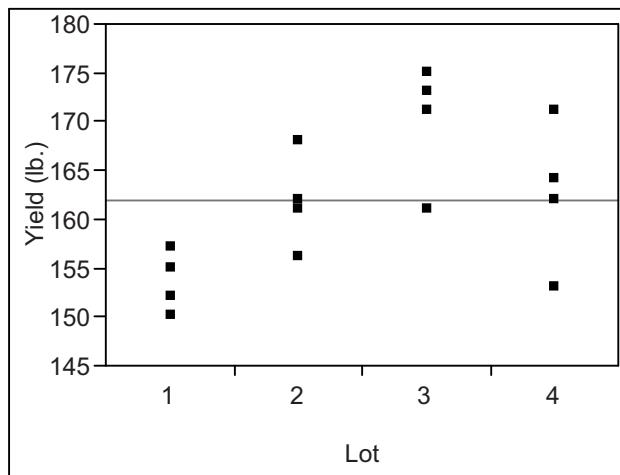


Fig. 3.5. Yield versus Lot for replicated 2^3

Since the three-factor interaction is confounded with lots, we fit the two-factor interaction model (1.3) augmented with a Lot (block) main effect. The resulting analysis of variance appears in Table 3.8.

Table 3.8. Analysis of variance for Example 3.2

Source	df	SS	MS	F-Ratio	p-Value
Lot	3	546.1875	182.0625	24.62	.0009
A	1	138.0625	138.0625	18.67	.0050
B	1	1.5625	1.5625	0.21	.6620
C	1	189.0625	189.0625	25.56	.0023
AB	1	1.5625	1.5625	0.21	.6620
AC	1	5.0625	5.0625	0.68	.4397
BC	1	3.0625	3.0625	0.41	.5437
Error	6	44.3750	7.3958		
Total	15	928.9375			

As expected by those conducting the experiment, interactions are not important. Somewhat surprising is the fact that the quality of grind main effect (**B**) is not statistically significant. The fine grind does not appear to improve yield, but increasing the amount of ammonium chloride does. Further, unit 1 performs more efficiently than unit 2. The fitted first-order model for yield is

$$\hat{y} = 161.94 + \text{Lot}_i + 2.94\mathbf{A} + 0.31\mathbf{B} - 3.44\mathbf{C}. \quad (3.2)$$

Adding 10% ammonium chloride is estimated to increase yield by $2b_{\mathbf{A}} = 5.88$ pounds per batch, and Unit 2 appears to yield 6.88 pounds/batch less than Unit 1. The experimental results in Table 3.8, combined with the experimenters' expectation of an additive model, affirms the reasonableness of sacrificing information about the three-factor interaction due to confounding with blocks. Additionally, the lot-to-lot variation in yield confirms the necessity of blocking on lots. The error variance within blocks is estimated to be 7.4, the MSE from Table 3.8. If one fits model (3.2) with the lot effect declared random (see Section 14.5 for details), the estimated lot-to-lot variance is 44.0. Adding these variance components, we see that the error variance would have been close to 50 if each run had come from a different blended material lot. Thus, blocking greatly improved the precision of Davies's experiment. Subsequent investigations should seek to learn why yield varies so much from batch to batch.

3.4.3 Example 3.3: Analysis for a randomized block experiment with partial confounding

Yates (1937) presented the results of a potato yield experiment involving four factors. Because soil differences affect potato harvest, the experiment was

performed at eight different portions of a field (blocks), with each site large enough for only four treatment combinations. With three fertilizer factors, each site contained only half of the 2^3 . There are at least two options for such an experiment. The first would be to confound **ABC** with blocks in each replicate. Such a design provides the maximum information for all main effects and two-factor interactions, and using the interblock analysis presented for Example 3.4, one could even estimate the three-factor interaction **ABC** from the eight block totals. However, an alternative design option was adopted. Yates chose to change the confounded interaction from replicate to replicate (see Figure 3.6). In replicate 1, **ABC** is confounded with blocks; for replicates 2–4, **ABC** is orthogonal to blocks, since in these replicates a different two-factor interaction was confounded with blocks. Thus, we may use three replicates to estimate each interaction and all four replicates to estimate the main effects.

Treatment combination	Replicate I		Replicate II		Replicate III		Replicate IV		Mean	
	Block 1 ABC = -1	Block 2 ABC = 1	Block 3 AB = 1	Block 4 AB = -1	Block 5 AC = 1	Block 6 AC = -1	Block 7 BC = 1	Block 8 BC = -1		
(1)	101		106			87		131		106.25
a		106			89		128	103		106.50
b		265			272	279			302	279.50
ab	291		306				334		272	300.75
c		312	324				323		324	320.75
ac	373			338	324				361	349.00
bc	398			407		423	445			418.25
abc		450	449			471		437		451.75
Mean	290.75	283.25	296.25	276.50	290.25	302.00	279.00	314.75		291.59

Fig. 3.6. Yates's (1937) potato yield experiment

By fitting a model containing the nominal blocks (numbered 1–8, and designated categorical) plus all the factorial effects, we obtain the analysis of variance presented in Table 3.9 and the parameter estimates presented in Table 3.10. The sum of squares in Table 3.9 do not sum to the total sum of squares because blocks and interactions are not orthogonal to one another; these 'partial' sum of squares (SS) represent the additional variation that is explained by a term in the model if it were the last term to be included. The estimates for interactions are similarly adjusted for differences between blocks (and vice versa). For instance, the estimate for **BC** is based on data from the first three replicates, where **BC** is orthogonal to blocks. From replicates I–III, the estimates are $-186/8 = -23.2$, $-189/8 = -23.6$ and $-151/8 = -18.9$, respectively, and the least squares estimate for β_{BC} is $(-186 - 189 - 151)/24 = -21.9$. Correspondingly, the standard error for each interaction coefficient is calculated as $\text{RMSE}/(24)^{1/2} = 3.65$, rather than $\text{RMSE}/(32)^{1/2} = 3.16$. These larger standard errors reflect the loss of information for each interaction

due to the confounding. From Tables 3.9 and 3.10, we have no evidence of block differences. In this case, it appears that we might have conducted the experiment using four complete blocks of size 8 rather than eight incomplete blocks. However, this is hindsight. The use of smaller blocks provides more insurance against potential variation. If smaller (four plot) sites produces more homogeneous conditions than is available for larger (eight plot) sites, then using incomplete blocks is beneficial.

Table 3.9. ANOVA for Example 3.3

Source	df	SS	MS	F-Ratio	p-Value
Blocks	7	2,638.5	376.9	1.18	.3636
A	1	3,465.3	3,465.3	10.86	.0043
B	1	161,170.0	161,170.0	505.21	.0000
C	1	278,817.8	278,817.8	873.99	.0000
Interactions	4	13,404.4	3,351.1	10.50	.0002
Error	17	5,423.3	319.0		
Total	31	466,779.7			

Table 3.10. Parameter estimates for Example 3.3

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	291.59	3.16	92.35	.0000
Block[1]	-2.22	9.11	-0.24	.8106
Block[2]	-6.97	9.11	-0.76	.4550
Block[3]	3.57	9.11	0.39	.6999
Block[4]	-14.01	9.11	-1.54	.1427
Block[5]	-10.01	9.11	-1.10	.2874
Block[6]	19.07	9.11	2.09	.0517
Block[7]	9.32	9.11	1.02	.3207
Block[8]	1.24	9.11	0.14	.8934
A	10.41	3.16	3.30	.0043
B	70.97	3.16	22.48	.0000
C	93.34	3.16	29.56	.0000
AB	1.08	3.65	0.30	.7700
AC	8.67	3.65	2.38	.0295
BC	-21.92	3.65	-6.01	.0000
ABC	-1.38	3.65	-0.38	.7107

3.4.4 Example 3.4: Analysis for a randomized block experiment, using both within-block and interblock information

Sheesley (1985) presented results for a confirmation experiment involving a new type of welded filament for use in light bulbs. The new type was expected to produce fewer missed lead failures during bulb assembly, because it typically had smaller weld knots. At the time, the manufacturer used both standard and high-speed assembly machines; each was involved in the comparison of the new welded filament versus the standard welded filament. This 2^2 factorial in Wire type (**A**) and Machine type (**B**) was arranged in a blocked design. Each replicate was assigned to a pair of adjacent days such as the following:

- Day 1: Standard machine with old wire type; high-speed machine with new wire type
- Day 2: Standard machine with new wire type; high-speed machine with old wire type

This experiment was repeated for four replicates. Note that the main effect contrasts are orthogonal to days, since each day we include both machine types and both wire types. Thus, the Machine*Wire (**AB**) interaction is confounded with blocks. For simplicity, we present here only the data for the first shift from a single facility. The response is number of defects per hour for a nominal 8-hour shift; see Table 3.11. Since these data are a Poisson-type count, we take $y = (\text{defect rate})^{1/2}$ to stabilize the variance. With counts of 50 or more for each shift, there is no need for the more complex Freeman–Tukey transformation discussed in Section 2.8.2.

Table 3.11. Sheesley lightbulb defects data for one facility

Replicate	Day	Wire type	Machine type	Defects/Hour
1	1	New	High-speed	6.7
	1	Old	Standard	17.6
	2	Old	High-speed	40.5
	2	New	Standard	12.4
2	3	New	High-speed	18.6
	3	Old	Standard	19.2
	4	Old	High-speed	37.8
	4	New	Standard	7.8
3	5	New	High-speed	12.7
	5	Old	Standard	21.4
	6	Old	High-speed	25.1
	6	New	Standard	25.6
4	7	New	High-speed	13.1
	7	Old	Standard	22.7
	8	Old	High-speed	49.4
	8	New	Standard	11.2

If we fit a model with effects for Days, Wire, and Machine, the analysis is straightforward. The resulting ANOVA for (defect rate)^{1/2} is provided in Table 3.12. This provides conclusive evidence for a statistically significant difference due to Wire type, whereas the main effect for Machine is not statistically significant.

Table 3.12. Initial ANOVA for Sheesley data

Source	df	SS	MS	F-Ratio	p-Value
Day	7	39.0185	5.57407	0.8386	.5933
Wire type	1	93.8556	93.8556	14.1206	.0094
Machine type	1	17.7487	17.7487	2.6703	.1534
Error	6	39.8804	6.6467		
Total (corrected)	15	190.5032	0.1367		

Now what about the Wire*Machine interaction? This term was not included in the above analysis because it is confounded with Days. To investigate the importance of a Wire*Machine interaction, we must fit a full factorial model in Replicate, Wire type, and Machine type. The resulting sums of squares and mean squares are presented in Table 3.13, rearranged to facilitate our analysis. The first three lines partition the between-day variation. Note that those sum of squares sum to (4.90 + 25.77 + 8.35 =) 39.02, the sum of squares for Day. In addition, the SS for the last two lines of Table 3.13 sum to 39.88, the error sum of squares from the previous ANOVA. Thus, Table 3.12's error mean square is the average of the Rep*Wire and Rep*Machine mean squares. This reflects the magnitude of error within a day and was used to test for Wire and Machine main effects in the initial ANOVA.

Table 3.13. Complete ANOVA for Sheesley data

Source	df	SS	MS
<u>Day-to-day differences:</u>			
Rep	3	4.8972	1.6324
Wire*Machine	1	25.7704	25.7704
Rep*Wire*Machine	3	8.3509	2.7836
<u>Within-day differences:</u>			
Wire	1	93.8556	93.8556
Machine	1	17.7487	17.7487
Rep*Wire	3	17.7078	5.9026
Rep*Machine	3	22.1726	7.3908

The Wire*Machine estimate is subject to more error variation than are the Wire and Machine main effects, since it is confounded with random day-to-

day differences—variation that is captured in the Rep*Wire*Machine mean square. The Rep*Wire*Machine variation may be understood in two ways. First, it captures inconsistency of the Wire*Machine interaction estimate from replicate to replicate. Second, it represents day-to-day differences in defect rate that are not attributable to the replicate main effect or the Wire*Machine interaction; it is the residual day-to-day variation and is used as the interblock error term for the Wire*Machine interaction effect. Here, the F -test for a Wire*Machine interaction is

$$F = 25.77/2.78 = 9.26,$$

which, with degrees of freedom of (1, 3), has a p -value of .056. Thus, the evidence for a Wire*Machine interaction is inconclusive. Since the replicate mean square (1.63) and day-to-day error mean square (2.78) are unexpectedly smaller than the within-day error mean square of 6.65, we gladly find no evidence for random day-to-day variation.

In the preceding example, the interaction effect was estimable only by treating day-to-day differences as a source of (random) variation. The benefit of using such block-to-block differences to estimate factorial effects confounded with blocks is greatest when the block sizes are small, since, in that case, a substantial percentage of the degrees of freedom are devoted to between-block differences. For example, for blocks of size 2, half of the degrees of freedom are between blocks. Yang and Draper (2003) illustrated computation of within-block and interblock estimates for an example with partial confounding. Their helpful numerical example is a replicated 2^3 factorial in 12 blocks of size 2 using one of their recommended designs. In the next section, we consider designs that confound main effects with blocks and, once again, obtain estimates for these effects by treating block-to-block differences as a second source of variation. Section 14.5 gives an introduction to mixed-model analysis, which provides a unifying foundation for all analyses of experiments containing block-to-block and within-block variation.

3.5 Split-Unit Designs

Split-unit designs are simply factorial designs in incomplete blocks where some main effects are confounded with blocks. Such designs are often much simpler to conduct, which justifies the corresponding sacrifice of precision for some factorial effects. Split-unit designs are most common in agricultural applications, where they are customarily named split-plot designs. To clarify the ideas, we present the following industrial example.

Bisgaard, Fuller, and Barrios (1996) described an experiment with plasma-treated paper. Four factors involving the reactor settings were studied: Pressure (**A**), Power (**B**), Gas flow rate (**C**), and Gas type (**D**). For the three quantitative factors, the levels are simply described as “low” and “high”; the

levels for Gas type are oxygen and SiCl₄, which we designate by -1 and $+1$, respectively. The sequence for the 2⁴ treatment combinations for the reactor was determined by randomization. A fifth factor studied was Paper type (**E**). Prior to each of the 16 reactor runs, 2 paper samples were placed in the reactor, 1 of each type. Following each run, the paper samples were removed and measured for “wettability” using a special microscope to determine the contact angle between a water drop and the paper surface. The coded treatment combinations and the wettability measurements are reported in the next subsection (Table 3.14), where they will be analyzed.

The experiment just described required essentially half the time to conduct that a completely randomized 2⁵ would have required. By collecting data on 2 paper specimens for each run, the reactor only had to be set up and operated 16 times. Since runs required up to 30 minutes simply to achieve a vacuum in the reactor before creating the plasma, halving the number of runs resulted in a substantial reduction in effort.

We now introduce some terminology that will help us describe split-unit experiments. Note that each run for the reactor could be consider a block of size 2. Here each block corresponds to a different combination for (**A**, **B**, **C**, **D**). Within each block, the two observations correspond to the two levels for **E**. Split-unit experiments involve two types of experimental units, commonly called whole units and split units. The reactor runs are called whole units rather than blocks. The factors **A–D** are named whole-unit factors, since they are confounded with whole units. The levels of whole-unit factors do not change for the runs within a whole unit. Factors such as **E** whose levels vary within each whole unit are called split-unit factors. Here, each piece of paper corresponds to a split unit.

Associated with each type of unit is a source of error. Here, split-unit error is a composite of sources that vary within a whole unit. For the plasma-treatment experiment, split-unit error is produced by differences within the reactor where the paper samples are located during a run, as well as random differences in the paper and measurement error. Other sources of error would be associated with the whole-unit variance component, namely run-to-run differences for the setup and operation of the reactor.

Let σ_w and σ_s denote standard deviations associated with the whole-unit and split-unit errors, and let M and N denote the number of whole units and the number of split units in the experiment, respectively. For our example, $M = 16$ and $N = 32$. Assuming that the error for each observation is the sum of a whole-unit error and a split-unit error, then differences between observations within a whole unit are unaffected by the whole-unit error. Thus, for any factorial effect that is based on within-whole-unit differences, the standard error for its regression coefficient will be $\sigma_s/N^{1/2}$. However, factorial effects that are confounded with whole units must be estimated from differences among the M whole units. Such regression coefficients will have a standard error of

$$\sqrt{\sigma_w^2/M + \sigma_s^2/N}. \quad (3.3)$$

The analysis must take into account that some effects have standard error (3.3) and others have standard error $\sigma_s/N^{1/2}$.

3.5.1 Analyzing split-unit designs

Equally replicated 2^k factorial experiments run as split-unit designs are straightforward to analyze, provided all of the effect estimates are based either on within-whole-unit differences or on whole-unit totals. The key to analyzing these split-unit designs is to separate the estimates into two groups. Effects based on differences within whole units are referred to as split-unit contrasts, and these will typically have less error than effects based on differences between different whole units. The analysis changes when there is no replication, as was the case for completely randomized designs. Example 3.5 is an unreplicated design, and so Lenth's method is utilized to construct two PSEs: one for whole-unit effects and another for split-unit effects. At the conclusion of this example, we describe how the analysis would have changed if we had replication.

3.5.2 Example 3.5: Analysis of split-unit design with 16 whole units of size 2

Table 3.14 contains the data for the plasma-treated paper experiment. The actual run order for the 16 whole units is not reported in Bisgaard et al. (1996), so we have chosen to sort the treatment combinations from high to low based on the whole-unit totals.

If we fit a full factorial model for the four whole-unit factors (**A**, **B**, **C**, **D**) to the 16 whole-unit means, we obtain estimates for the whole-unit effects. This saturated model produces a PSE = 2.475. The estimates and Lenth t statistics are displayed in Table 3.15 and plotted in Figure 3.7. The three Lenth t statistics that exceed 2.156 (the $c_{.05}^{\text{IER}}$ critical value from the table in Appendix C) correspond to a full factorial model in Pressure (**A**) and Gas type (**D**), and the next largest Lenth t is 1.38 for the four-factor interaction. Thus, the choice of a reduced model for whole-unit effects is clear.

Table 3.14. Contact angles for plasma-treated paper split-unit experiment

A	B	C	D	E = -1	E = +1	Mean
-1	1	-1	-1	55.8	62.9	59.35
1	-1	-1	1	56.8	56.2	56.50
-1	-1	-1	-1	48.6	57.0	52.80
1	1	-1	-1	53.5	51.3	52.40
-1	1	1	-1	47.2	54.6	50.90
1	1	1	1	49.5	48.2	48.85
1	1	1	-1	48.7	44.4	46.55
1	-1	1	-1	47.2	44.8	46.00
1	-1	1	1	47.5	43.2	45.35
-1	-1	1	-1	37.6	43.5	40.55
1	1	-1	1	41.8	37.8	39.80
1	-1	-1	-1	41.2	38.2	39.70
-1	1	-1	1	25.6	33.0	29.30
-1	-1	1	1	13.3	23.7	18.50
-1	1	1	1	11.3	23.9	17.60
-1	-1	-1	1	5.0	18.1	11.55

Table 3.15. Saturated whole-unit model for plasma split-unit experiment

Term	Estimate	PSE	Lenth <i>t</i>
Intercept	40.98	2.475	16.56
A	5.91	2.475	2.39
B	2.11	2.475	0.85
C	-1.69	2.475	-0.68
D	-7.55	2.475	-3.05
AB	-2.11	2.475	-0.85
AC	1.49	2.475	0.60
AD	8.28	2.475	3.35
BC	-0.42	2.475	-0.17
BD	-1.66	2.475	-0.67
CD	0.84	2.475	0.34
ABC	1.43	2.475	0.58
ABD	-1.65	2.475	-0.67
ACD	-1.16	2.475	-0.47
BCD	0.62	2.475	0.25
ABCD	3.43	2.475	1.38

The next task is to select a model for the split-unit effects [i.e., the effects involving Paper type (**E**)]. If we fit a model that contains only the split-unit factor **E** and the 15 interactions that involve this factor, then we may use

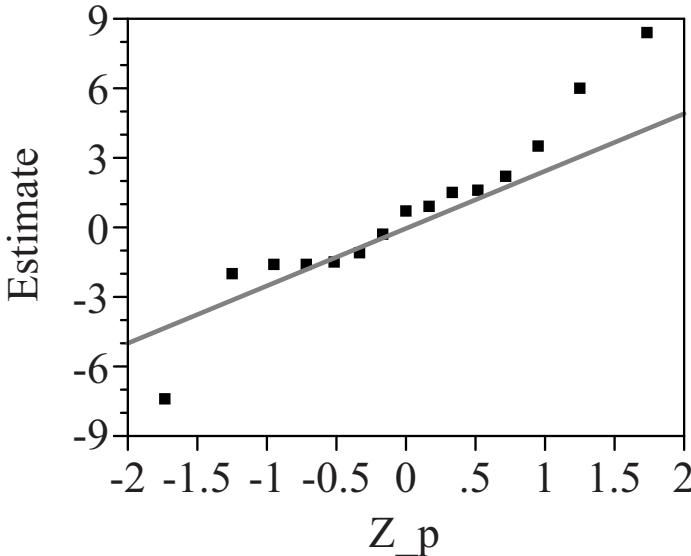


Fig. 3.7. Normal effects plot for plasma experiment whole-unit estimates

Lenth's PSE from this model to estimate the standard error $\sigma_s/N^{1/2}$. To make the details very clear, we show some of the intermediate computations in Table 3.16. Figure 3.8 provides a normal plot for the split-unit effect estimates. The largest three estimates are statistically significant, although b_{DE} is so small that it might be ignored.

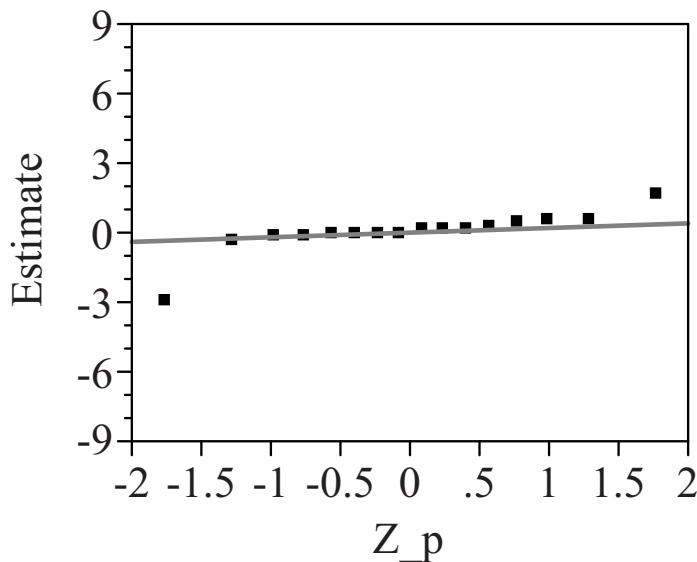
Including three whole-unit effects and three split-unit effects, all statistically significant, the reduced model indicated by our analysis is

$$y = \beta_0 + \beta_A \mathbf{A} + \beta_D \mathbf{D} + \beta_E \mathbf{E} + \beta_{AD} \mathbf{AD} + \beta_{AE} \mathbf{AE} + \beta_{DE} \mathbf{DE} + \epsilon_w + \epsilon_s$$

(although one might include the nonsignificant **B** and **C** main effects rather than allow these to contribute to whole-unit error). Because the model has two error terms, the usual residual plot for the reduced model is not useful for assessing the suitable fit. Instead, we construct two residual plots: one for the whole-unit error and a second for the split-unit error. The whole-unit residual plot is constructed by computing the mean for each whole unit and fitting a reduced model with only the necessary whole-unit effects. This model will have either 12 or 10 df for error, depending on whether we include only {**A**, **D**, **AD**} or also include **B** and **C**. This residual plot appears in Figure 3.9. To isolate the split-unit error, we fit a new model to the 32 observations that represents a saturated model for the whole-unit factors and a reduced model for the split-unit effects; that is, we fit a full factorial in {**A**, **B**, **C**, **D**}, plus the **E**, **AE**, and **DE** terms. The residuals from this fitted model are plotted in Figure 3.10. Both residual plots appear to show homogeneity of variability.

Table 3.16. Sorted split-unit estimates for plasma experiment, with computation of split-unit PSE

Term	Estimate	s_0	PSE	Lenth t
AE	-2.9500	0.234	0.216	-13.68
E	1.5687	0.234	0.216	7.28
DE	0.5125	0.234	0.216	2.38
BCE	0.4500	0.234	0.216	2.09
BCDE	0.4437	0.234	0.216	2.06
ADE	-0.4062	0.234	0.216	-1.88
ABCE	-0.2188	0.234	0.216	-1.01
CDE	0.1625	0.234	0.216	0.75
BE	-0.1500	0.234	0.216	-0.70
ABDE	0.1375	0.234	0.216	0.64
ACDE	-0.1313	0.234	0.216	-0.61
ABCDE	0.1250	0.234	0.216	0.58
BDE	-0.0937	0.234	0.216	-0.43
ACE	-0.0875	0.234	0.216	-0.41
CE	-0.0688	0.234	0.216	-0.32
ABE	0.0563	0.234	0.216	0.26

**Fig. 3.8.** Normal effects plot for plasma experiment split-unit estimates

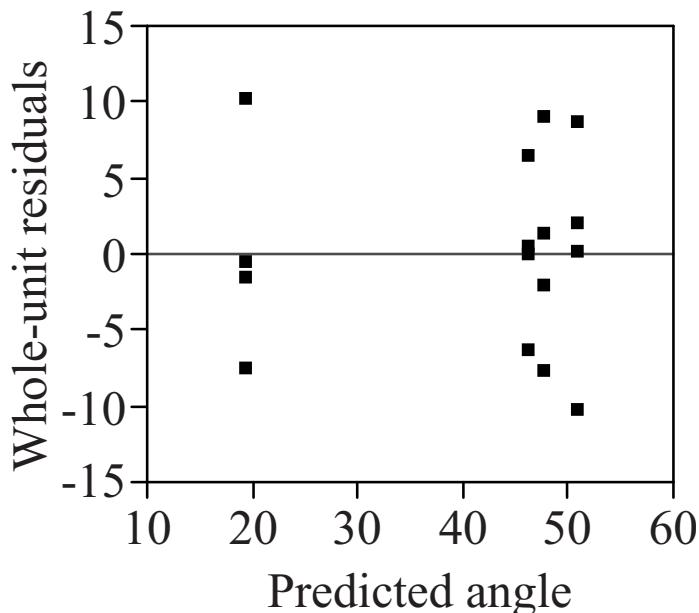


Fig. 3.9. Residual plot for plasma experiment whole-unit errors

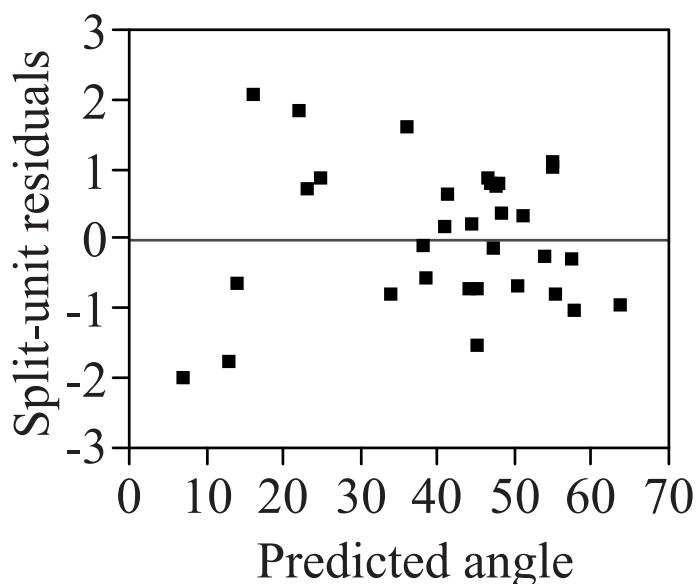


Fig. 3.10. Residual plot for plasma experiment split-unit errors

The split-unit regression coefficients are estimated with greater precision than are the whole-unit regression coefficients, unless $\sigma_w = 0$. How does the precision of these coefficients compare with what would have resulted had we conducted a completely randomized design? If the error variance σ^2 associated with a completely randomized design equals the sum of σ_w^2 and σ_s^2 , then comparisons can be easily made. To make this discussion concrete, suppose $\sigma_w^2 = 15$ and $\sigma_s^2 = 2$. Then for a split-unit design with $M = 16$ and $N = 32$, the standard error for whole-unit and split-unit factorial effects equal 1 and 0.25, respectively. Under the assumption that $\sigma^2 = \sigma_w^2 + \sigma_s^2 = 15 + 2 = 17$, the standard error of 1 and 0.25 correspond to completely randomized designs of sizes 17 and 68, respectively; that is, having runs with two types of paper adds very little information for the whole-unit factors, but it adds a great deal of information for the split-unit effects. Although the actual equivalent sample sizes for a completely randomized design depend on the ratio of σ_w and σ_s , the equivalent sample size for whole-unit effects will always be between M and N , whereas the equivalent sample size for split-unit effects will always exceed N . Recall again that this assumes the completely randomized design's error variance is the sum of σ_w^2 and σ_s^2 . If $\sigma^2 < \sigma_w^2 + \sigma_s^2$, then the benefits of the split-unit design are reduced.

Suppose that after the data in Table 3.14 were collected, a second replicate of the same experiment was performed, using a new random ordering for the furnace runs and for placement of Paper type within runs. This results in a replicated 2^5 with $N = 32$ whole units and $M = 64$ split units. There are $N - 1 = 31$ df for whole-unit variation and $M - N = 32$ df for split-unit variation. For the whole-unit effect tests, one may fit a full factorial model in Replicate and **A–D** and compute Lenth's PSE from these 31 whole-unit contrasts, or one may pool the 15 interactions involving Replicate to form a MSE for whole-unit contrasts. Similarly, for the 16 split-unit effects (**E** and every interaction involving **E**), one may create a mean square for split-unit contrasts by pooling together the 16 interactions involving **E*Replicate**, or one may compute Lenth's PSE from the 32 contrasts involving **E**. If $M/N > 2$, then Lenth's method is no longer applicable, and one constructs an ANOVA with one mean square for whole-unit error and another for split-unit error. An analysis using the mean for each whole unit readily provides the appropriate whole-unit error mean square, and, upon fitting a reduced model, the whole-unit residuals. Then, to analyze the split-unit effects, always include a term called "whole units" with N levels, so that all whole-unit variation is excluded from the split-unit error.

3.6 Multiway Blocking

Sections 3.3–3.5 address design and analysis where the runs within each block (or whole unit) are completely randomized. Here, we enumerate more elaborate restrictions to the treatment combinations for 2^k factorial designs. The

prevalence of these designs in practice is field-specific. For instance, dietary studies for animals are frequently based on crossover studies in which individual animals are assigned different treatment combinations across multiple time periods. This class of experiments, where two or more blocking variables have a crossed (i.e., factorial) structure, is described and illustrated in Section 3.6.1. Blocking can also be sequential (synonyms are *nested* and *hierarchical*). This is the unit structure for split-split-unit designs that are somewhat common for industrial experiments for processes involving multiple steps. Such design structures are discussed in Section 3.6.3. In these sections we focus primarily on the designs themselves. Only one numerical example is analyzed in each of the following sections. However, the same blocking structures appear in Chapter 10—the parallel material for fractional factorial designs. For more elaboration about unit structures, see Brien and Bailey (2006) and the subsequent discussion.

3.6.1 Crossover, Latin squares, and other designs with crossed blocking factors

Crossover designs are the most common type with crossed blocking factors. Consider the experiment displayed in Figure 3.11 involving three factors and two 4×4 Latin squares. Note that each square contains four replicates of half of the treatment combinations in the 2^3 . The three-factor interaction is confounded with squares and so is not estimable. The use of crossover designs is common in animal science, where different animals are assigned to the columns and different time periods are assigned to the rows.

Square 1 ($\mathbf{ABC} = -1$)			
(1)	bc	ac	ab
ac	(1)	ab	bc
bc	ab	(1)	ac
ab	ac	bc	(1)

Square 2 ($\mathbf{ABC} = +1$)			
abc	a	b	c
b	abc	c	a
a	c	abc	b
c	b	a	abc

Fig. 3.11. Latin square design for replicated three-factor experiment

Yang, Beauchemin, and Rode (2001) conducted an experiment based on Plan 8.1b from Cochran and Cox (1957, p. 328), which is similar to Figure 3.11. Yang et al. (2001) studied three dietary factors on the milk production of cows: Grain (**A**), Forage (**B**), and Forage/Concentrate Ratio (**C**). Rather than assigning one treatment combination to each cow (for which cow-to-cow variation would make σ large), a four-period crossover study was performed, with each cow receiving four treatment combinations. Yang et al. ran one square with four cows for four 3-week periods. Although the diet was controlled for 3-week periods, milk production data was recorded for only the last 10 days of each period (hopefully to eliminate any lingering effects from the previous diet condition). Later, they performed a second square with four cows, two of which were the same as in the first square. The exact design, kindly furnished by the authors, appears in Table 3.17, along with two response variables: mean milk production and milk fat %, computed across each set of 10 days. Note that each main effect and two-factor interaction involves differences that cancel out any (additive) cow or time period effects. These data are analyzed later in this section.

There is a potentially important advantage to the design in Figure 3.11 versus the Cochran and Cox (1957) design in Table 3.17. If the rows of Figure 3.11 denote consecutive periods, then every ordered pair of treatment combinations appears once in each square. Such squares, originally due to Williams (1949), are designed to be balanced with respect to simple carryover effects. With four treatment combinations, there are 12 ordered pairs (such as “ac, bc”), and each appears in one of the columns. Such balance is particularly important for sensory testing, where often there are many subjects assigned by repeated use of a few standard designs (Wakeling, Hasted, and Buck 2001). A carryover balanced 8×8 square for a 2^3 is given in Figure 3.12.

Cochran and Cox (1957, p. 328) provides 8×8 plans for four to six two-level factors. When each treatment combination appears once in each row and column of a square, it is a Latin square design. If instead each treatment combination appears in only one-half or one-fourth of the rows and columns, it is labeled a quasi-Latin square design. Examples of quasi-Latin square designs are shown in Figures 3.13 and 3.14. These designs are not balanced for carryover effects. Figure 3.13 provides four replications of a 2^4 factorial and Figure 3.14 is two replicates of a 2^5 . For the four-factor design, each pair of columns forms one replicate, and **ABCD** is confounded with columns in every replicate. Each pair of rows also forms one replicate, but a different three-factor interaction is confounded with rows for each replicate. Thus, each three-factor interaction is partially confounded with rows, being orthogonal to rows in three of the four replicates. If this design were used for a taste test experiment, then 8 sequences will be utilized and each subject will taste 8 of the 16 treatment combinations. For advice regarding the randomization of designs involving such complicated restrictions, see Grundy and Healy (1950).

Table 3.17. Milk production data from Yang et al. (2001)

Square	ID	Cow			A	B	C	Milk	Butterfat
		Period	t.c.					(kg/day)	(%)
1	450	1	ac	Fine	Long	High		25.66	3.69
1	450	2	(1)	Coarse	Long	Low		24.43	4.40
1	450	3	ab	Fine	Short	Low		21.70	4.05
1	450	4	bc	Coarse	Short	High		20.74	3.61
1	458	1	bc	Coarse	Short	High		24.97	4.45
1	458	2	ab	Fine	Short	Low		24.89	4.29
1	458	3	(1)	Coarse	Long	Low		20.66	4.74
1	458	4	ac	Fine	Long	High		20.25	4.71
1	510	1	ab	Fine	Short	Low		23.67	3.76
1	510	2	bc	Coarse	Short	High		17.34	4.64
1	510	3	ac	Fine	Long	High		17.04	4.15
1	510	4	(1)	Coarse	Long	Low		14.46	4.52
1	513	1	(1)	Coarse	Long	Low		29.09	2.76
1	513	2	ac	Fine	Long	High		25.91	3.14
1	513	3	bc	Coarse	Short	High		22.49	2.93
1	513	4	ab	Fine	Short	Low		21.73	3.08
2	458	5	a	Fine	Long	Low		19.15	4.57
2	458	6	c	Coarse	Long	High		16.84	5.04
2	458	7	b	Coarse	Short	Low		15.84	4.68
2	458	8	abc	Fine	Short	High		15.08	4.69
2	513	5	b	Coarse	Short	Low		21.24	3.06
2	513	6	abc	Fine	Short	High		19.88	2.89
2	513	7	a	Fine	Long	Low		20.64	3.00
2	513	8	c	Coarse	Long	High		17.26	3.31
2	525	5	c	Coarse	Long	High		33.74	4.16
2	525	6	b	Coarse	Short	Low		36.58	3.34
2	525	7	abc	Fine	Short	High		35.02	3.48
2	525	8	a	Fine	Long	Low		35.34	3.36
2	528	5	abc	Fine	Short	High		29.99	4.10
2	528	6	a	Fine	Long	Low		31.00	3.43
2	528	7	c	Coarse	Long	High		27.46	3.81
2	528	8	b	Coarse	Short	Low		26.38	3.58

For the five-factor design, the first four columns (rows) constitute one replicate and the last four columns (or rows) constitute a second replicate. For each replicate, a different set of 3 factorial effects is used to partition the 32 treatment combinations into 4 sets of 8.

Period	1	c	a	ab	bc	b	(1)	ac	abc
2	(1)	ac	abc	c	a	ab	bc	b	
3	abc	c	a	ab	bc	b	(1)	ac	
4	ac	abc	c	a	ab	bc	b	(1)	
5	ab	bc	b	(1)	ac	abc	c	a	
6	bc	b	(1)	ac	abc	c	a	ab	
7	a	ab	bc	b	(1)	ac	abc	c	
8	b	(1)	ac	abc	c	a	ab	bc	

Fig. 3.12. Carryover balanced Latin square design for replicated 2^3 experiment

ABCD		ABCD		ABCD		ABCD		
c	abcd	b	ad	a	bd	abc	cd	
abd	(1)	bcd	bc	acd	ac	d	ab	ABC
d	bc	a	abcd	b	cd	abd	ac	
bcd	ad	acd	bd	abc	ab	c	(1)	ABD
a	bd	c	ab	d	abcd	acd	bc	
abc	ac	abd	cd	bcd	(1)	b	ad	ACD
b	ab	d	ac	c	ad	bcd	abcd	
acd	cd	abc	(1)	abd	bc	a	bd	BCD

Fig. 3.13. Quasi-Latin square design for replicated 2^4 experiment

ACE, BCD, ABDE				ACD, BDE, ABCE				
(1)	abe	bc	ace	abd	acd	bcde	de	
bce	ac	e	ab	bcd	d	abde	acde	ABC, ADE,
cde	abcd	bde	ad	abce	ae	b	c	BCDE
bd	ade	cd	abcde	be	ce	abc	a	
abc	ce	acde	bcd	(1)	bde	ad	abe	ABD,
acd	bcde	abce	c	ade	ab	e	bd	BCE,
abde	d	a	be	cde	bc	ace	abcd	ACDE
ae	b	abd	de	ac	abcde	cd	bce	

Fig. 3.14. Quasi-Latin square design for replicated 2^5 experiment

3.6.2 Example 3.6: Analysis of experiment based on crossed blocking factors

We now illustrate the analysis of experiments with crossed blocking factors for the milk production data in Table 3.17. The ANOVA will partition the variation into five parts:

- Difference between squares (1 df)
- Differences between animals within each square [$2(4 - 1) = 6$ df]

- Differences between periods within each square [$2(4 - 1) = 6$ df]
- Main effects and two-factor interactions involving the three factors (6 df)
- Residual variation ($31 - 1 - 6 - 6 - 6 = 12$ df)

This basic ANOVA appears in Table 3.18. Note that the residual degrees of freedom arise from assuming additivity of animal, period, and factor effects. For simplicity, we ignore that some animals appeared in both squares. For this response, the error variance is estimated to be $1.05 (\text{kg/d})^2$, much smaller than if we had not excluded from error animal-to-animal (or period-to-period) variation. Given this small MSE, the benefits of finely processed grain and low F/C ratio are evident. The opposite levels appear desirable for maximizing fat % (or fat kg/d), especially for long forage. This analysis is left to the reader.

Table 3.18. ANOVA for milk production data from Yang et al. (2001)

Source	df	SS	MS	F-Ratio	p-Value
Model	19				
Square	1	67.31	67.31	64.02	.000
Animal[square]	6	950.28	158.38	150.63	.000
Period[square]	6	120.29	20.05	19.07	.000
Grain	1	9.49	9.49	9.03	.011
Forage	1	0.06	0.06	0.06	.815
F/C	1	9.17	9.17	8.72	.012
Grain*Forage	1	0.68	0.68	0.65	.436
Grain*F/C	1	0.07	0.07	0.06	.807
Forage*F/C	1	0.52	0.52	0.50	.494
Error	12	12.62	1.05		
Total (corrected)	31	1170.49			

3.6.3 Split-unit experiments with multiway blocking

For designs described in Section 3.6.1, each column represents either a block containing a full factorial (as in Figure 3.12) or an incomplete block (as in Figure 3.11, where each column contains half of a full factorial). Rather than completely randomizing the run order within each column, the order is constrained to achieve balance with respect to the row (time period) factor. When each column is an incomplete block, interactions are confounded with columns to determine the subset of treatment combinations in a given column. This is akin to Section 3.3, in which blocks were created using interaction contrasts so that we sacrifice information for effects of little importance.

We now explore an extension of Section 3.5, where for the sake of simplicity of experimentation, main effects are confounded with blocks. Consider factorial experiments with two sets of factors, where the treatment combinations for

one set of factors are assigned to different rows and treatment combinations for the second set of factors are assigned to different columns. Two examples are given, one agricultural and the second industrial. For an experiment involving plots in a field, one set of factors might be different fertilizers and the second set might be different planting and/or variety factors. The fertilizer combinations may be randomly assigned to the columns in a field; that is, each entire column receives one combination of fertilizer levels. Similarly each row in the field receives a particular planting/variety combination. Such designs are commonly labeled “strip block” designs in agriculture. A similar structure can occur in industrial experiments that involve two different stages in a process. For instance, suppose we are investigating the optimum settings for washing and drying of clothes to avoid wrinkling, with eight washing machines and four dryers available. We begin by assigning each washer to a different combination of levels for the washer factors. When the eight loads are completed, the wet clothes from each load are distributed to the four different dryers. Provided the articles of clothing are labeled according to their assigned washer/dryer combination, one may perform $8 \times 4 = 32$ different washer/dryer treatment combinations running each washer and dryer just once. Such a washer/dryer experiment is another example of a strip-block design. The alternative terminology, “multiway split unit” is sometimes used for such industrial experiments, especially when there are more than two stages (sets of factors). The analysis of such experiments is illustrated in Part II (see Section 10.3.4).

How do strip-block and multiway split-unit designs differ from the split unit designs discussed and analyzed in Section 3.5? Both are experiments with a factorial treatment structure and generally a sequential assignment of treatment combinations to the different types of experimental units. The difference regards whether the unit structure is crossed or nested. If the washer loads contain clothes that will be assigned to different dryers, and the dryer loads contain clothes from multiple washer loads, then the washer units and dryer units are crossed. This is the structure for strip-block designs just described. If, instead, the wet clothes from each washer load are assigned to different dryers, but each dryer load contains a small load all from the same washer load, then the dryer loads are nested within washer loads. This is the structure of a typical split-unit design. Just as multiway split-unit designs can have more than two sets of factors assigned to units that have a crossed structure, designs with more than two sets of factors can be assigned to units that are nested in more than two levels. Such designs are called split-split-unit designs, and they are described in the next subsection.

3.6.4 Split-split-unit designs and others with nested blocking factors, with Example 3.7

Consider the following meat loaf experiment conducted by Baardseth, Bjerke, Aaby, and Lielnik (2005). The purpose of the experiment was to identify

factors causing rancidity of meat loaf that has been stored. The production of meat loaf is a multistep process, and this experiment involved factors appearing in different steps. The four factors that preceded making, baking, and storage of cooked loaves were Meat (pork or turkey), Processing (pieces or ground), Salting (before or after storage), and Packaging (air or vacuum sealed). For simplicity, the experiment began with a single large portion of pork and turkey. Each portion was split in half, creating one split portion for each Meat/Processing combination. Similarly each of the four Meat/Processing units was split in half, creating eight units for the Salting phase. Finally, each of these 8 units was split in half, creating 16 storage units. This illustrates what we mean by a nested unit structure: 2 Meat units split to form a total of 4 processing units, 8 salting units and 16 storage units. Such a design is called a split-split-split-unit design, since each final unit results from three sequential splits.

There are both benefits and disadvantages to such a design. In addition to the obvious benefit of convenience is the higher precision associated with factors at the bottom of the hierarchy. For instance, the error associated with the effects involving storage should be quite small. However, the downside of convenience is the lack of replication at the top of the hierarchy of splits. The fact that 16 storage units will provide 16 (or more) measurements of rancidity does not alter the fact that this is an unreplicated design for Meat, with everything based on a single unit of pork and a single unit of turkey. Even if the entire experiment were replicated, using a second unit of each raw material, we would not be able to effectively confirm any Meat factor effect.

We now consider an analysis of the rancidity data from Baardseth et al. (2005). After 3 months of storage, a meat loaf was made from each of the 16 units following a prescribed recipe. Each loaf was baked, refrigerated, frozen for 6 months, and, finally, thawed and tested for rancidity. Coding for the 4 factors is shown in Table 3.19 and the rancidity score for each of the 16 treatment combinations is shown in Table 3.20. We are neglecting some less important details about the experiment to permit its straightforward analysis here. For instance, the scores for the first and last treatment combination are averages of scores from two loaves.

Table 3.19. Coding of factors for the meat loaf experiment

Factors	Levels	
	-1	1
A Meat	Pork	Turkey
B Processing	Pieces	Ground
C Salting	Later	Now
D Packaging	Air	Vacuum

Table 3.20. Rancidity scores for meat loaf experiment

A	B	C	D	Rancidity
-1	-1	-1	-1	2.46
-1	-1	-1	1	2.30
-1	-1	1	-1	10.85
-1	-1	1	1	2.45
-1	1	-1	-1	1.60
-1	1	-1	1	1.90
-1	1	1	-1	4.65
-1	1	1	1	2.50
1	-1	-1	-1	3.40
1	-1	-1	1	3.85
1	-1	1	-1	5.90
1	-1	1	1	3.40
1	1	-1	-1	2.90
1	1	-1	1	3.15
1	1	1	-1	8.30
1	1	1	1	2.05

The analysis of the meat loaf experiment is a simple extension of the analysis presented earlier for split-unit experiments. For Example 3.5 there were two different standard errors, a larger one for whole-unit effect estimates (Table 3.15) and a smaller one for split-unit effect estimates (Table 3.16). Here, there are four standard errors, one for each tier; see Table 3.21. Due to the lack of replication, we rely on Lenth's procedure for estimating the standard errors. Only in the lowest tier can Lenth's procedure be effective. There, two of the eight estimates stand out, **D** ($t = -2.03, p = .06$) and **CD** ($t = -2.22, p = .05$). Assuming independently distributed errors for the different sources of variation, the true standard error is highest for the whole-unit level and decreases as one moves down the tiers. Thus, although four estimates at the split-split-unit level is insufficient to test for significance using Lenth's method, the fact that the estimates for **A**, **B**, and **AB** are all smaller than the estimate for **C** supports the conclusion that **A** and **B** have no apparent effect but that **C** does. The reduced model for expected rancidity score is

$$\hat{y} = 3.854 + 1.159\mathbf{C} - 1.154\mathbf{D} - 1.259\mathbf{CD}.$$

We conclude that the “salting now, air packaging” combination is vastly inferior, with expected score around 7.4, whereas the other three (**C**, **D**) combinations have expected scores of 2.6 to 2.8.

Table 3.21. Rancidity score estimates and Lenth t statistics

Term	Estimate	PSE	Lenth t
Intercept	3.854		
A	0.265		
(Split units)			
B	-0.473	0.695	-0.68
AB	0.454	0.695	0.65
(Split-split units)			
C	1.159	0.608	1.90
AC	-0.365	0.608	-0.60
BC	-0.165	0.608	-0.27
ABC	0.446	0.608	0.73
(Split-split-split units)			
D	-1.154	0.567	-2.03
AD	0.148	0.567	0.26
BD	0.173	0.567	0.30
ABD	-0.666	0.567	-1.17
CD	-1.259	0.567	-2.22
ACD	0.077	0.567	0.14
BCD	0.140	0.567	0.25
ABCD	-0.584	0.567	-1.03

If the estimates for **A** had been large, we would not have known whether this was due to the systematic effect of Meat type or due to random variation affecting the whole units. Determining which was in fact the case would require additional data with replication. Simply adding more levels of splitting, which produces many data points, does not help. However, it does help confirm that whatever the effect is for Meat (or Processing), these effects do not depend on the levels for Salting and Packaging.

Split-split-unit designs have experimental units in a nested structure, with main effects confounded with blocks. It is also possible to have a nested structure for blocks when interactions, not main effects, are confounded with blocks. For instance, suppose an unreplicated 2^5 experiment requires 2 days to complete and that we utilize 4 batches of raw material per day. We partition the 32 runs into 8 blocks of size 4, confounding with batches the 7 factorial effects listed in Appendix E: **ABC**, **ADE**, **BCDE**, **BD**, **ACD**, **ABE**, **CE**. Rather than randomly assigning four of the eight blocks to the first day and the remainder to the second day, we choose to confound the longest interaction **BCDE** with day. In this manner, the other six interactions confounded with batches are orthogonal to days. A normal plot of these six effect estimates might be used to assess whether one of these interactions is in fact active.

Table 3.22. Three sets of effects for Holm and Sidik split-split-unit design

Unit	Effects
Whole	F, G, FG, ABCD, ABCDF, ABCDG, ABCDFG
Split	E, EF, EG, EFG, ABCDE, ABCDEF, ABCDEG, ABCDEFG
Split-split	A, B, C, D and all 108 remaining interactions

Holms and Sidik (1971) proposed a 2^7 split-split plot experiment involving a nuclear reactor. Four fluid variables (**A**, **B**, **C**, **D**) can be easily changed within a fuel cycle. The power (**E**) can be changed within a cycle but not as often. Two mechanical variables (**F**, **G**) can only be changed between cycles. For this scenario, an 8-cycle design is proposed, with 16 runs per cycle. Using eight cycles provides replication of the whole-unit treatment combinations, something that was not the case for the rancidity experiment. Holms and Sidik confounded **F**, **G**, and **ABCD** with cycles. Power is a split-unit factor, being reversed at the midpoint of each fuel cycle. The four fluid factors are split-split-unit factors, since they are allowed to change run to run. The three sets of factorial effects that differ in precision are listed in Table 3.22. Rather than randomly assign whole-unit blocks to the eight cycles, the order for the runs is determined systematically; see Holms and Sidik (1971, p. 569). This was done due to the expected termination of some cycles before all 16 runs are completed. By choosing a systematic run order, they hope to ensure that effects of interest will be estimable even if early termination of some cycles causes data to be missing.

This concludes the discussion of full factorial designs with randomization restrictions. Randomization restrictions are discussed again in Chapter 10, in the context of fractional factorial designs.

More Full Factorial Design Examples

This chapter contains the analysis of three interesting experiments reported in the literature. The sections are as follows:

- Section 4.1. Example 4.1: Replicated 2^3 With Subsampling Within Runs
- Section 4.2. Example 4.2: 2^9 Factorial for Peptide Research
- Section 4.3. Example 4.3: 2^5 with Centerpoint Runs for Ceramic Strength

4.1 Example 4.1: Replicated 2^3 With Subsampling Within Runs

Lamb, Boos, and Brownie (1996) analyzed a replicated 2^3 factorial to illustrate methods for identifying factors affecting variability. The measured response is tablet weight for a pharmaceutical product. The target weight was 0.5 grams, and the objective was to achieve better consistency of tablet weights. The tablets are formed by putting a powder into holes of a disk that turns and compacting the powder twice before it is ejected. The three factors are Turning speed (x_1), Initial compression (x_2), and Final compression (x_3). The actual low and high levels for each factor are not reported in the article by Lamb et al.; we denote the levels as -1 and $+1$, respectively. The 2^3 factorial was replicated six times. Although details are lacking, it appears that the 48 runs were performed by blocking on replicate; that is, the 8 treatment combinations of the 2^3 were each performed once, in random order, to complete the first replicate. This process was repeated to obtain the second replicate, and so forth, until all six sets of eight treatment combinations were obtained. We analyze the data by including a Replicate main effect. Within each run, 25 tablets were sampled and weighed. Such sampling within a run, very different from replication of treatment combinations, is quite common in industrial examples in which the objective is to improve consistency about a target.

We begin the data analysis by plotting the 1200 data points in a boxplot and histogram. The mean and standard deviation are 0.4974 and 0.0072 grams, respectively. Figure 4.1 shows 24 “outliers,” the majority of which are on the low extreme.

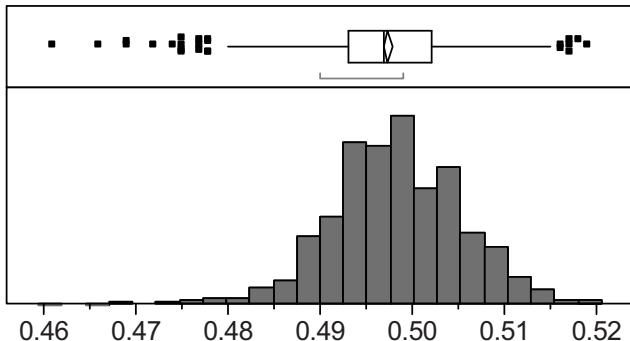


Fig. 4.1. Histogram and box-plot for 1200 tablet weights

There are several questions to be answered with these data:

- Which factors, if any, affect the mean tablet weight?
- Which source of variation is the greatest?
- Which factors, if any, affect the greatest source of variation?
- Are the previous conclusions influenced by just a few outliers?

4.1.1 Which factors, if any, affect the mean?

This question is most easily answered by computing the mean for each sample and analyzing these 48 means using a full factorial model in the 3 factors of interest, augmented with a main effect for Replicates. Note that in this analysis we have one data point (y_i = a sample mean) for each run. Table 4.1 provides these means, along with sample standard deviations, which are used later. Figure 4.2 summarizes this fitted model. The coefficients b_1 (Speed), b_3 (Final compression), and the interaction $b_{1,3}$ are statistically significant and all near 0.0009; this implies that the high-speed, high-final-compression combination produces larger weights on average. We are disappointed to see that Replicate is also significant (Replicate mean square = 0.0000213, $F = 3.27$, $p = .016$), which indicates undesirable variability over time.

Table 4.1. Means and standard deviations for within-run samples of size 25

Replicate	x_1	x_2	x_3	Mean	StdDev	ln(StdDev)
1	-1	-1	-1	0.49568	0.00352	-5.649
1	-1	-1	1	0.49956	0.00503	-5.293
1	1	-1	-1	0.50008	0.00516	-5.266
1	1	-1	1	0.50196	0.00503	-5.293
1	-1	1	-1	0.49664	0.00491	-5.317
1	-1	1	1	0.49992	0.00524	-5.251
1	1	1	-1	0.49760	0.00953	-4.654
1	1	1	1	0.49868	0.00618	-5.086
2	-1	-1	-1	0.49456	0.00471	-5.358
2	-1	-1	1	0.49300	0.00435	-5.438
2	1	-1	-1	0.49732	0.00784	-4.848
2	1	-1	1	0.49632	0.00767	-4.870
2	-1	1	-1	0.49708	0.00654	-5.030
2	-1	1	1	0.49400	0.00362	-5.622
2	1	1	-1	0.49824	0.00614	-5.093
2	1	1	1	0.49152	0.00656	-5.027
3	-1	-1	-1	0.49784	0.00636	-5.058
3	-1	-1	1	0.49960	0.00492	-5.314
3	1	-1	-1	0.49288	0.00900	-4.711
3	1	-1	1	0.50252	0.00773	-4.863
3	-1	1	-1	0.49628	0.00743	-4.902
3	-1	1	1	0.49804	0.00544	-5.213
3	1	1	-1	0.49576	0.00976	-4.630
3	1	1	1	0.50152	0.00635	-5.059
4	-1	-1	-1	0.49896	0.00509	-5.280
4	-1	-1	1	0.49512	0.00448	-5.407
4	1	-1	-1	0.49616	0.00703	-4.957
4	1	-1	1	0.49968	0.00639	-5.053
4	-1	1	-1	0.49256	0.00598	-5.119
4	-1	1	1	0.49456	0.00627	-5.072
4	1	1	-1	0.49080	0.01007	-4.598
4	1	1	1	0.50056	0.00623	-5.079
5	-1	-1	-1	0.49608	0.00457	-5.388
5	-1	-1	1	0.49752	0.00771	-4.865
5	1	-1	-1	0.49944	0.00848	-4.771
5	1	-1	1	0.50080	0.00632	-5.064
5	-1	1	-1	0.50032	0.00589	-5.135
5	-1	1	1	0.50072	0.00672	-5.002
5	1	1	-1	0.49744	0.00948	-4.658
5	1	1	1	0.50312	0.00621	-5.081
6	-1	-1	-1	0.49556	0.00827	-4.795
6	-1	-1	1	0.49164	0.00508	-5.282
6	1	-1	-1	0.49660	0.00613	-5.094
6	1	-1	1	0.50196	0.00552	-5.200
6	-1	1	-1	0.49520	0.00729	-4.922
6	-1	1	1	0.49572	0.00524	-5.251
6	1	1	-1	0.49408	0.01008	-4.597
6	1	1	1	0.50280	0.00569	-5.168

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	12	0.00023944	0.00001995	3.059
Error	35	0.00022827	0.00000652	Prob > F
C. Total	47	0.00046772		0.005

Expanded Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	0.49738	0.00037	1349.30	0.000
Rep[1]	0.00139	0.00082	1.69	0.101
Rep[2]	-0.00212	0.00082	-2.57	0.015
Rep[3]	0.00068	0.00082	0.82	0.415
Rep[4]	-0.00132	0.00082	-1.61	0.117
Rep[5]	0.00206	0.00082	2.49	0.018
Rep[6]	-0.00068	0.00082	-0.82	0.415
x1	0.00087	0.00037	2.36	0.024
x2	-0.00016	0.00037	-0.43	0.667
x3	0.00099	0.00037	2.69	0.011
x1*x2	-0.00041	0.00037	-1.10	0.277
x1*x3	0.00088	0.00037	2.40	0.022
x2*x3	0.00022	0.00037	0.60	0.551
x1*x2*x3	-0.00007	0.00037	-0.20	0.840

Fig. 4.2. Analysis of mean weight per run for 48 runs**4.1.2 Which source of variation is the greatest?**

Sampling 25 tablets within each run enables us to estimate well the short-term variability in weights. Replicating each treatment combination numerous times (presumably spread out over time) permits assessment of run-to-run variation. If one conducts an analysis of variance (ANOVA) of the 1200 tablet weights, using nominal effects for Replicate (5 df), Treatments (7 df), and Replicate*Treatment (35 df), one can quickly identify which sources of variation are largest. Figure 4.3 shows the mean squares for each of these terms. The factor Replicate represents a sampling of the tablet-making process at different points in time. Such factors (and their interactions) are best treated as random effects in a model, as discussed in Section 14.5. If the terms Replicate and Replicate*Treatment are declared random, statistical software allowing this feature will estimate their variances from the mean squares. For these balanced data, method of moments estimates and restricted maximum likelihood (REML) estimates are identical. Note that the residual variance estimate equals the mean square error, and the other two variance components are estimated as follows:

$$\hat{\sigma}_{\text{Rep*Treatment}}^2 = [\text{MS}_{\text{Rep*Treatment}} - \text{MSE}] / 25 = 0.000005,$$

$$\hat{\sigma}_{\text{Rep}}^2 = [\text{MS}_{\text{Rep}} - \text{MS}_{\text{Rep} * \text{Treatment}}] / [8(25)] = 0.000002.$$

Both of these variance component estimates are negligible relative to the residual variance, $\text{MSE} = 0.000044$. Although there is variation over time, as reflected in the variance components involving Replicate, nearly all of the variation in tablet weight for these data is within-run variation. (For more details about inference from models with multiple sources of variation, see Section 14.5.)

Analysis of Variance

<u>Source</u>	<u>DF</u>	<u>Sum of Squares</u>	<u>Mean Square</u>	<u>F Ratio</u>
Model	47	0.01169293	0.000249	5.6251
Error	1152	0.05095032	0.000044	Prob > F
C. Total	1199	0.06264325		<.0001

Effect Tests

<u>Source</u>	<u>DF</u>	<u>Sum of Squares</u>	<u>Mean Square</u>
Rep	5	0.00266599	0.0005332
Treatment	7	0.00332008	0.0004743
Rep*Treatment	35	0.00570686	0.0001631

Variance Component Estimates

<u>Random Effect</u>	<u>Var Component</u>	<u>% of Total</u>
Rep	.000002	3.64
Rep*Treatment	.000005	9.35
<u>Residual</u>	<u>.000044</u>	<u>87.01</u>
Total	.000051	100.00

Fig. 4.3. ANOVA and variance component estimates for tablet weight

4.1.3 Which factors, if any, affect the greatest source of variation?

The above analyses (for the 48 means and the 1200 tablet weights) fit models that assume common variation across all the treatment combinations. Whether this is the case is a question of keen interest, since if some treatment combination has smaller within-run variation, this is an operational advantage. Is the within-run variation impacted by the factors Speed (x_1), Initial compression (x_2), or Final compression (x_3)? We answer this question by plotting the 1200 weights and by modeling the 48 within-run standard deviations in Table 4.1. Figure 4.4 shows a plot of the 1200 tablet weights arranged by treatment combinations. The smallest 17 tablet weights appear in bold in the plot. All come from the higher Speed. This is somewhat surprising, since the tablet weight mean is larger at high Speed.

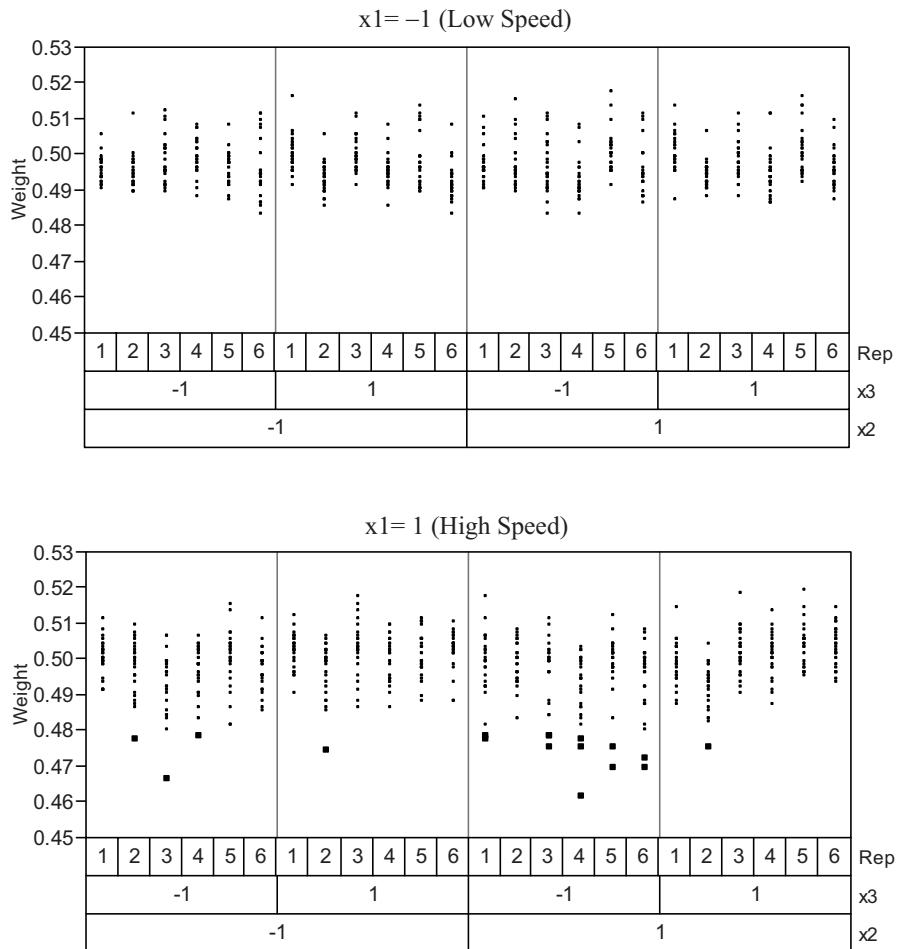


Fig. 4.4. Variability graph for 1200 tablet weights

To see what other factors, if any, affect within-run variability, we model the natural log of the within-run sample standard deviations, which appear in the last column of Table 4.1. (The logarithm is a variance-stabilizing transformation for sample variances and standard deviations; see Section 2.8.3.) Figure 4.5 shows that both x_1 (Speed) and x_3 (Final compression) affect the variability of weight, with higher speed and lower compression producing more variability. Furthermore, $b_2 = 0.053$ and $b_{2,3} = -0.052$ have p -values approaching .05. When Final compression is high, these terms cancel one another, so that the level of x_2 makes no difference. However, when Final compression is low ($x_3 = -1$), the terms combine to equal $0.105x_2$, imply-

ing that high Initial compression and low Final compression is a particularly variable combination.

The F -test for Replicate is nearly significant as well ($F = 2.35$, $p = .061$). The data suggest that averaging across the eight treatment combinations, the first couple of replicates had smaller within-run variability than the later replicates.

Analysis of Variance for $\ln(\text{StdDev})$

<u>Source</u>	<u>DF</u>	<u>Sum of Squares</u>	<u>Mean Square</u>	<u>F Ratio</u>
Model	12	1.8918369	0.157653	4.7142
Error	<u>35</u>	<u>1.1704756</u>	<u>0.033442</u>	Prob > F
C. Total	47	3.0623125		0.0002

Expanded Estimates for $\ln(\text{StdDev})$

<u>Term</u>	<u>Estimate</u>	<u>Std Error</u>	<u>t Ratio</u>	<u>Prob> t </u>
Intercept	-5.077	0.026	-192.336	<.0001
Rep[1]	-0.149	0.059	-2.530	0.016
Rep[2]	-0.084	0.059	-1.424	0.163
Rep[3]	0.108	0.059	1.829	0.076
Rep[4]	0.006	0.059	0.104	0.918
Rep[5]	0.081	0.059	1.377	0.177
Rep[6]	0.038	0.059	0.645	0.523
x1	0.130	0.026	4.928	<.0001
x2	0.053	0.026	2.012	0.052
x3	-0.084	0.026	-3.175	0.003
x1*x2	-0.001	0.026	-0.023	0.982
x1*x3	-0.040	0.026	-1.508	0.141
x2*x3	-0.052	0.026	-1.966	0.057
<u>x1*x2*x3</u>	<u>-0.014</u>	<u>0.026</u>	<u>-0.521</u>	<u>0.605</u>

Fig. 4.5. Analysis of $\ln(\text{StdDev})$ per run for 48 runs

Searching for dispersion effects (i.e., factor effects on the variability) is important for many process improvement applications. This topic is taken up in more detail in Sections 13.3 and 14.3. We are not investigating here whether the run-to-run variance depends on any of the factors because this variance is negligible compared to the within-run variation. Lamb, Boos, and Brownie (1996) did check for inequality of run-to-run variability, assuming no replicate effect, and found no significant differences.

4.1.4 Are the previous conclusions influenced by just a few outliers?

The analysis of 48 means is little affected by the tablets with extremely low weights. Likewise, eliminating the lowest tablet weights would not alter the conclusion that the predominate source of variation is within runs. However,

exclusion of these outlier observations might possibly alter the conclusion that low Speed and high Final compression is necessary to minimize variability.

Tablets with extreme weights ought to be examined physically. Low weights may arise due to low density or due to being undersized or broken. Understanding better the cause for the low weights will supplement the insight from a statistical model that indicates which factor levels are preferred. The lowest 2% ($24/1200$) of the weights all occurred at high Speed and 22 of the 24 occurred at low Final compression. Clearly, the cause is linked to this combination of levels. If one were to eliminate the 17 tablet weights below 0.48 and repeat the analysis of $\ln(\text{StdDev})$, the signs of the main effect estimates are unchanged, but only b_1 is statistically significant.

The preference for low Speed is fully supported by the data, whether we exclude the low weights or not. Considering only the 600 tablets produced at low Speed, the compression factors do not appear to affect the mean or variance for weight. There is some variation from run to run, but the within-run variability remains the dominant variance. The coefficient of variation for weight is about 1.2% at low speed. If this is considered too large, further work to reduce the within-run variability is needed.

4.2 Example 4.2: 2^9 Factorial for Peptide Research

Wang, Dipasquale, Bray, Maeji, and Geysen (1993) presented results involving the binding capacity of the neuropeptide substance P (SP) systematically modified by replacing native L-amino acids with D-amino acids in nine positions of the amino acid sequence of SP. The authors measured the percentage of inhibition of a reagent labeled SP for the 2^9 sequences. These data were analyzed by Young and Hawkins (1995) after correcting two typographical errors in Wang et al.'s (1993) Table 1. We denote the factors by x_i , $i = 1, \dots, 9$. Wang et al. (1993) remarked that x_1-x_5 are N-terminal residues, whereas x_6-x_9 , the last four positions, are C-terminal residues. This is an important distinction in that these factor groups affect the response differently. We code each L-amino acid as the high level +1 and each D-amino acid as the low level -1. Higher inhibition percentages are preferred, and this is expected for sequences with predominantly L-amino acids. The percent inhibition values (P) are given in Table 4.2, where each row corresponds to a different x_1-x_5 combination, and each column corresponds to a different x_6-x_9 combination.

Table 4.2. Percent inhibition values for 512 amino acid sequences from Wang et al. (1993)

	$x_6:-$	+	-	-	-	+	+	-	+	-	+	+	+	-	+
	$x_7:-$	-	+	-	-	+	-	+	-	+	-	+	+	-	+
	$x_8:-$	-	-	+	-	-	+	+	-	-	+	+	-	+	+
	$x_1x_2x_3x_4x_5$	$x_9:-$	-	-	-	+	-	-	+	+	+	-	+	+	+
-----	15	22	17	3	29	1	0	3	5	9	11	45	56	54	86
+-+-+-	10	8	17	3	11	14	0	20	27	7	27	23	37	41	77
-+---	12	33	14	15	19	20	11	2	13	20	26	22	30	42	68
++---	31	10	0	11	15	7	3	0	18	14	9	19	19	22	52
--+--	22	0	0	0	29	18	12	14	15	5	13	36	43	31	70
+--+--	28	3	0	11	2	3	13	24	17	23	19	27	21	16	39
-++--	13	12	10	2	3	4	3	11	5	22	17	0	15	17	80
+++--	0	19	14	11	5	18	0	8	6	12	21	29	39	41	74
--+-+--	20	28	0	31	8	29	16	26	24	37	30	58	74	80	82
+-+-+-	10	0	5	4	2	19	26	14	20	32	24	50	42	60	69
-+-+-+-	23	10	1	4	0	4	0	8	8	25	6	51	55	42	57
++-+-	0	21	13	26	26	9	6	7	5	6	21	42	43	49	69
--++-	17	14	16	20	27	27	23	15	36	26	32	43	60	53	69
+-++-	10	6	11	11	9	13	6	22	17	14	25	8	9	18	41
-+++-	19	2	1	13	5	14	0	15	1	0	0	59	73	73	75
+++-	18	7	11	9	14	3	11	44	12	16	8	72	63	72	76
--+-+--	21	17	14	15	0	21	15	27	22	39	9	59	70	74	73
+-+- -	17	17	0	21	0	5	14	26	12	0	4	32	32	69	0
-+-+-+	17	4	5	3	18	10	10	13	20	14	16	52	47	62	54
++-+-	4	18	18	5	4	20	13	17	16	15	6	48	52	50	36
--+--	18	18	14	4	17	23	33	25	4	24	23	39	34	41	63
+-+--	12	5	11	10	13	18	22	18	23	17	19	0	18	39	36
-++-+	0	0	3	2	16	0	0	0	0	23	32	0	0	29	97
+++-	8	11	12	19	16	29	13	14	46	19	13	28	50	83	97
--+-+ +	0	7	0	2	22	22	10	18	19	15	16	81	33	80	68
+-+-++	7	12	14	17	17	0	0	4	3	19	24	43	64	79	79
-+-+-+	13	6	6	15	17	2	0	17	17	22	20	15	66	76	100
++-++	20	8	3	9	7	25	16	29	31	23	52	49	35	54	67
--++ +	19	7	13	18	21	6	3	14	2	20	17	65	62	67	81
+-++ +	8	12	10	21	18	0	0	0	0	0	0	39	60	91	91
-++ + +	0	0	0	51	62	7	1	0	0	0	0	87	92	95	83
++ + +	14	0	20	4	9	1	0	0	0	0	0	89	80	85	70
Column No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
															16

Whereas the rows of Table 4.2 are arranged in standard order, the columns are sorted by the number of L-amino acids at the C-terminal end. Grouping the columns in this manner reveals the similarity of the data for columns

with like number of L-amino acids at the C-terminal end. Figure 4.6 shows a histogram for P for each group of columns.

Apart from two outliers, note how little variation exists among the 32 responses from Column 16. This alerts us to a feature of the measurements. The percentage inhibition values in Table 4.2 were determined by measuring the amount of bound reagent labeled SP using a multigamma counter. We surmise from the article that this Poisson-type count c was converted to the inhibition percentages in Table 4.2 as

$$P = \max\{0, 100[1 - (c/c_{100\%})]\}, \quad (4.1)$$

where $c_{100\%}$ is the radioactive count using only the reagent labeled SP in the assay; that is, P near 100 corresponds to small counts, and P near 0 corresponds to (more variable) large counts. As discussed in Section 2.8.2, the square root transformation stabilizes the variance for Poisson responses. Thus, we propose using the transformation

$$T = \sqrt{1 - P/100} = \min\{1, \sqrt{c/c_{100\%}}\}. \quad (4.2)$$

If P were confined to the range $[0, 50]$, this transformation would have little effect. However, here P ranges from 0 to 100, and the transformation has the effect of compressing values corresponding to P near zero; see Figure 4.7. When $P = 0$, the actual ratio $c/c_{100\%}$ is not recorded but is censored (i.e., truncated) at 1. This accounts for the prevalence of zeros in Table 4.2. This censoring has a negligible impact on the response, especially after our transformation. Furthermore, the investigators were disinterested in peptides with $P < 25\%$.

4.2.1 Choosing a reduced model for Wang's 2^9 factorial data

Suppose we fit a saturated model with 511 terms for the response T . Lenth's PSE = 0.00412, and 54 of the estimates exceed 1.97(.00412), where $c_{.05}^{\text{IER}} = 1.97$ is the .05 IER critical value from the first table in Appendix C. If there were no systematic effects, one would expect about half that number, since $(.05)511 = 25.55$. With such a large number of potential Type I errors, we should be more stringent in declaring effects to be statistically significant, lest we complicate the model unnecessarily. One approach is to control the EER; the other is to control the false discovery rate, based on the method described in Section 14.2.2. Using Appendix C's EER critical value of $c_{.20}^{\text{EER}} = 3.572$, so that the probability of making one or more Type 1 errors experimentwise is at most 20%, we find 20 significant effects (see Table 4.3). For controlling the false discovery rate (FDR) for this example, see Section 14.2.2. The p -values in Table 4.3, which are needed for the FDR procedure, are based on 40,000 simulations of 511 estimates. None of the $511 \times 40,000 = 20.44$ million estimates exceeded 6.23, the 12th largest Lenth t in Table 4.3. The last column is the p -value based on the $t_{511/3}$ distribution; the agreement is remarkable.

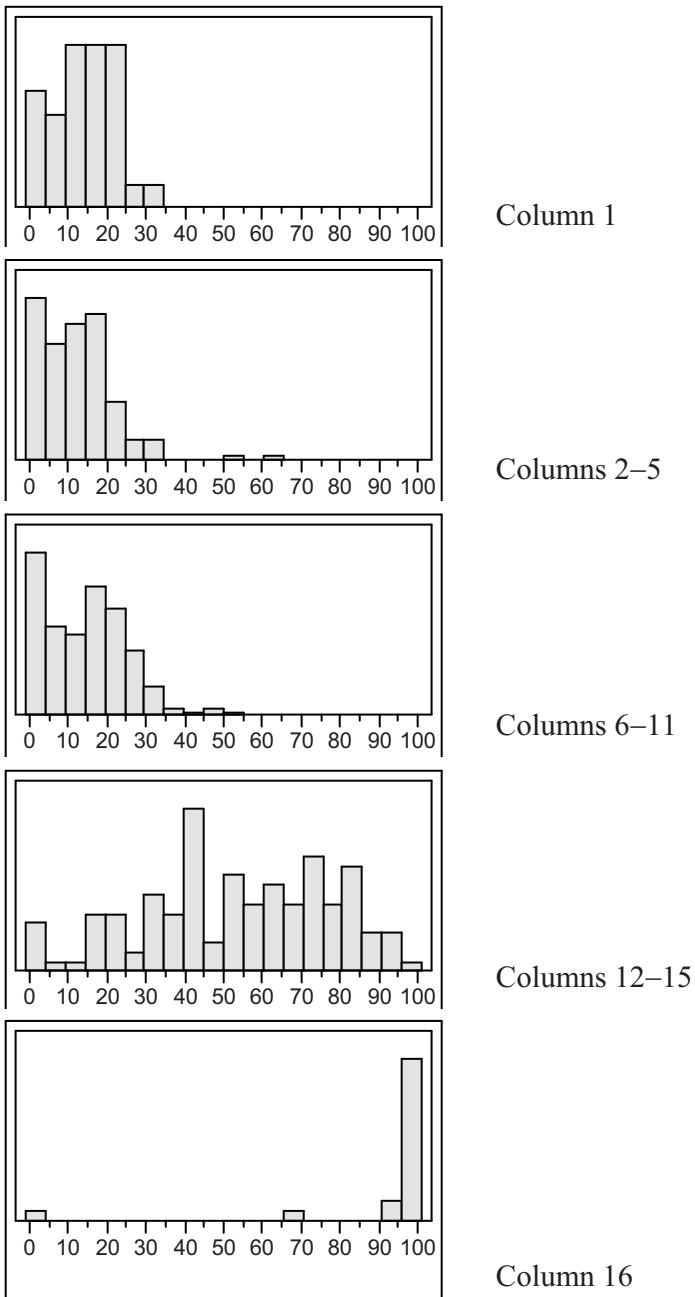


Fig. 4.6. Histograms for P for subsets of the data in Table 4.2

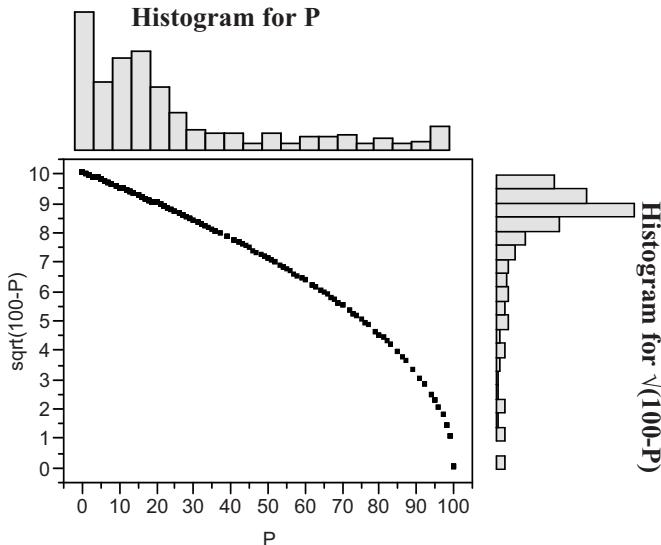


Fig. 4.7. Transformation and histograms for response in Wang et al. (1993) data

This indicates that for large m , Lenth's proposal of using $m/3$ df is accurate for extreme p -values.

One clearly needs a saturated model involving the four C-terminal factors, x_6-x_9 , since 13 of 15 terms involving only these 4 factors are significant, including the four-factor interaction. This model would account for the differences between columns for Table 4.2. In addition, x_4 has an effect that varies from column to column, as reflected in the three terms involving x_4 and the C-terminal factors. If one estimates the effect for x_4 using each of the 16 columns of Table 4.2 individually, the first 11 columns produced negligible estimates for b_4 ranging from 0.010 to -0.021 , whereas the last five columns produced b_4 values ranging from -0.070 to -0.097 . There are three additional terms listed in Table 4.3: x_1*x_2 , plus two interactions involving x_4*x_5 . These suggest effects related to pairs of adjacent N-terminal amino acids.

To construct a hierarchical model involving the 20 effects in Table 4.3 would require 75 terms in all; this is considered too complicated to be useful. Instead, if we use an EER $\alpha = .05$, only the largest 15 estimates in Table 4.3 are significant, and these involve only x_4 and x_6-x_9 . A full factorial model for T in these five significant factors has $R^2 = 82\%$. A plot of observed T versus \hat{T} reveals three prominent outliers with Studentized residuals of 7.62, 5.13, and -4.93 (see Figure 4.8). The last column of Table 4.2 contains $P = 0$ and $P = 64\%$; these outliers with larger-than-expected T appear in the upper-left corner in Figure 4.8. The third outlier is the $P = 100$ value in column 15. This appears as $T = 0$, $\hat{T} = 0.49$ in the figure. Without the variance-stabilizing

transformation, this point is not recognized as an outlier. Omitting these three values and refitting the full factorial model in five factors, the R^2 value increases to nearly 86%.

Table 4.3. Largest estimates from saturated model for T ; 20 exceed the EER critical value from Appendix C for $\alpha = .20$

Source	Estimate	PSE	Lenth t	Empirical p-Value	$t_{511/3}$ p-Value
Intercept	0.8181	0.0041			
x_9	-0.0974	0.0041	-23.65	.0000000	1.1e-55
x_8	-0.0881	0.0041	-21.38	.0000000	4.3e-50
x_7	-0.0779	0.0041	-18.91	.0000000	1.1e-43
x_6	-0.0636	0.0041	-15.45	.0000000	3.2e-34
$x_8 * x_9$	-0.0621	0.0041	-15.07	.0000000	3.7e-33
$x_7 * x_8$	-0.0533	0.0041	-12.93	.0000000	4.3e-27
$x_7 * x_9$	-0.0524	0.0041	-12.72	.0000000	1.7e-26
$x_6 * x_9$	-0.0428	0.0041	-10.39	.0000000	6.7e-20
$x_6 * x_8$	-0.0384	0.0041	-9.31	.0000000	6.2e-17
$x_6 * x_7$	-0.0350	0.0041	-8.49	.0000000	9.9e-15
x_4	-0.0257	0.0041	-6.23	.0000000	3.62e-9
$x_7 * x_8 * x_9$	-0.0237	0.0041	-5.75	.0000000	3.98e-8
$x_6 * x_8 * x_9$	-0.0189	0.0041	-4.60	.0000087	.0000083
$x_4 * x_6 * x_7 * x_8 * x_9$	0.0181	0.0041	4.38	.0000222	.0000204
$x_4 * x_8$	-0.0178	0.0041	-4.32	.0000284	.0000263
$x_4 * x_9$	-0.0163	0.0041	-3.95	.0001182	.0001132
$x_1 * x_2$	-0.0161	0.0041	-3.87	.0001574	.0001531
$x_4 * x_5 * x_6 * x_7 * x_8 * x_9$	0.0157	0.0041	3.81	.0001961	.0001910
$x_6 * x_7 * x_8 * x_9$	0.0153	0.0041	3.72	.0002747	.0002682
$x_1 * x_2 * x_4 * x_5$	0.0148	0.0041	3.59	.0004429	.0004347
$x_2 * x_3 * x_6 * x_7 * x_8 * x_9$	0.0145	0.0041	3.53	.0005467	.0005376
$x_3 * x_4 * x_6 * x_7 * x_8 * x_9$	0.0143	0.0041	3.46	.0006965	.0006830
$x_5 * x_6$	-0.0133	0.0041	-3.22	.0015715	.0015416
$x_1 * x_3 * x_4$	0.0132	0.0041	3.20	.0016492	.0016193
$x_4 * x_6$	-0.0126	0.0041	-3.07	.0025576	.0025198
$x_4 * x_7$	-0.0126	0.0041	-3.05	.0026866	.0026454

The three outliers noted above might be reexamined to see whether the recorded data is in fact correct. If we refit the full factorial model in 5 factors to the remaining 509 observations, there are an additional 10 Studentized residuals with magnitude of 3 or more. A normal plot reveals a distribution with much heavier tails than a normal (see Figure 4.9). The most likely explanation is that factors x_1-x_3 and x_5 have effects that our current model ignores.

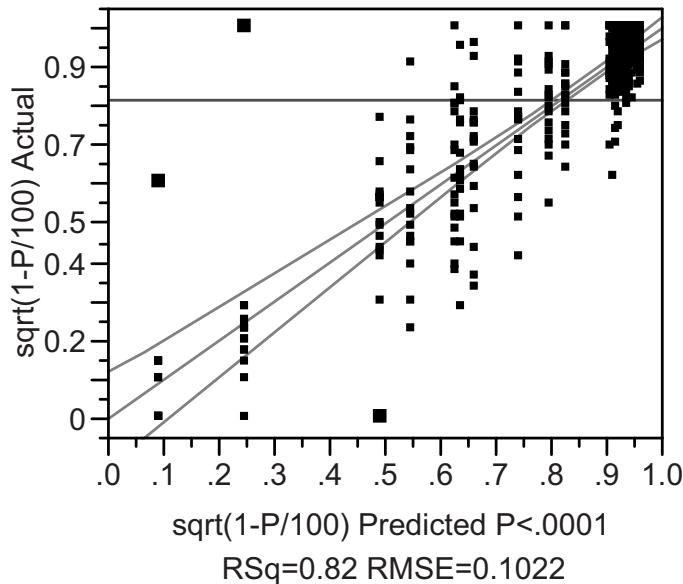


Fig. 4.8. Observed versus predicted $T = \sqrt{1 - P/100}$ for full factorial model in x_4 , x_6 , x_7 , x_8 , and x_9

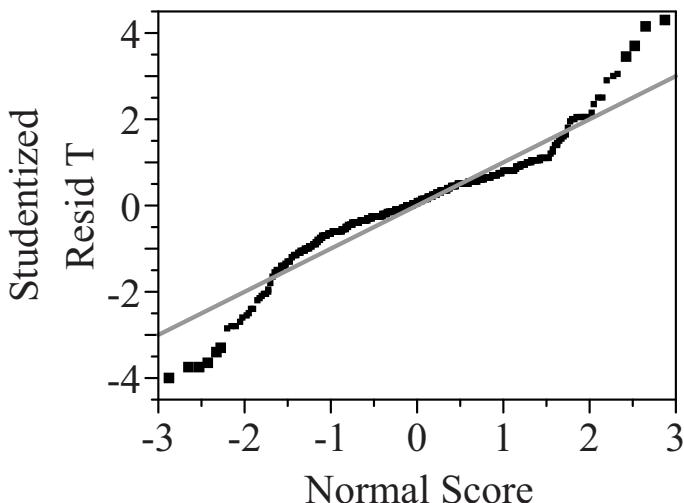


Fig. 4.9. Studentized residuals for full factorial model in x_4 , x_6 , x_7 , x_8 , and x_9 after three extreme outliers are omitted

How do we explore the need for additional terms with the three outliers removed? Given the lack of orthogonality and our interest in parsimonious models, we choose stepwise regression, adding terms without any constraints for the model to be hierarchical. If we specify $\alpha = .001$ as the probability to enter, 22 terms are added. The first four terms are main effects for x_6 – x_9 , followed by all their two-factor interactions. Eight of the final 12 terms come from the five-factor model fit earlier. The other four terms are $x_1 * x_2$ and $x_1 * x_2 * x_4 * x_5$, plus two six-factor interactions. Using $\alpha = .01$ results in a model with 37 terms; the last 15 are all interactions, including several with seven or eight factors. No simple model appears to account for all the systematic variation. Although a model involving just five of the nine factors explains 86% of the variation in T , we recognize that some additional systematic effects remain.

4.2.2 Classification tree analysis

When a single simple model cannot account for all of the variation, it is sometimes advisable to split the data into segments and fit simpler models to each segment. Young and Hawkins (1995) analyzed the response P using a classification tree. Their initial fitted tree had 16 terminal nodes (or leaves) and utilized 6 factors (x_4 – x_9). This fit identified five suspected outliers. Fitting a tree with these outliers removed produced a tree with 17 leaves utilizing 8 of the 9 factors, and having $R^2 = 81\%$. We now fit a classification tree for the response T , and find that the variance-stabilizing transformation assists in determining a simpler model.

Beginning with the full 2^9 and using JMP 7.0's Partition Modeling to fit the response T , the "Actual versus Predicted" plot reveals the three extreme outliers as soon as the model contains five or more leaves. Excluding these three observations, we obtain a tree with 20 leaves and $R^2 = 87\%$, which requires only 6 of the 9 factors. Figure 4.10 displays the tree, and Table 4.4 provides a summary of each leaf, sorted by the predicted T . The tree is exceedingly simply to interpret, largely because the splits at each level of the hierarchy involve the same factor: first x_9 , then x_8 , x_7 , x_6 , and x_4 , and, finally, x_5 . Note that this structure was not imposed, nor is it typical of classification tree models. Further, in Figure 4.10 we see that the branch to the right, corresponding to higher values of T , is the D-amino acid in every case except for leaves 5 and 6. According to this model, the one exception where replacement of the native L-amino acid decreases T (i.e., increases P) is when both x_4 and x_6 have the D-amino acid.

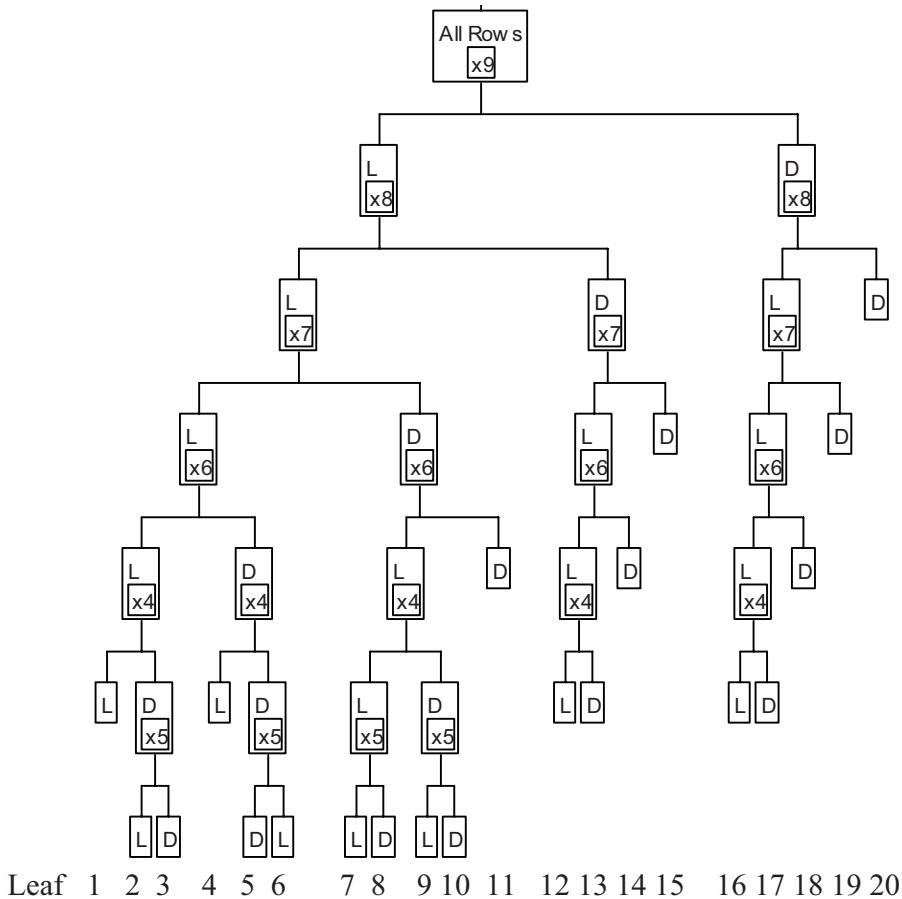


Fig. 4.10. Classification tree for T , excluding three outliers, with $R^2 = 87\%$

Table 4.4. Leaf summary for model in Figure 4.10, sorted by \hat{T}

Leaf No.	Leaf Label	No. of C-Terminal			Std. Dev. for T	
		D's	Count	\hat{T}	\hat{P}	
1	$x_9=L, x_8=L, x_7=L, x_6=L, x_4=L$	0	15	0.058	99.7	0.065
2	$x_9=L, x_8=L, x_7=L, x_6=L, x_4=D, x_5=L$	0	8	0.154	97.6	0.075
3	$x_9=L, x_8=L, x_7=L, x_6=L, x_4=D, x_5=D$	0	7	0.243	94.1	0.031
7	$x_9=L, x_8=L, x_7=D, x_6=L, x_4=L, x_5=L$	1	8	0.445	80.2	0.145
4	$x_9=L, x_8=L, x_7=L, x_6=D, x_4=L$	1	15	0.520	72.9	0.111
5	$x_9=L, x_8=L, x_7=L, x_6=D, x_4=D, x_5=D$	1	8	0.550	69.8	0.132
12	$x_9=L, x_8=D, x_7=L, x_6=L, x_4=L$	1	16	0.637	59.4	0.161
8	$x_9=L, x_8=L, x_7=D, x_6=L, x_4=L, x_5=D$	1	8	0.649	57.8	0.149
16	$x_9=D, x_8=L, x_7=L, x_6=L, x_4=L$	1	16	0.662	56.2	0.178
9	$x_9=L, x_8=L, x_7=D, x_6=L, x_4=D, x_5=L$	1	8	0.669	55.2	0.185
6	$x_9=L, x_8=L, x_7=L, x_6=D, x_4=D, x_5=L$	1	8	0.705	50.4	0.196
13	$x_9=L, x_8=D, x_7=L, x_6=L, x_4=D$	1	16	0.797	36.4	0.114
10	$x_9=L, x_8=L, x_7=D, x_6=L, x_4=D, x_5=D$	1	8	0.815	33.6	0.085
17	$x_9=D, x_8=L, x_7=L, x_6=L, x_4=D$	1	16	0.827	31.6	0.099
11	$x_9=L, x_8=L, x_7=D, x_6=D$	2	32	0.911	17.0	0.064
14	$x_9=L, x_8=D, x_7=L, x_6=D$	2	32	0.918	15.8	0.059
15	$x_9=L, x_8=D, x_7=D$	2–3	64	0.924	14.6	0.068
18	$x_9=D, x_8=L, x_7=L, x_6=D$	2	32	0.924	14.5	0.058
20	$x_9=D, x_8=D$	2–4	128	0.942	11.4	0.045
19	$x_9=D, x_8=L, x_7=D$	2–3	64	0.945	10.8	0.055

Further insight is gained by scoring the 20 leaves in terms of the number of D-amino acid replacements in the C-terminal factors x_6 – x_9 . Note that two or more replacements (the last six leaves in Table 4.4) have $\hat{T} > 0.9$ and very little unexplained variability. These six leaves account for 352 of the 512 combinations. At the other extreme, the first three leaves have no replacements in the C-terminal factors, and $\hat{T} < 0.25$, again with little unexplained variation. The other 11 leaves correspond to leaves with a single D-amino acid replacement among x_6 – x_9 . Although these leaves account for only one-fourth of the data, they correspond to the most variation (both explained and unexplained); predicted P for these leaves ranges from 32% to 80%.

More properly accounting for the error variation in the response has led to a more parsimonious model (in terms of number of factors) than the classification tree obtained by Young and Hawkins (1995). Our model for T explains the majority of the variation in T , at the same time highlighting the need for more insight regarding the factor effects when a single C-terminal L-amino acid is replaced with D. For a final look at the Wang et al. data, in the next subsection we examine models for the 128 treatment combinations with 1 C-terminal replacement.

Other models might be useful, including regression trees that fit an additive regression model at each leaf (see Loh 2006). We also considered the method of Filliben and Li (1997), who suggested an alternative empirical means for

fitting different regression models for segments of the data, based on analysis of the two-factor interactions. However, their approach did not provide much insight for Wang's data.

4.2.3 Further analysis of Wang et al. (1993) data with a single C-terminal amino acid replacement

Figure 4.11 shows a plot of T versus the number of D-amino acid replacements among the four factors x_6-x_9 . Five outliers are highlighted. The classification tree predicts T well for cases with no replacements or with two or more replacements, but leaves much variation unexplained when there is a single replacement with a D-amino acid for x_6-x_9 . Considering only the 127 observations with 1 replacement (excluding the outlier), the 11 leaves of the classification tree explained only 42% of the variation. Here we investigate other models for just this portion of the data (columns 12–15 of Table 4.2).

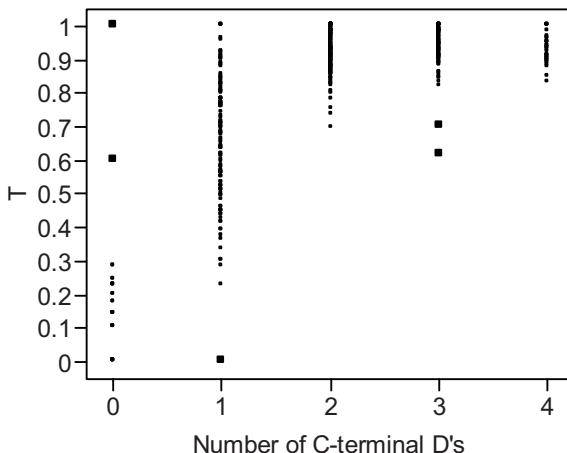


Fig. 4.11. T versus the number of D-amino acid replacements for x_6-x_9

We construct a variable “C-terminal D” with the four levels {6, 7, 8, 9} to indicate for columns 12–15 which factor has the D-amino acid. Fitting a full factorial model in this factor and x_1-x_5 reveals that two-factor interactions for factors in adjacent positions on the peptide are statistically significant, whereas other two-factor interactions are not. Since the significant two-factor interaction estimates are all negative, this implies that, apart from the additive main effects, having adjacent amino acids match is beneficial; that is, it decreases $E(T)$ and so increases P . Table 4.5 provides the ANOVA for a reduced model with $R^2 = 60\%$, and Table 4.6 lists the estimated coefficients. Note that, surprisingly, the estimate for x_1 is positive, indicating that a D-amino acid replacement here is actually beneficial. Furthermore, since the

main effects for x_2 and x_3 are not significant, D-amino acids in the first three positions can be beneficial, by virtue of $b_{1,2} = -0.034$ and $b_{2,3} = -0.52$. The RMSE for this model is 0.12, which is an improvement over the within-leaf standard deviations in Table 4.4 for this portion of the data.

Table 4.5. ANOVA for a reduced model for T , with data from Table 4.2, columns 12–15

Source	df	SS	MS	F-Ratio	p-Value
Model	13	2.429	0.1868	12.818	< .0001
Error	113	1.648	0.0146		
Total	126	4.076			

Table 4.6. Estimated coefficients for main effects model, plus two-factor interactions of adjacent factors

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	0.671	0.011	62.56	.000
x_1	0.026	0.011	2.41	.018
x_2	-0.006	0.011	-0.57	.570
x_3	-0.002	0.011	-0.15	.882
x_4	-0.078	0.011	-7.26	.000
x_5	-0.017	0.012	-1.37	.172
C-terminal D[6]	-0.095	0.019	-5.06	.000
C-terminal D[7]	-0.026	0.019	-1.41	.162
C-terminal D[8]	0.047	0.019	2.53	.013
C-terminal D[9]	0.074	0.019	4.00	.000
x_1*x_2	-0.034	0.011	-3.22	.002
x_2*x_3	-0.052	0.011	-4.89	.000
x_3*x_4	-0.036	0.011	-3.32	.001
x_4*x_5	-0.023	0.011	-2.10	.038
x_5*x_6	-0.036	0.012	-2.90	.004

4.2.4 Summary of our analysis

We have analyzed Wang's 2^9 data in four separate steps. First, we plotted the P data, observed differences in variability, and chose T rather than P as the response, based on our understanding the nature of the measurement. Second, we fit a full factorial model for T , using an EER of 0.05 to keep the model from being too complex to interpret. We found that a full factorial model in the factors x_4 and x_6-x_9 explained 86% of the variation in T , once three extreme outliers were removed. This initial analysis proved the primary

importance of the C-terminal factors x_6-x_9 . The prevalence of outliers for our reduced model suggested that this five-factor model, although useful, still omitted some systematic effects. Thus, we tried another modeling approach, fitting a classification tree as Young and Hawkins (1995) had done. By using T rather than P as the response, we obtained a simpler model than Young and Hawkins and one that was easily interpreted, especially after sorting the leaves by the predicted response. The classification tree revealed that the number of D-amino acid replacements for x_6-x_9 prominently affects the response. The tree predicted the response well, except for the data in columns 12–15 of Table 4.2, where there is a single position with a D-amino acid among the last four. The final analysis focused on these four columns and found that although not all main effects were important, consistency of adjacent positions did have a beneficial effect on the response. In this final step, we focused on obtaining a model that was interpretable, excluding some higher-order interactions that were statistically significant.

In Chapter 8, we reanalyze fractions of the 2^9 data to illustrate the use of various (less common) fractional factorial designs. By doing so, we will learn that many of the conclusions obtained from the full 2^9 would have been ascertained from a small subset of the 512 treatment combinations.

4.3 Example 4.3: 2^5 with Centerpoint Runs for Ceramic Strength

Section 2.3 explained how centerpoint replicates provide an estimate for the error variance and permit testing for nonlinearity in the factor effects. Bouler et al. (1996)'s ceramic compressive strength data in Table 2.3 were used to illustrate the computations. Here, we review the authors' findings and perform a thorough reanalysis.

The experiment involved five steps and five factors:

1. HA and β -TCP powders were blended according to a specified ratio (x_1) to form the biphasic calcium phosphate (BCP).
2. A prescribed mass (x_2) of naphthalene particles of a given diameter (x_3) are mixed with the BCP powder from step 1.
3. Two milligrams of the mixture are isostatically compacted at a specified pressure (x_4).
4. The compacted cylinder is heated to 500°C to eliminate the naphthalene and then sintered at a higher temperature (x_5).
5. The sintered ceramic is then measured for compressive strength and checked for purity.

See Table 2.2 for the factor levels.

Bouler et al. concluded that four of the five factors impact compressive strength; only Isostatic compaction pressure (x_4) was found to have no effect. In fact, in confirmation runs, they found that a wide range of pressures

sufficiently compacted the mixture without fracturing the particles. Although some important interactions were found, main effects dominate; compressive strength increases with Sintering temperature and Diameter of macropores but decreases with higher proportions of naphthalene and HA.

4.3.1 Analysis of Bouler's data as a completely randomized design

We added the x_4 main effect to the reduced model selected by Bouler et al. to obtain the residual plot in our Figure 2.5b. Although this model has an $R^2 = 84\%$, its lack-of-fit is highly significant ($F = 25.75$, $p = .0003$). If one fits a saturated model instead, the t statistic for the five-factor interaction $b_{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 1.77$ is $t = 7.35$ (p -value = .003), and all but 6 of the 31 factorial effects are statistically significant at $\alpha = .05$. Contrary to Bouler et al.'s article, the test for curvature is also significant (p -value = .020). No simple model satisfactorily fits these data. Nor does using a transformation of the response remedy this problem, although taking $(\text{Strength})^{1/2}$ would improve the residual plot for a reduced model (see Figure 2.5c).

Perhaps the lack-of-fit found in this initial analysis is indicative of the complicated combination of effects. More likely, it indicates that the pure error mean square, calculated from the centerpoint replicates, seriously underestimates run-to-run experimental error. The seven centerpoint data values may in fact correspond to observations produced from a single mixture of ingredients and concurrently sintered in an oven. If so, then the pure error mean square excludes pertinent sources of variation and is not valid for testing the significance of model effects. Bouler et al. were silent regarding the preparation and testing sequence followed for the 39 cylinders. It is indicated that each cycle of the sintering process, including cool down, took approximately 24 hours. Given that each test involved such a small quantity of material, it is only reasonable that some sets of runs were sintered in the same cycle and/or taken from the same mixture batch. For this reason, we will proceed as if this were an unreplicated experiment. The seven "replicates" at the center will be averaged and treated as a single observation.

Histograms for Strength and $(\text{Strength})^{1/2}$ are shown in Figure 4.12. For (untransformed) Strength, 30 of the 33 observations fall below the midrange. Taking the square root produces a more even distribution of response values, and is our choice for this analysis.

Fitting a full factorial model to the 2^5 $(\text{Strength})^{1/2}$ values produces Lenth's PSE = 0.223; only four effects have Lenth t statistics exceeding 2.064, the IER critical value for $\alpha = .05$. Adding one main effect to make the model hierarchical, the fitted model is

$$\widehat{(\text{Strength})^{1/2}} = 2.47 - 0.45x_1 - 1.41x_2 + 0.68x_3 + 0.80x_5 - 0.54x_1x_5. \quad (4.3)$$

This simple model explains 74% of the variation. Figure 4.13 provides a plot of actual versus predicted values for $(\text{Strength})^{1/2}$.

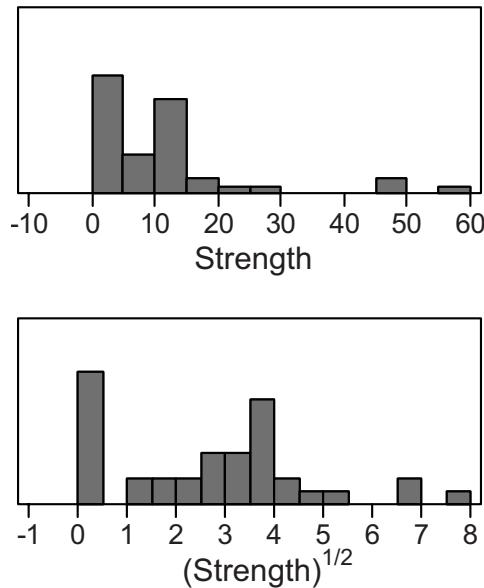


Fig. 4.12. Histograms for Strength and $(\text{Strength})^{1/2}$

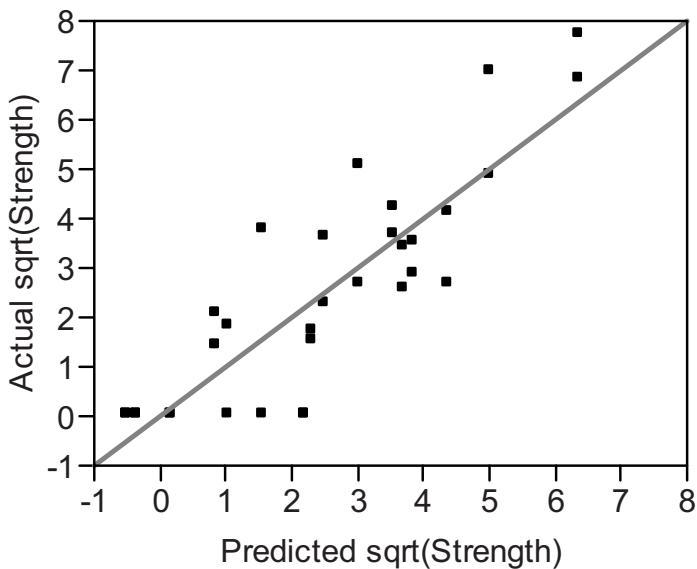


Fig. 4.13. Predicted versus observed $(\text{Strength})^{1/2}$ for reduced model (4.3)

This model fits well, with the exception of not accounting for the observations with zero strength. Ten of the 32 factorial treatment combinations produced no measurable compressive strength. In fact, the ceramic disintegrated at all eight runs with $x_2 = 1$, $x_3 = -1$. (Bouler et al. explained that this combination produces unsuitably narrow bridges between the macropores.) This outcome is not accounted for by our current model. The next subsection will consider models fitted to data omitting this combination.

4.3.2 Omitting one-fourth of the 2^5

A cube plot of the 2^5 strength values reveals why no simple model can account for these data (see Figure 4.14). Whenever a high quantity of small naphthalene particles are included, the ceramic has no measurable strength, regardless of the levels of the other factors. Whatever model applies elsewhere in the experimental region does not apply when $(x_2 = 1, x_3 = -1)$. We now proceed by analyzing portions of the data that exclude the eight observations for this quarter of the factorial, to obtain a suitable model for the restricted region where the process does perform more favorably.

Considering only the data with $x_2 = -1$ (i.e., the two left cubes in Figure 4.14), we have an unreplicated 2^4 . Fitting a full factorial model for $(\text{Strength})^{1/2}$, one main effect and one interaction have statistically significant Lenth t statistics. A hierarchical model is

$$\widehat{(\text{Strength})^{1/2}} = 3.9 - 0.5x_1 + 1.2x_5 - 0.9x_1x_5. \quad (4.4)$$

When Naphthalene % (x_2) is low, high Sintering temperature (x_5) is best, especially at low HA (x_1), and neither x_3 or x_4 appear to have any affect.

Now consider only the data with $x_3 = 1$ (i.e., the two lower cubes in Figure 4.12). Fitting a full factorial model for $(\text{Strength})^{1/2}$, only the x_2 main effect is statistically significant, although the three terms in (4.4) are the next largest estimates. Including these produces the fitted model

$$\widehat{(\text{Strength})^{1/2}} = 3.1 - 0.8x_1 - 1.0x_2 + 1.0x_5 - 0.7x_1x_5. \quad (4.5)$$

Higher volume of Naphthalene (x_2) definitely lowers Strength, even when the particles (x_3) are large. When naphthalene particles are large, high Sintering temperature (x_5) produces higher Strength, especially at low HA (x_1), and, again, x_4 has no significant effect.

The fitted models (4.4) and (4.5) are quite consistent, which implies that a simple model applies for three-quarters of the 2^5 . We now exclude only the eight observations at the $(x_2 = 1, x_3 = -1)$ combination and fit a saturated model. We cannot estimate the $x_2 * x_3$ interaction or any of the seven higher-order interactions involving both of these factors. However, if we omit these seven interactions, the remaining terms form a hierarchical model that can be estimated. Fitting a model for $(\text{Strength})^{1/2}$ to just the remaining 24 factorial

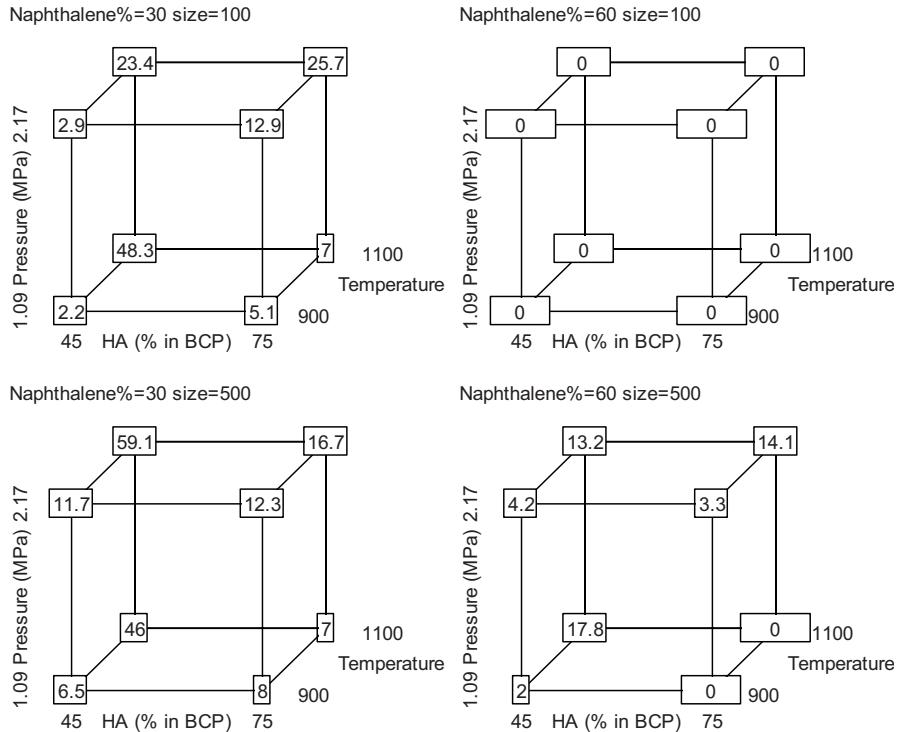


Fig. 4.14. Cube plot for Strength

combinations, the estimated coefficients all have the same standard error. (This is true whenever one omits a quarter fraction and fits a saturated model; see Section 8.3 for related designs.) Table 4.7 contains these estimates, sorted from largest to smallest.

We use Lenth's method to compute a standard error for the estimates in Table 4.7, even though they are correlated (one-eighth of the correlations are $\pm .5$, and the rest are zero). For such cases, Lenth's method is slightly conservative (Edwards and Mee 2008). From Appendix C, the $c_{.05}^{\text{IER}}$ critical value for 23 contrasts is 2.097. Thus, main effects for Naphthalene % and Sintering temperature are declared active. Although the next four terms are not statistically significant based on Lenth's PSE, they do form a simple model involving interactions with the percentage of HA (x_1). At the risk of overfitting the model, we consider the following equation for predicting Strength, obtained by fitting a reduced model with six effects:

$$(2.99 - 0.61x_1 - 0.88x_2 + 0.47x_4 + 1.07x_5 + 0.48x_1x_4 - 0.72x_1x_5)^2. \quad (4.6)$$

Table 4.7. Estimated coefficients for $(\text{Strength})^{1/2}$ for model fit to 24 factorial treatment combinations

Term	Estimate	PSE	Lenth t
Intercept	2.839	0.466	6.09
x_5	1.050	0.466	2.25
x_2	-1.039	0.466	-2.23
$x_1 * x_4$	0.698	0.466	1.50
$x_1 * x_5$	-0.574	0.466	-1.23
x_4	0.468	0.466	1.00
x_1	-0.447	0.466	-0.96
$x_1 * x_4 * x_5$	0.411	0.466	0.88
$x_1 * x_3$	-0.373	0.466	-0.80
$x_1 * x_2 * x_5$	0.353	0.466	0.76
$x_1 * x_3 * x_4$	-0.332	0.466	-0.71
$x_1 * x_2 * x_4$	0.324	0.466	0.69
x_3	0.311	0.466	0.67
$x_1 * x_3 * x_4 * x_5$	-0.168	0.466	-0.36
$x_2 * x_5$	-0.161	0.466	-0.34
$x_1 * x_2 * x_4 * x_5$	0.152	0.466	0.33
$x_3 * x_4 * x_5$	0.127	0.466	0.27
$x_3 * x_4$	0.127	0.466	0.27
$x_2 * x_4$	0.108	0.466	0.23
$x_1 * x_2$	0.102	0.466	0.22
$x_3 * x_5$	-0.098	0.466	-0.21
$x_1 * x_3 * x_5$	-0.085	0.466	-0.18
$x_4 * x_5$	-0.033	0.466	-0.07
$x_2 * x_4 * x_5$	-0.005	0.466	-0.01

Model (4.6) indicates that the smaller the HA% (x_1), the greater the benefit from high Sintering temperature (x_5), which is consistent with both (4.4) and (4.5). At low HA%, there is no benefit to higher Compaction pressure (x_4); however, when the HA% is high, the higher Compaction pressure seems to increase Strength. The reduced model with four main effects and two interactions explains 85% of the variation in $(\text{Strength})^{1/2}$, whereas the model with only two main effects explains just 49% of the variation for these 24 observations. Comparing Table 4.7 with the reduced model (4.6), note that the estimates changed when we omitted insignificant terms, due to correlations among the estimates in the saturated model. The standard error for b_0 and $b_2 = -0.88$ is 0.193, and the other coefficients in (4.6) have a standard error of 0.182; these likely overstate the precision (and statistical significance) of the coefficients, because they are based on the reduced model's MSE.

Figure 4.15 shows the residual plot for the reduced model (4.6). We prefer the fitted model (4.6) fitted to a portion of the data, rather than the model (4.3) obtained using all the factorial data. Since the model (4.3) does not fit

well when $(x_2 = 1, x_3 = -1)$, where Strength = 0, it is preferable to ignore this region when fitting a model. For the reduced data set, the residuals support the assumption of common error variance.

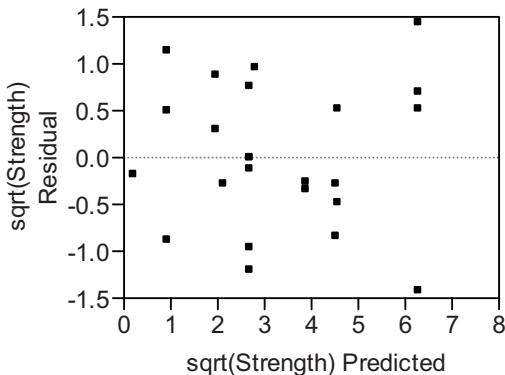


Fig. 4.15. Residual versus predicted plot for $(\text{Strength})^{1/2}$ based on (4.6)

4.3.3 Factor relationship diagram and final discussion of Bouler et al. design

We concluded earlier that the seven centerpoint replicates were likely produced together. If this is the case, then other sets of runs might also have been blended or sintered as a group, producing correlations among the errors. If such an experiment is much more convenient, then design planning needs to take this into account. For instance, since the sintering step requires a full 24 hours, it may be expedient to sinter four to eight observations together at a time. If 8 at a time, then the 32 factorial combinations would be partitioned into 4 blocks of size 8, blocking on x_5 and the five-factor interaction; the centerpoint runs would then constitute a fifth block. Within each block, the batches of BCP need to be mixed individually for each run, even the common centerpoint runs. For such a design, the centerpoint runs would be used to estimate the split-unit variance, whereas testing nonlinearity would involve a whole-unit (i.e., interblock) contrast. We now consider a diagram that helps document such subtle design differences.

Bergerud (1996) introduced *factor relationship diagrams* that effectively display the treatment combinations for factorial designs whether they are run with either blocking or the hierarchical unit structure of split-unit designs. There is no need to display the unit structure for completely randomized designs. However, for designs with randomization restrictions, Bergerud's factor relationship diagrams (FRDs) are particularly useful.

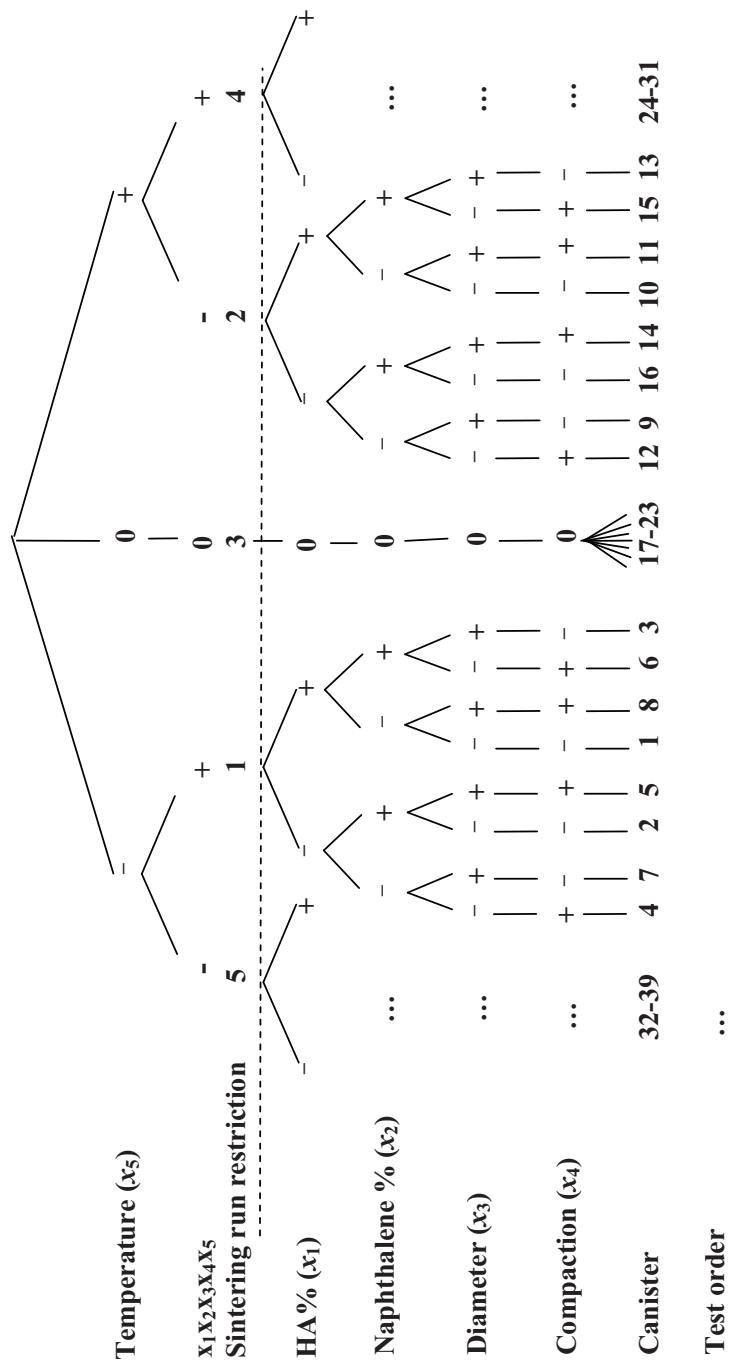


Figure 4.16. Factor relationship diagram for Bouler et al. experiment run as a split-unit design with five whole units

Figure 4.16 provides a FRD of the treatment and unit structure for Bouler et al.'s (1996) 39-run experiment, assuming this 2^5 factorial, with 7 center-point replicates, was performed as a split-unit design with 1 whole-unit factor (Temperature) and 5 whole units. The unit structure for split-unit designs is hierarchical. Here, furnace runs are the largest unit grouping, with canisters nested within runs. This unit structure is displayed by identifying the randomized order of the five furnace runs and drawing a dashed line under this to emphasize that canisters are grouped within runs. Sanders and Coleman (2003) proposed using such lines to indicate "restrictions on variation."

Sintering temperature is the only factor that varies from furnace run to furnace run, so its levels are placed on top. Since two furnace runs are made at low temperature and two are made at high temperature, another branch is displayed for the interaction used to create this blocking. Note that the diagram could have been drawn with either $x_1x_2x_3x_4$ or the five-factor interaction, since both are confounded with blocks. Factors x_1-x_4 distinguish the different compositions. Because no additional restriction lines are drawn, it is assumed that each of the 39 compositions is formed individually.

If Bouler et al.'s (1996) experiment were performed in the manner displayed in Figure 4.16, the pure error mean square could be used to test all of the effects except for those confounded with blocks: x_5 , $x_1 * x_2 * x_3 * x_4$, the five-factor interaction, and the pure quadratic term. If testing for the significance of x_5 's main effect is essential, then one should perform the experiment with more blocks (i.e., more furnace runs).

If the material for the seven centerpoint canisters were blended as a batch and material for the other 32 canisters were blended individually prior to compaction and sintering, then another restriction line would be drawn just below the level for x_3 , emphasizing that the composition step was performed just 33 times. The seven branches at the center would fall below this line, emphasizing that these seven canisters were formed from a single blend. Whether the results would differ for these two manners of preparing the centerpoint canisters depends on the extent of variation associated with raw material heterogeneity or mixing. FRDs are particularly useful for clarifying and communicating such changes to an experimental plan.

Part II

Fractional Factorial Designs

Fractional Factorial Designs: The Basics

This chapter presents the essential ideas of regular fractional factorial designs. Its sections are as follows:

Section 5.1. Initial Fractional Factorial Example

Section 5.2. Introduction to Regular Fractional Factorial Designs

Section 5.3. Basic Analysis of Regular Fractional Factorial Designs

Following this introductory chapter on fractional factorials are six more chapters with additional details and examples.

Regarding notation for factors, sometimes we identify factors using subscripts (e.g., x_1 for the first factor). In other cases, it is more convenient to avoid the use of subscripts, in which case we will label coded factors using bold uppercase letters or numerals (e.g., **A** or **1**).

5.1 Initial Fractional Factorial Example

Fractional factorial designs permit investigation of the effects of many factors in fewer runs than a full factorial design. To illustrate a typical fractional factorial experiment, consider the following 16-run experiment for 5 factors. Hu and Bai (2001) investigated how to control the phosphorus content of nickel–phosphorus deposits electroplated from a modified nickel bath. Their initial experiment involved five factors, each at two levels (refer to Table 5.1).

Rather than completing a full 2^5 factorial, they performed just 16 treatment combinations, as shown in Table 5.2, which are half of the possible 32 combinations. Hu and Bai stated that the nickel bath was freshly prepared for each experimental run. Although many other procedural details are given, they fail to mention the run order used. (The row order shown is a standard order for listing treatment combinations, not the run order that should be followed for experimentation. Randomization of the order typically reduces

the impact of any undesirable correlation or trend in the random errors.) A single percent phosphorus (%P) measurement is reported for each run.

Table 5.1. Factors and levels for electroplating experiment

Factors	Levels	
	-1	1
A Temperature (°C)	20	50
B Current density (A/m ²)	500	2500
C pH	1	4
D NaH ₂ PO ₂ concentration (M)	0.5	1
E Stirring rate (rev/min)	200	400

Table 5.2. Treatment combinations (t.c.) and percent phosphorus (%P) for Hu and Bai's electroplating experiment

t.c.	A	B	C	D	E	%P
1	-1	-1	-1	-1	1	0.51
2	1	-1	-1	-1	-1	1.54
3	-1	1	-1	-1	-1	2.38
4	1	1	-1	-1	1	12.20
5	-1	-1	1	-1	-1	5.93
6	1	-1	1	-1	1	5.83
7	-1	1	1	-1	1	2.90
8	1	1	1	-1	-1	4.73
9	-1	-1	-1	1	-1	0.49
10	1	-1	-1	1	1	1.02
11	-1	1	-1	1	1	10.59
12	1	1	-1	1	-1	12.00
13	-1	-1	1	1	1	6.50
14	1	-1	1	1	-1	4.87
15	-1	1	1	1	-1	1.86
16	1	1	1	1	1	4.49

Quick inspection of the data reveals the wide range for %P achieved in this experiment. A histogram of the 16 %P values is a useful first step in the analysis (see Figure 5.1). Note that three treatment combinations (4, 11, and 12) produced substantially higher %P than the rest.

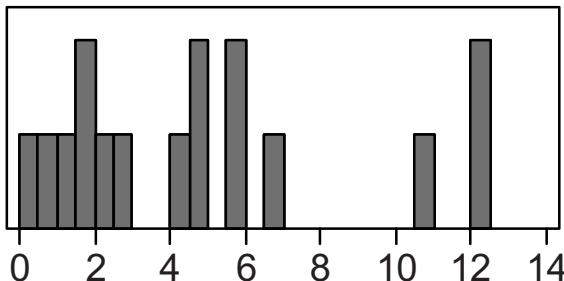


Fig. 5.1. Histogram for %P from electroplating experiment

We continue our analysis by fitting a saturated model containing all 5 main effects and 10 two-factor interactions. The 15 estimates are displayed in a Pareto plot (Figure 5.2) and a half-normal plot (Figure 5.3). Lenth's (1989) pseudo-standard-error for these estimates is calculated to be

$$\text{PSE} = 1.5(0.590 + 0.571)/2 = 0.871.$$

Based on the assumption of effect sparsity implicit in Lenth's method, only one effect stands out as significant: the interaction of Current density and pH, with Lenth $t = -2.672/0.871 = -3.07$. (The Lenth t statistic $1.529/0.871 = 1.76$ for Current density has p -value = .09.)

Term	Estimate
B*C	-2.67250
B	1.52875
A*B	0.99125
A	0.97000
E	0.64000
A*C	-0.62875
A*D	-0.60250
A*E	-0.59000
C*D	-0.57125
B*E	0.51125
B*D	0.47875
D	0.36250
C*E	-0.34875
C	-0.22625
D*E	-0.21750

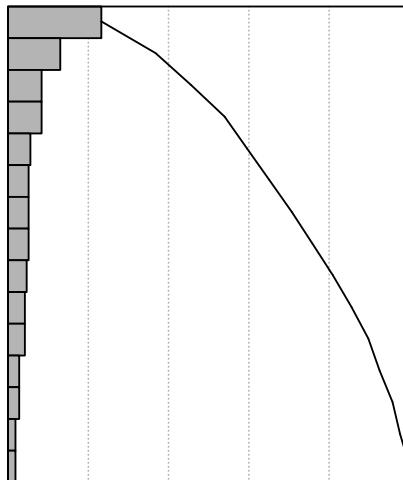


Fig. 5.2. Pareto plot for estimates of main effects and two-factor interactions

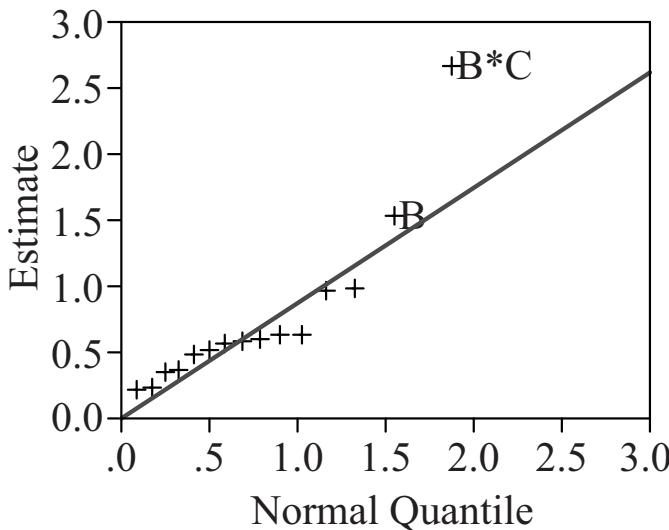


Fig. 5.3. Half-normal plot for estimates of main effects and two-factor interactions

The interaction graph in Figure 5.4 corresponds to the simple hierarchical model with main effects for Current density and pH and their interaction. The eight observations at low pH are denoted with circles, and the observations at high pH are denoted with pluses. Clearly, %P can be increased by holding pH low and increasing Current density. However, one observation at this combination fails to match the pattern—Run 3, where Temperature, Concentration, and Stirring rate are all at their low level. Thus, although this simple model with only two factors explains much of the variation in %P ($R^2 = 67.7\%$), either we have one aberrant response value or there exists systematic variation that this simple model fails to capture.

The researchers Hu and Bai (2001) followed this initial fractional factorial experiment with a second experiment in which the levels of the last three factors were held fixed (pH = 1, Concentration = 1 M, and Stirring rate = 400 rev/min) while searching along larger values for both Temperature and Current density. This follow-up experiment may be justified as follows:

- If we assume the effect of pH on %P is negative when Current density is high and that one would not use pH below 1, then pH = 1 is the optimal level for maximizing %P.
- The initial experiment does suggest that using higher temperatures will increase %P if Current density is high. The third and fourth largest estimates are the main effect for Temperature and its interaction with Current density. Although not individually significant, these estimates are synergistic; that is, the conditional effect for Temperature at high Current density is $0.970 + 0.991(1) = 1.961$.

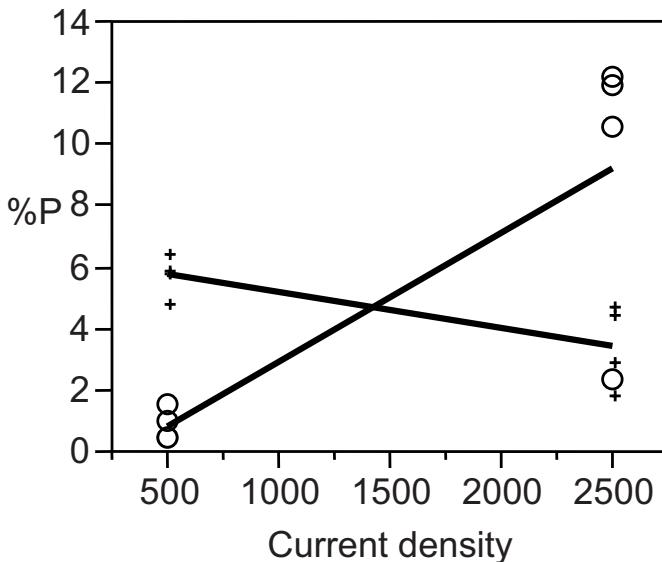


Fig. 5.4. Interaction plot for Current density and pH

- There is little evidence that Concentration and Stirring rate affect %P, except for the unexpectedly low response for Run 3. Setting both of these factors at their high levels is reasonable as one investigates the impact of raising both Temperature and Current density.

Their “steepest ascent” search in Temperature and Current density produced %P values from 19% to 22%, with the highest outcome at 60°C and 4000 A/m² (corresponding to **A** = 1.67 and **B** = 2.5). This follow-up experiment will be discussed again in Section 9.3.

The Hu and Bai initial 16-run experiment illustrates four points pertaining to the use of fractional factorial designs.

- **More factors economically.** Fractional factorial designs enable one to increase the number of factors without increasing the number of runs. In this case, the researchers opted for experimenting with five factors in half the number of runs required by a full factorial. If one only uses full factorials, a 16-run experiment would limit the choice of factors to 4 or fewer. Since the success of an experiment depends on inclusion of important factors, being able to examine five factors rather than four without requiring additional runs is quite useful.
- **Assuming model simplicity.** Fractional factorial designs permit estimation of the relevant effects, provided the true situation can be usefully described by a simple model. For the electroplating experiment, conducting only a half-fraction of the full factorial did not cause much confusion

as the results were analyzed. Although not perfect, a simple model did account for most of the variation, and several insights were achieved regarding how to increase the response.

- **Projection into a few important factors.** Two of the variables investigated in the electroplating experiment seemed less important. After accounting for the effects of Temperature, Current density, and pH, little variation in %P remained. This is often the case when experimenting with many factors. In such cases, one does not need a full factorial design in all of the factors.
- **Informed follow-up.** The design of follow-up experiments always takes into account what was learned from an initial fractional factorial design. For the electroplating investigation, collecting more data (e.g., completing a full 2^5) in the initial experimental region would not be as informative as augmenting the initial frugal experiment with runs in a new region where the predicted response is higher.

Exploring more factors economically is a key feature of fractional factorial designs. Although not suited for every situation, fractional factorial designs have wide applicability, especially for initial experiments in situations where follow-up experiments are feasible. We conclude this chapter with another example to illustrate further the analysis of fractional factorial designs. First, however, we present the key terminology and concepts needed to understand regular fractional factorial designs more fully.

5.2 Introduction to Regular Fractional Factorial Designs

In Chapter 3, we saw how the treatment combinations of a 2^k factorial could be partitioned into 2^b blocks by confounding b independent factorial effects, and their generalized interactions, with blocks. Consider, for instance, the 2^5 factorial in 2^2 ($= 4$) blocks of size $2^5/2^2 = 8$. Confounding with blocks the **ABD** and **ACE** effects, and their generalized interaction **ABD · ACE = BCDE**, the 32 treatment combinations of the full factorial are partitioned into blocks as follows:

A	B	C	D	E	ABD	ACE	BCDE	Block
-1	-1	-1	-1	-1	-1	-1	1	1
1	-1	-1	-1	-1	1	1	1	4
-1	1	-1	-1	-1	1	-1	-1	2
1	1	-1	-1	-1	-1	1	-1	3
-1	-1	1	-1	-1	-1	1	-1	3
1	-1	1	-1	-1	1	-1	-1	2
-1	1	1	-1	-1	1	1	1	4
1	1	1	-1	-1	-1	-1	1	1
-1	-1	-1	1	-1	1	-1	-1	2
1	-1	-1	1	-1	-1	1	-1	3
-1	1	-1	1	-1	-1	-1	1	1
1	1	-1	1	-1	1	1	1	4
-1	-1	1	1	-1	1	1	1	4
1	-1	1	1	-1	-1	-1	1	1
-1	1	1	1	-1	-1	1	-1	3
1	1	1	1	-1	1	-1	-1	2
-1	-1	-1	-1	1	-1	1	-1	3
1	-1	-1	-1	1	1	-1	-1	2
-1	1	-1	-1	1	1	1	1	4
1	1	-1	-1	1	-1	-1	1	1
-1	-1	1	-1	1	-1	-1	1	1
1	-1	1	-1	1	1	1	1	4
-1	1	1	-1	1	1	-1	-1	2
1	1	1	-1	1	-1	1	-1	3
-1	-1	-1	1	1	1	1	1	4
1	-1	-1	1	1	-1	-1	1	1
-1	1	-1	1	1	1	1	1	3
1	1	-1	1	1	1	-1	-1	2
-1	-1	1	1	1	1	-1	-1	2
1	-1	1	1	1	-1	1	-1	3
-1	1	1	1	1	-1	-1	1	1
1	1	1	1	1	1	1	1	4

The eight treatment combinations in the fourth block appear in Table 5.3. This subset of treatment combinations is defined by $\mathbf{ABD} = \mathbf{ACE} = \mathbf{BCDE} = +1$; that is, only for this quarter of the full 2^5 factorial are these three interaction columns identical to the intercept column. This construction of incomplete blocks, which was introduced in Chapter 3, has a close connection to regular fractional factorial designs. Now we introduce some terminology for fractional factorial designs, first for this particular example and then in full generality.

Table 5.3. One of four blocks of treatment combinations from a 2^5

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

5.2.1 Defining relation, defining contrast subgroup, resolution, and aliasing

The eight treatment combinations in Table 5.3 constitute one-fourth of a full 2^5 factorial design, which we denote by 2^{5-2} . The first term in the exponent indicates the number of factors; here we have a design for five factors. The second term in the exponent indicates what fraction of the full factorial; here “ -2 ” indicates a $2^{-2} = 1/4$ fraction. This particular quarter fraction is defined by

$$\mathbf{I} = \mathbf{ABD} = \mathbf{ACE} = \mathbf{BCDE}, \quad (5.1)$$

where \mathbf{I} denotes the identity column of +1’s. Expression (5.1) is known as the *defining relation* for this subset of the factorial, and the elements (\mathbf{I} , \mathbf{ABD} , \mathbf{ACE} , \mathbf{BCDE}) form its *defining contrast subgroup*. The other three subsets of the full 2^5 factorial created on the previous page have the following defining relations:

$$\begin{aligned} \text{Block 1: } & \mathbf{I} = -\mathbf{ABD} = -\mathbf{ACE} = \mathbf{BCDE}, \\ \text{Block 2: } & \mathbf{I} = \mathbf{ABD} = -\mathbf{ACE} = -\mathbf{BCDE}, \\ \text{Block 3: } & \mathbf{I} = -\mathbf{ABD} = \mathbf{ACE} = -\mathbf{BCDE}. \end{aligned}$$

Each block by itself represents another 2^{5-2} fractional factorial from the *same family*, with the elements of the defining relation differing only in sign.

The elements of the defining contrast subgroup and the defining relation are called *words*. The number of factors in each interaction or word determines the *length* of the word. Ignoring the identity element, which has length 0, these defining relations each have two length-3 words and one word of length 4. The shortest word in any defining relation determines a fractional factorial design’s *resolution*, which is typically denoted by Roman numerals. These 2^{5-2} fractions have resolution III, since the defining relation contains no words of length 1 or length 2, but does contain words of length 3. The concept of resolution is useful, since it reflects the ability of the design to discriminate between effects. The larger the resolution the better, although, in general, larger resolution necessitates more treatment combinations. Detailed

discussion of various fractional factorial designs over the next three chapters is organized based on resolution.

The importance of resolution will become evident as we consider which (combinations of) effects can be estimated from a particular fraction. A full factorial model for a 2^5 contains 32 regression coefficients. With all 32 treatment combinations, we can estimate all 32 coefficients [refer to the saturated model (1.4)]. However, with only eight treatment combinations, we estimate linear combinations of coefficients.

Consider the eight treatment combinations defined by

$$\mathbf{I} = -\mathbf{ABD} = -\mathbf{ACE} = \mathbf{BCDE} : \quad (5.2)$$

A	B	C	D	E	AB	AC	BC	AD	BD	CD	... BCDE	ABCDE	
-1	-1	-1	-1	-1	1	1	1	1	1	1	1 ...	1	-1
1	1	1	-1	-1	1	1	1	-1	-1	-1	-1 ...	1	1
-1	1	-1	1	-1	-1	1	-1	-1	1	-1	-1 ...	1	-1
1	-1	1	1	-1	-1	1	-1	1	-1	1	-1 ...	1	1
1	1	-1	-1	1	1	-1	-1	-1	-1	1	-1 ...	1	1
-1	-1	1	-1	1	1	-1	-1	1	1	-1	-1 ...	1	-1
1	-1	-1	1	1	-1	-1	1	1	-1	-1	-1 ...	1	1
-1	1	1	1	1	-1	-1	1	-1	1	1	-1 ...	1	-1

With only eight observations, one can estimate an overall average and seven regression coefficients. Thus, there must be linear dependencies among the above columns. For instance, columns **A** and **ABCDE** are identical, and columns **BD** and **CE** are the negative of these; that is, $\mathbf{A} = -\mathbf{BD} = -\mathbf{CE} = \mathbf{ABCDE}$. We refer to **A**, $-\mathbf{BD}$, $-\mathbf{CE}$, and **ABCDE** as *aliases* since these are different names for the same column of ± 1 's. Such linear dependencies among the columns can be recognized from the defining relation for the fraction. Multiplying the defining relation (5.2) by factor **A** and simplifying,

$$\begin{aligned} \mathbf{A} \times (\mathbf{I} = -\mathbf{ABD} = -\mathbf{ACE} = \mathbf{BCDE}) &\Rightarrow \mathbf{A} = -\mathbf{A}^2\mathbf{BD} = -\mathbf{A}^2\mathbf{CE} = \mathbf{ABCDE} \\ &\Rightarrow \mathbf{A} = -\mathbf{BD} = -\mathbf{CE} = \mathbf{ABCDE} \end{aligned}$$

since any column of ± 1 's times itself becomes the identify column **I** and any column times **I** is unchanged. Note that if one were to multiply the defining relation by any alias of **A**, one would get the same equation. For example, if we multiply the defining relation by $-\mathbf{BD}$, the result is

$$\begin{aligned} -\mathbf{BD} \times (\mathbf{I} = -\mathbf{ABD} = -\mathbf{ACE} = \mathbf{BCDE}) \\ \Rightarrow -\mathbf{BD} = \mathbf{A} = \mathbf{ABCDE} = -\mathbf{CE}. \end{aligned}$$

So $-\mathbf{BD}$, $-\mathbf{CE}$, and **ABCDE** are the only aliases of **A**.

The other six alias sets for this fraction, obtained by multiplying different effects by the defining relation, are as follows:

- $\mathbf{B} = -\mathbf{AD} = -\mathbf{ABCE} = \mathbf{CDE}$

- $C = -ABCD = -AE = BDE$
- $D = -AB = -ABCE = BCE$
- $E = -ABDE = -AC = BCD$
- $BC = -ACD = -ABE = DE$
- $CD = -ABC = -ADE = BE$

These seven sets of aliases, plus the defining relation contain the labels for all 32 columns for the model matrix of the full factorial model. From this subset of eight treatment combinations, one cannot estimate any coefficients individually; rather one can estimate the following combinations:

- $\beta_0 - \beta_{ABD} - \beta_{ACE} + \beta_{BCDE}$
- $\beta_A - \beta_{BD} - \beta_{CE} + \beta_{ABCDE}$
- $\beta_B - \beta_{AD} - \beta_{ABCE} + \beta_{CDE}$
- $\beta_C - \beta_{ABCD} - \beta_{AE} + \beta_{BDE}$
- $\beta_D - \beta_{AB} - \beta_{ABCE} + \beta_{BCE}$
- $\beta_E - \beta_{ABDE} - \beta_{AC} + \beta_{BCD}$
- $\beta_{BC} - \beta_{ACD} - \beta_{ABE} + \beta_{DE}$
- $\beta_{CD} - \beta_{ABC} - \beta_{ADE} + \beta_{BE}$

What use is it to be able to estimate these combinations of effects, when in fact we are interested in estimating each main effect and perhaps some two-factor interactions? The answer is that data from fractional factorial designs are interpreted based on the following two assumptions:

- **Sparsity of important effects.** Only a few of the many possible effects are prominent. Even if many effects are nonzero, we expect a few to stand out as much larger than the rest.
- **Simplicity of important effects.** Main effects and/or two-factor interactions are more likely to be important than higher-order interactions. This is also known as the hierarchical ordering principle.

Most full factorial examples in Chapters 1–4 illustrate the reasonableness of these assumptions.

If one assumes that all three-factor (or higher) interactions are negligible, then from these eight observations one can estimate the following:

- β_0
- $\beta_A - \beta_{BD} - \beta_{CE}$
- $\beta_B - \beta_{AD}$
- $\beta_C - \beta_{AE}$
- $\beta_D - \beta_{AB}$
- $\beta_E - \beta_{AC}$
- $\beta_{BC} + \beta_{DE}$
- $\beta_{CD} + \beta_{BE}$

Perutka and Martell (2001) conducted an oxidation experiment with defining relation (5.2). The factor names and levels are given in Table 5.4, and the

results for one response appear in Table 5.5. Higher values of turnover number (TN) indicate greater efficiency for the catalyst. Perutka and Martell assumed all interactions to be zero, fitted a model for the five main effects, and obtained the following model for TN:

$$\widehat{\text{TN}} = 45.25 + 7\mathbf{A} + 17.25\mathbf{B} + 15\mathbf{C} - 6.25\mathbf{D} - 23.25\mathbf{E}. \quad (5.3)$$

With a root mean square error of 19.5 (based on 2 df), the standard error for each coefficient is $19.5/8^{1/2} = 6.9$. Using $\alpha = .20$, the authors reported that three factors have statistically significant effects. If their assumption about interactions is correct, this initial experiment suggests that by increasing **B** and **C** and reducing **E**, one can dramatically increase this response. If interactions are not all negligible, then the conclusions about which factors are important and the direction of their effects may be in error.

Table 5.4. Factors and levels for Perutka and Martell's (2001) experiment

Factors	Levels	
	-1	1
A Pyridine (mL)	2	5
B Adamantane (mmol)	1.25	5
C Oxygen (mL/min)	4	20
D Hydrogen sulfide (mL/min)	1	4
E Catalyst: dinuclear iron complex (mmol)	.005	.02

Table 5.5. Perutka and Martell's (2001) 2^{5-2} experiment for TN

A	B	C	D	E	TN
-1	1	1	1	1	46
-1	1	-1	1	-1	52
1	1	1	-1	-1	127
-1	-1	1	-1	1	14
1	-1	1	1	-1	54
1	1	-1	-1	1	25
1	-1	-1	1	1	3
-1	-1	-1	-1	-1	41

5.2.2 Need for additional data

Resolution III designs such as the eight runs just analyzed often leave ambiguity as to which effects are actually important. In addition to the risk of attributing a significant effect to the wrong coefficient in a set of aliases, a more insidious mistake will arise if two important coefficients that are aliased

have opposite signs and the sum near zero is construed to indicate that both of these aliased effects are negligible. There are two remedies to these potential difficulties:

- **Avoid use of resolution III fractional factorial designs;** that is, only use full factorial designs, or fractions with higher resolution.
- **Follow resolution III designs with additional experimentation.** Which additional treatment combinations are to be explored may be chosen after the initial experiment is completed and analyzed. Suitable follow-up experiments are discussed in Chapter 9.

For Perutka and Martell's (2001) experiment, what follow-up is recommended? One option is to collect data at just a couple of treatment combinations expected to produce the best outcomes. This is how Perutka and Martell proceeded, using the initial experiment combined with an understanding of the chemistry to select a follow-up run yielding TN = 215, much better than any outcome in Table 5.5! Alternatively, if one were less certain about the results and could afford a second eight-run experiment, one may add another quarter fraction [e.g., block 4 with defining relation (5.1)]. Combining these two quarter-fractions results in a half-fraction with defining relation

$$\begin{array}{r} (\mathbf{I} = -\mathbf{ABD} = -\mathbf{ACE} = \mathbf{BCDE}) \\ + (\mathbf{I} = \mathbf{ABD} = \mathbf{ACE} = \mathbf{BCDE}) \\ \hline (\mathbf{I} = \mathbf{BCDE}). \end{array}$$

The advantage of combining blocks 1 and 4 is that the resulting half-fraction is resolution IV rather than resolution III. It is always possible to increase the resolution from III to IV by adding a second fraction from the same family as the first. The details are presented in Section 9.4.

5.2.3 Construction of fractional factorial designs

There are two methods for constructing regular fractional factorial designs:

- **Blocked full factorial construction.** This method was illustrated in obtaining Table 5.3. Each of the four blocks of the full 2^5 corresponds to a resolution III 2^{5-2} fraction. This construction method clearly identifies all alternative fractions from the same family, which is useful if one is considering running a sequence of fractions.
- **Fractional factorial generator construction.** This is the most convenient construction method for a single fraction. We illustrate this construction for the 2^{5-2} fraction with defining relation $\mathbf{I} = -\mathbf{ABD} = -\mathbf{ACE} = \mathbf{BCDE}$. These eight treatment combinations represent a full factorial in any set of three factors besides $\{\mathbf{A}, \mathbf{B}, \mathbf{D}\}$ and $\{\mathbf{A}, \mathbf{C}, \mathbf{E}\}$. For instance, we have a full factorial in the first three factors. These three factors are designated *basic* factors, and we begin the construction of the 2^{5-2} by constructing a 2^3 factorial in them:

A	B	C
-1	-1	-1
1	-1	-1
-1	1	-1
1	1	-1
-1	-1	1
1	-1	1
-1	1	1
1	1	1

We now complete the design construction as follows. The additional factors will always be aliased with interactions of the basic factors. Here, $\mathbf{D} = -\mathbf{AB}$ and $\mathbf{E} = -\mathbf{AC}$. Thus, we compute the columns for these aliases by multiplying together the appropriate basic columns and relabeling them as factors \mathbf{D} and \mathbf{E} .

A	B	C	$-\mathbf{AB} = \mathbf{D}$	$-\mathbf{AC} = \mathbf{E}$
-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

These are the same eight treatment combinations that appear by (5.2), although in a different order. Thus, this fraction is identified either by its defining relation $\mathbf{I} = -\mathbf{ABD} = -\mathbf{ACE} = \mathbf{BCDE}$ or, equivalently, by the pair of generators $\mathbf{D} = -\mathbf{AB}$ and $\mathbf{E} = -\mathbf{AC}$.

5.2.4 General results for 2^{k-f} fractional factorial designs with defining relations

A regular 2^{k-f} fractional factorial design with no repeated treatment combinations has the following properties:

- Has k factors;
- Is a full factorial in some set of $k - f$ basic factors;
- Has a defining relation and defining contrast subgroup with 2^f elements;
- Aliases the remaining factorial effects in $2^{k-f} - 1$ sets of size 2^f ;
- Can be constructed either by
 - Partitioning the 2^k factorial into 2^f blocks, confounding the factorial effects in the defining contrast subgroup and choosing one block, or

- Creating a full factorial in $k - f$ basic factors, and then appending the f additional factors using as generators their aliases. These aliases are interactions in the basic factors.

We now enumerate all regular two-level fractional factorial designs of size 8 and use these to introduce notation for tabulating recommended designs of larger size. Consider the following table of seven orthogonal columns:

	A	B	AB	C	AC	BC	ABC
Col. No.	1	2	3	4	5	6	7
	-1	-1	1	-1	1	1	-1
	1	-1	-1	-1	-1	1	1
	-1	1	-1	-1	1	-1	1
	1	1	1	-1	-1	-1	-1
	-1	-1	1	1	-1	-1	1
	1	-1	-1	1	1	-1	-1
	-1	1	-1	1	-1	1	-1
	1	1	1	1	1	1	1

Columns labeled **A**, **B**, and **C** are the basic columns, and the other four columns are constructed from interactions of **A**, **B**, and **C**. Rather than listing the three basic columns first, they are listed in an order that will readily extend to larger cases. Additional basic columns will be needed for fractional factorial designs with more runs and will be numbered as successive powers of 2 (i.e., 8, 16, 32, etc.). Interactions of the basic columns are numbered corresponding to the sum of the basic column numbers; for example,

- Column 3 is the interaction of columns 1 and 2
- Column 5 is the interaction of columns 1 and 4
- Column 7 is the interaction of columns 1, 2, and 4

Using this notation, every type of orthogonal eight-run fractional factorial design may be constructed using columns 1, 2, 4, and one or more of columns 3, 5, 6, and 7. There are five distinct design types:

- Resolution IV 2^{4-1} : Basic columns + column 7
- Resolution III 2^{4-1} : Basic columns + column 3 (or column 5 or 6)
- Resolution III 2^{5-2} : Basic columns + any two other columns
- Resolution III 2^{6-3} : Basic columns + any three other columns
- Resolution III 2^{7-4} : Basic columns + all four other columns

Note that for the four-factor designs, it matters which column is chosen for the fourth factor. Clearly, defining **D** = **ABC** produces a different design than choosing **D** = **AB**; the designs have different resolution.

That it makes no difference which two columns one chooses for the 2^{5-2} designs is not obvious. Consider two possibilities. If one were to choose columns 3 and 5, the generators would be **D** = **AB** and **E** = **AC**, and the defining relation is **I** = **ABD** = **ACE** = **BCDE**. Alternatively, if one were to choose

columns 3 and 7, the generators are $\mathbf{D} = \mathbf{AB}$ and $\mathbf{E} = \mathbf{ABC}$, and the defining relation is $\mathbf{I} = \mathbf{ABD} = \mathbf{ABCE} = \mathbf{CDE}$. By swapping the letters \mathbf{A} and \mathbf{D} , one may see that these two designs are equivalent. Two regular fractional factorial designs are said to be *isomorphic* (i.e., equivalent) if by swapping letters (and if necessary, reversing the signs of letters) the defining relation for the first design can be made to match the other defining relation exactly.

For 16 runs, fractional factorial designs of resolution III or higher exist for $k = 5, 6, \dots, 15$ factors. How many distinct design types exist? The possibilities are listed in Chen, Sun, and Wu (1993, Table 2) and summarized in Table 5.6. Note that for 12 or fewer factors, it matters which columns are selected as generators. For eight or fewer factors, the resolution of the design is affected by the choice of generators. For 9–12 factors, although all designs are resolution III, the number of length-3 words in the defining relation depends on the choice of generators. Tables that list designs often list just a single “best” design. The most commonly used criterion for ranking fractional factorial designs, in addition to resolution, is *aberration*. To understand this property, we must define the *word length pattern* for a design.

Table 5.6. Number of nonisomorphic regular fractional factorial designs of size 16

No. Factors	Res. III	Res. IV	Res. V
5	1	1	1
6	3	1	0
7	4	1	0
8	5	1	0
9	5	0	0
10	4	0	0
11	3	0	0
12	2	0	0
13	1	0	0
14	1	0	0
15	1	0	0

5.2.5 Word length pattern and minimum aberration

Consider the four possible 2^{6-2} fractions:

- Design 6-2.1 with $\mathbf{I} = \mathbf{ABCE} = \mathbf{ABDF} = \mathbf{CDEF}$
- Design 6-2.2 with $\mathbf{I} = \mathbf{ABE} = \mathbf{ACDF} = \mathbf{BCDEF}$
- Design 6-2.3 with $\mathbf{I} = \mathbf{ABE} = \mathbf{CDF} = \mathbf{ABCDEF}$
- Design 6-2.4 with $\mathbf{I} = \mathbf{ABE} = \mathbf{ACF} = \mathbf{BCEF}$

Design 6-2.1 is resolution IV, and the other three designs are resolution III. The rank ordering of these designs from best to worst is in terms of the number

of short words in the defining relation. The *word length pattern* (wlp) for a regular fractional factorial design is the vector of frequencies of words of each length. Define

$$\text{wlp} = (A_3, A_4, \dots, A_k),$$

where A_j denotes the number of words of length j in the defining relation ($j = 3, \dots, k$). The word length patterns for these four designs are as follows:

- Design 6-2.1 with wlp = (0, 3, 0, 0)
- Design 6-2.2 with wlp = (1, 1, 1, 0)
- Design 6-2.3 with wlp = (2, 0, 0, 1)
- Design 6-2.4 with wlp = (2, 1, 0, 0)

The designs are sorted first based on the number of length-3 words. Using A_3 alone, Design 6-2.1 is better than Design 6-2.2, and both are preferred to Designs 6-2.3 and 6-2.4. Since the last two designs are tied with respect to A_3 , they are then compared on A_4 , where Design 6-2.3 is better than Design 6-2.4. Designs with more short words are said to have more aberration (i.e., more distortion as we seek to distinguish which effects are present). The design that ranks first on this criterion is called the *minimum aberration design*.

Appendix F lists the column number labels (up to 127) for each interaction of the basic factors. Appendix G uses these labels to identify generators for each minimum aberration design of size 8, 16, 32, 64, and 128. These tables are based on four references: Franklin (1984), Chen, Sun, and Wu (1993), Butler (2003a), and Block and Mee (2005). The use and analysis of these and other designs will appear in the next three chapters.

- **Chapter 6: Fractional Factorial Designs for Estimating Main Effects:** Resolution III fractional factorial designs are useful as initial experiments seeking to investigate the effects of many factors and for which follow-up experimentation is reasonably convenient. With these designs, the presence of interactions the same magnitude as main effects may well mislead the experimenter in terms of the relative importance of main effects.
- **Chapter 7: Designs for Estimating Main Effects and Some Two-Factor Interactions:** Resolution IV fractional factorial designs avoid aliasing between main effects and two-factor interactions. Thus, if all interactions involving three or more factors are zero, then a resolution IV design will permit unbiased estimation of all main effects. Additional degrees of freedom will be available for estimating two-factor interactions, although some sets of two-factor interactions will be aliased together. If the alias sets for two-factor interactions are large, then one cannot learn much about individual two-factor interactions, but at least the main effect estimates are not biased by these interactions.
- **Chapter 8: Resolution V Fractional Factorial Designs:** These permit estimation of all main effects and two-factor interactions, assuming

three-factor and higher-order interactions are negligible. If there are more than five factors, regular 2^{k-f} designs of resolution V are quite large. Chapter 8 presents smaller alternative orthogonal and nonorthogonal designs as well.

Chapters 9–11 complete the discussion of how to use these fractional factorial designs.

- **Chapter 9: Augmenting Fractional Factorial Designs.** The option to perform additional experimentation is often essential to the successful application of fractional factorial designs. Chapter 9 discusses several approaches to augmentation, including confirmation runs, foldover, semi-folding, and optimal design methods.
- **Chapter 10: Fractional Factorial Designs with Randomization Restrictions.** These designs are similar to those discussed in Chapter 3 for full factorial treatment structures. However, the presence of aliasing adds complexity to the construction of randomized block and split-unit designs. In Chapter 10 we address both design and analysis issues.
- **Chapter 11: More Fractional Factorial Design Examples.** These examples are presented to reinforce and illustrate the concepts of the previous six chapters.

Some orthogonal fractional factorial designs do not have defining relations. These include the 12-run Plackett–Burman design and other similar designs. Analysis of data from these *nonregular* designs differs somewhat from that presented below for regular fractional factorial designs. Examples of their analysis will be given later in Sections 6.3 and 7.1.5.

5.3 Basic Analysis for Regular Fractional Factorial Designs

This section illustrates a five-step analysis method for interpreting the results of regular fractional factorial experiments without centerpoint runs. The five steps are as follows:

1. Plot the response data.
2. Determine resolution, aliasing, and effects that can be estimated.
3. Fit a saturated model and use output to select a tentative reduced model.
4. Examine fit and diagnostics for the reduced model. Consider modifications to the model until a satisfactory summary is obtained.
5. Report results for the final model.

These steps are illustrated now for an 8-factor, 32-run experiment reported by Martin and Cuellar (2004) involving the coating of stainless-steel microbeads with a polymeric layer. Stainless-steel spheres with diameters 53–75 μ_m were coated in a suspension of styrene and divinylbenzene in water. The

authors listed approximately two dozen factors that affect the characteristics of the final polymeric particles, which they considered too many to investigate in a single experiment. Based on their experience and existing literature, Martin and Cuellar chose eight factors to vary, and they set all the other variables to selected fixed values. Their eight experimental factors are listed in Table 5.7 and the 32 treatment combinations of their experiment appear in Table 5.8.

Table 5.7. Factors and levels for Martin and Cuellar's (2004) polymeric coating experiment

Factors	Levels	
	-1	1
A Double polymerization	No	Yes
B Temperature (°C)	80	90
C Stirring speed (rev/min)	550	650
D Aqueous phase/organic phase	5	10
E Percentage of cross-linker	8	16
F Ammonium hydroxide	No	Yes
G Prepolymerization	No	Yes
H Initiator-metal contact	No	Yes

Since this design is a 2^{8-3} fraction, three generators were needed to construct the additional columns. Martin and Cuellar used $\mathbf{F} = \mathbf{BCD}$, $\mathbf{G} = \mathbf{CDE}$, and $\mathbf{H} = \mathbf{BDE}$, which produces an inferior design that is not minimum aberration. To recreate their design, construct a full 2^5 factorial in **A–E**, add the three interaction columns **BCD**, **CDE**, and **BDE**, and relabel these **F**, **G**, and **H**, respectively. These 32 treatment combinations were performed in the sequence indicated by the second column of Table 5.8.

Three responses were reported, each based on density calculations for the batch of coated spheres:

- R/r denotes the average ratio of the radius of the coated spheres R to the radius of the uncoated stainless-steel spheres r .
- Yield denotes the percentage (by weight) of the metal retained after excluding the agglomerated metal-polymeric particles
- P/S denotes the ratio of the polymer mass on the nonagglomerated particles to the total mass of steel for all particles before coating.

The experiment was intended to identify how to achieve as much coating as possible while still maintaining high Yield. Low values for R/r correspond to high Yield values, since balls with little or no coating are less likely to clump; the correlation between them is -0.832 . The third response is proportional to $[(R/r)^3 - 1] * \text{Yield}$. Maximizing this response is an attempt to find a compromise that simultaneously achieves high R/r and reasonable Yield.

Table 5.8. Martin and Cuellar's polymeric coating experiment

t.c.	Order	Run								R/r	Yield	P/S
		A	B	C	D	E	F	G	H			
1	30	-1	-1	-1	-1	-1	-1	-1	-1	1.0015	94.89	0.0006
2	32	1	-1	-1	-1	-1	-1	-1	-1	1.0238	60.86	0.0060
3	13	-1	1	-1	-1	-1	1	-1	1	1.0010	98.31	0.0004
4	18	1	1	-1	-1	-1	1	-1	1	1.0051	94.76	0.0020
5	24	-1	-1	1	-1	-1	1	1	-1	1.0000	96.69	0.0000
6	7	1	-1	1	-1	-1	1	1	-1	1.0041	70.38	0.0012
7	15	-1	1	1	-1	-1	-1	1	1	1.0128	50.63	0.0027
8	12	1	1	1	-1	-1	-1	1	1	1.0137	37.42	0.0021
9	19	-1	-1	-1	1	-1	1	1	1	1.0016	97.65	0.0006
10	1	1	-1	-1	1	-1	1	1	1	1.0049	61.88	0.0012
11	9	-1	1	-1	1	-1	-1	1	-1	1.0120	50.62	0.0025
12	29	1	1	-1	1	-1	-1	1	-1	1.0175	5.12	0.0004
13	17	-1	-1	1	1	-1	-1	-1	1	1.0034	98.66	0.0014
14	16	1	-1	1	1	-1	-1	-1	1	1.0058	94.03	0.0022
15	2	-1	1	1	1	-1	1	-1	-1	1.0000	96.22	0.0000
16	14	1	1	1	1	-1	1	-1	-1	1.0039	93.63	0.0015
17	25	-1	-1	-1	-1	1	-1	1	1	1.0225	45.95	0.0043
18	21	1	-1	-1	-1	1	-1	1	1	1.0261	28.52	0.0031
19	31	-1	1	-1	-1	1	1	1	-1	1.0164	70.67	0.0048
20	28	1	1	-1	-1	1	1	1	-1	1.0117	38.35	0.0019
21	23	-1	-1	1	-1	1	1	-1	1	1.0030	97.37	0.0012
22	20	1	-1	1	-1	1	1	-1	1	1.0026	94.10	0.0010
23	27	-1	1	1	-1	1	-1	-1	-1	1.0040	90.03	0.0015
24	26	1	1	1	-1	1	-1	-1	-1	1.0210	47.04	0.0041
25	11	-1	-1	-1	1	1	1	-1	-1	1.0030	98.50	0.0012
26	5	1	-1	-1	1	1	1	-1	-1	1.0067	93.15	0.0026
27	6	-1	1	-1	1	1	-1	-1	1	1.0025	95.27	0.0010
28	22	1	1	-1	1	1	-1	-1	1	1.0163	67.10	0.0045
29	10	-1	-1	1	1	1	-1	1	-1	1.0178	57.67	0.0043
30	8	1	-1	1	1	1	-1	1	-1	1.0274	22.24	0.0025
31	3	-1	1	1	1	1	1	1	1	1.0032	93.67	0.0012
32	4	1	1	1	1	1	1	1	1	1.0035	73.08	0.0010

Now we describe how to analyze these data.

Step 1. Plot the response data

Histograms and scatterplots are useful for understanding what variation is to be explained and whether any attractive outcomes were achieved. Figure 5.5 reveals that all three responses have skewed distributions. By highlighting the 16 observations with Yield > 90%, one can see that high Yield outcomes all added less than 1% to the radius ($R/r < 1.01$) and have a mass ratio of at

most 0.003. The greatest coating achieved with Yield $\geq 90\%$ is for observation 26 with $R/r = 1.0067$, Yield = 93.15, and P/S = 0.0026.

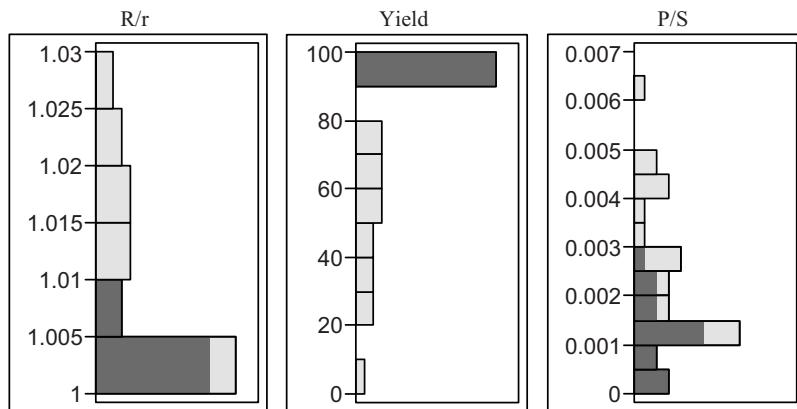


Fig. 5.5. Histograms for R/r , Yield, and P/S. Responses with Yield $> 90\%$ are highlighted

A thorough analysis would require building models for all three responses and then identifying trade-offs necessary to achieve sufficient coating with the maximum Yield. Here, we will focus on the steps to building a satisfactory model for Yield.

If we fit a model for the skewed response Yield, the data values with Yield $< 80\%$ will dominate because they represent most of the variation. Since we are interested in higher values for Yield, it makes sense to fit a model for a transformation that compresses values for low Yield and spreads out the values for high Yield. One reasonable choice is

$$\text{Sqrt}(\text{Loss}) = (1 - \text{Yield}/100)^{1/2}. \quad (5.4)$$

Figure 5.6 shows how this transformation dampens the variation where Loss is the greatest. Log(Loss) would be an alternative (and stronger) transformation.

Step 2. Determine resolution, aliasing and effects that can be estimated

This experiment is a one-eighth fraction, so the defining relation contains eight terms. The three generators create the length-4 words **BCDF**, **CDEG**, and **BDEH**. Multiply these together to obtain the entire defining relation:

$$(\mathbf{I} = \mathbf{BCDF}) \times (\mathbf{I} = \mathbf{CDEG}) \times (\mathbf{I} = \mathbf{BDEH}) \Rightarrow$$

$$\mathbf{I} = \mathbf{BCDF} = \mathbf{CDEG} = \mathbf{BEFG} = \mathbf{BDEH} = \mathbf{CEFH} = \mathbf{BCGH} = \mathbf{DFGH}$$

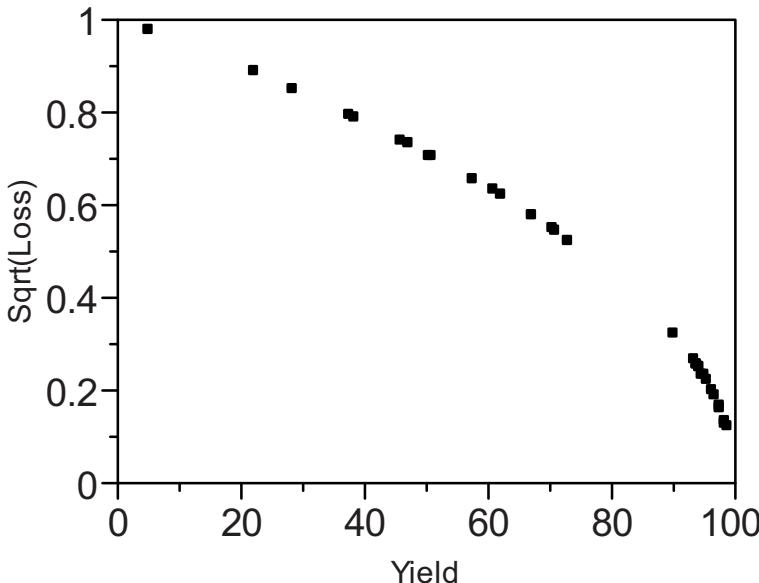


Fig. 5.6. $\text{sqrt}(\text{Loss})$ vs. Yield for 32 data values

This is a resolution IV design with seven length-4 words, four more than the minimum aberration 2^{8-3} design (refer to Table G.3 of Appendix G). Here, factor **A** does not appear in the defining relation. Therefore, this design may be written as a product of a 2^1 design for factor **A** and a 2^{7-3} fraction for factors **B – H**.

Given the defining relation, we now identify which (combinations of) effects can be estimated. Fitting a model containing all 8 main effects and 28 two-factor interactions results in the following linear dependencies among the columns of the model matrix:

$$\begin{aligned}
 \mathbf{BF} &= \mathbf{CD} = \mathbf{EG} \\
 \mathbf{BD} &= \mathbf{CF} = \mathbf{EH} \\
 \mathbf{BH} &= \mathbf{CG} = \mathbf{DE} \\
 \mathbf{BG} &= \mathbf{CH} = \mathbf{EF} \\
 \mathbf{BC} &= \mathbf{DF} = \mathbf{GH} \\
 \mathbf{CE} &= \mathbf{DG} = \mathbf{FH} \\
 \mathbf{BE} &= \mathbf{DH} = \mathbf{FG}
 \end{aligned}$$

Thus, with this experiment, we can estimate the eight main effects, the seven two-factor interactions involving factor **A** (since they are not aliased with other effects in the model), plus seven sums of coefficients (such as $\beta_{BF} + \beta_{CD} + \beta_{EG}$), assuming three-factor and higher-order interactions are negligible. The remaining nine degrees of freedom may be used to estimate the error variance, as well as check the assumption of no higher-order interactions.

Step 3. Fit a saturated model and use output to select a tentative reduced model

For most fractional factorial designs, including this resolution IV 2^{8-3} , the simplest way to fit a saturated model is to specify a full factorial model in the basic factors. For experiments without replication, fitting a saturated model will produce an estimate for the standard error of each coefficient (via Lenth's PSE), based only on an assumption of the sparsity of important effects. After identifying the number of important effects, one can use the defining relation and the corresponding aliasing to interpret which effect(s) in each significant alias set is most plausibly present. We now illustrate these steps for the response $\text{Sqrt}(\text{Loss})$. A Pareto plot of the 31 effect estimates is shown in Figure 5.7. Remember that each of these terms has seven aliases, which we will take into account after determining which estimates are large enough to include.

Using Lenth's t statistics, six effects are found to be statistically significant at $\alpha = .05$, three dominant effects and three modest ones. Pleasingly, five of the six significant effects are main effects, including all three dominant estimates, since $\mathbf{CDE} = \mathbf{G}$, $\mathbf{BCD} = \mathbf{F}$, and $\mathbf{BDE} = \mathbf{H}$. The simple model with only \mathbf{A} , \mathbf{F} , and \mathbf{G} explains 80.4% of the variation in $\text{Sqrt}(\text{Loss})$. Because of the dominance of the main effects for \mathbf{A} , \mathbf{F} , and \mathbf{G} , we consider the significant estimate associated with the aliases

$$\mathbf{ABE} = \mathbf{ADH} = \mathbf{AFG}$$

as evidence for the three-factor interaction \mathbf{AFG} . This judgment is consistent with Martin and Cuellar (2004, p. 2100), who construed \mathbf{FG} to be active rather than its aliases \mathbf{BE} and \mathbf{DH}). To ensure that our model is hierarchical, inclusion of \mathbf{AFG} in the model requires that we add \mathbf{AF} , \mathbf{AG} , and \mathbf{FG} as well.

Using the .05 level of significance, one would include only five of the eight main effects. However, since all 8 main effects are among the 10 largest estimates, many analysts would include \mathbf{C} , \mathbf{D} , and \mathbf{E} as well. At some risk of overfitting the model, we adopt the hierarchical model containing eight main effects and four interactions.

Some software automates Steps 2 and 3. For example, JMP's Modeling – Screening platform requires only a list of the factors and the responses. For a regular 2^{k-f} design, it will then add interactions to produce a saturated model, compute the PSE, Lenth t statistics, and their corresponding p -values, show aliasing, highlight terms with p -values $< .10$, and produce a half-normal plot. Figure 5.8 shows the results for the Martin and Cuellar's data, with (5.4) as the response. Actually, JMP lists aliasing up to four-factor interactions here, because it required a four-factor interaction to create a saturated model. For the alias list in Figure 5.8, we omitted the main effects' three-factor interaction aliases; we also omitted four-factor interaction aliases of the two-factor interactions. Note that JMP arranges the 31 terms for the saturated model intelligently. First, the eight main effect estimates are listed from largest to smallest. Second, the 14 two-factor interaction estimates are

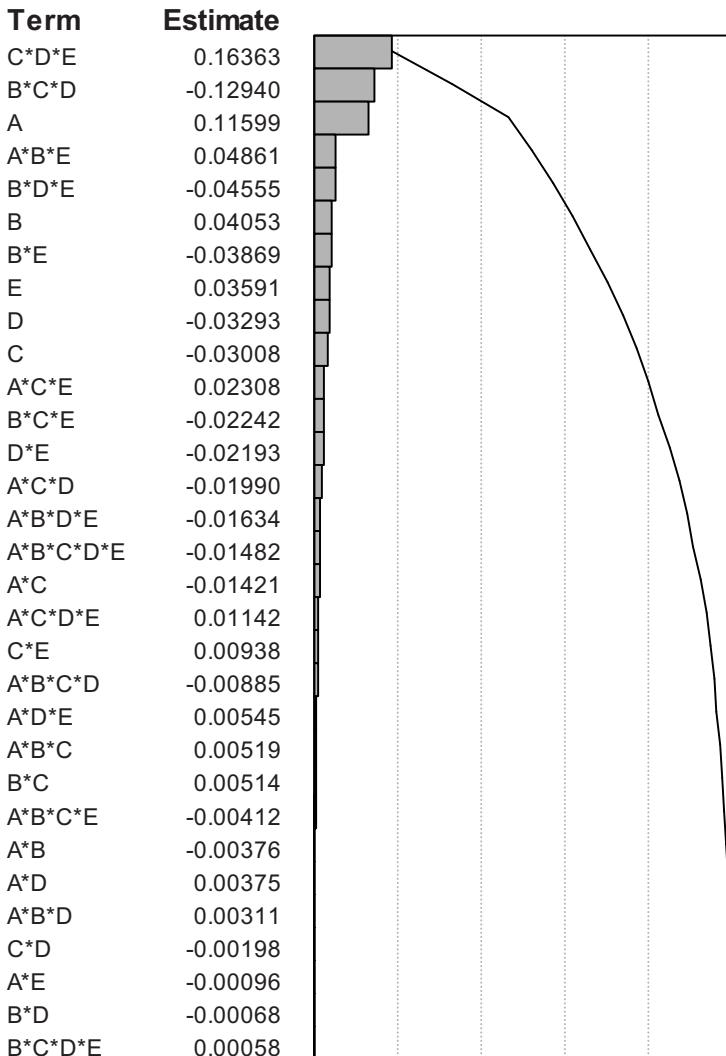


Fig. 5.7. Pareto plot of estimates from saturated model for $\text{Sqrt}(\text{Loss})$

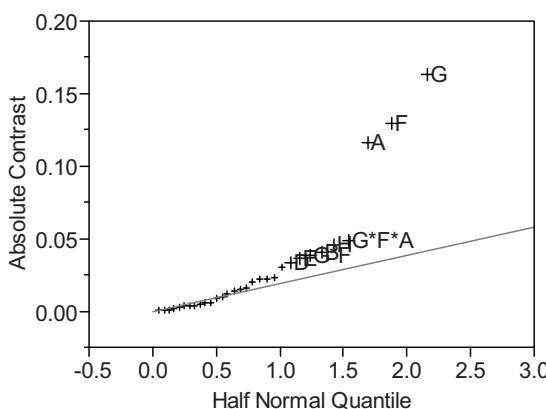
listed, not from largest to smallest, but in an order determined by the magnitudes of the main effects. Since **G** and **F** have the largest estimates, their interaction is listed first and labeled as **GF**, with aliases **BE** and **HD**. Next come three-factor interactions. Note that **GFA** is labeled and ordered as it is because it involves the three factors with the largest main effects.

Screening for Sqrt(Loss)

<u>Term</u>	<u>Contrast</u>	<u>Lenth t-Ratio</u>	<u>Individual p-Value</u>	<u>Aliases</u>
G	0.164	8.51	<.0001	
F	-0.129	-6.73	0.0002	
A	0.116	6.03	0.0003	
H	-0.046	-2.37	0.0302	
B	0.041	2.11	0.0483	
E	0.036	1.87	0.0735	
D	-0.033	-1.71	0.0951	
C	-0.030	-1.56	0.1250	
G*F	-0.039	-2.01	0.0576	B*E, H*D
G*A	0.011	0.59	0.5718	
F*A	-0.009	-0.46	0.6629	
G*H	0.005	0.27	0.8053	F*D, B*C
F*H	0.009	0.49	0.6443	G*D, E*C
A*H	-0.016	-0.85	0.3787	
G*B	0.001	0.03	0.9796	F*E, H*C
F*B	-0.002	-0.10	0.9265	G*E, D*C
A*B	-0.004	-0.20	0.8583	
H*B	-0.022	-1.14	0.2444	E*D, G*C
A*E	-0.001	-0.05	0.9648	
H*E	-0.001	-0.04	0.9760	B*D, F*C
A*D	0.004	0.20	0.8587	
A*C	-0.014	-0.74	0.4462	
G*F*A	0.049	2.53	0.0239	A*B*E, A*H*D
G*A*H	0.005	0.27	0.8025	F*A*D, A*B*C
F*A*H	0.023	1.20	0.2257	G*A*D, A*E*C
G*A*B	-0.015	-0.77	0.4274	F*A*E, A*H*C
F*A*B	-0.020	-1.04	0.2839	G*A*E, A*D*C
F*H*B	-0.022	-1.17	0.2349	G*H*E, G*B*D, etc.
A*H*B	0.005	0.28	0.7914	A*E*D, G*A*C
A*H*E	0.003	0.16	0.8846	A*B*D, F*A*C
F*A*H*B	-0.004	-0.21	0.8449	G*A*H*E, G*A*B*D, etc.

Lenth PSE=0.01922

P-Values derived from a simulation of 10000 Lenth t ratios.

**Fig. 5.8.** JMP's Modeling–Screening analysis for Sqrt(Loss)

Step 4. Examine fit and diagnostics for the reduced model; consider modifications until a satisfactory summary is obtained

A residual versus predicted plot (Figure 5.9) from this proposed reduced model shows constant variability across the range of predicted responses. Next, we plot residuals versus run order (Figure 5.10) to check on stability of the error distribution over time. No evidence of trend or autocorrelation appears, so this reduced model appears to be satisfactory.

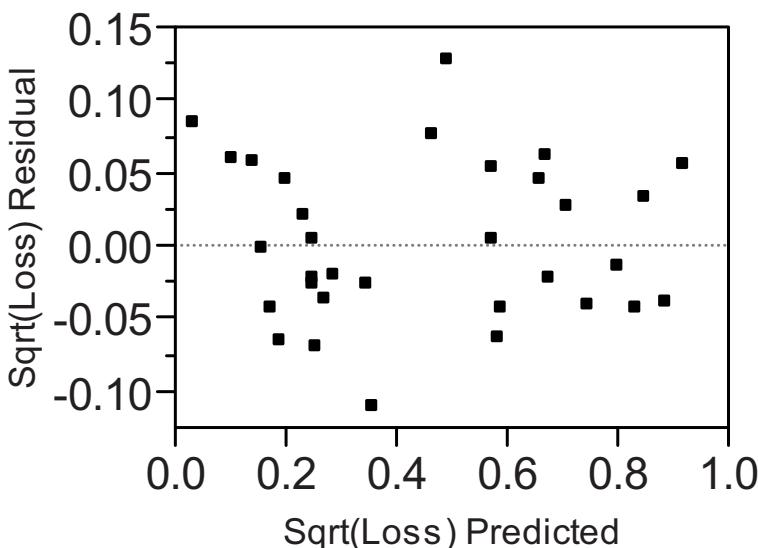
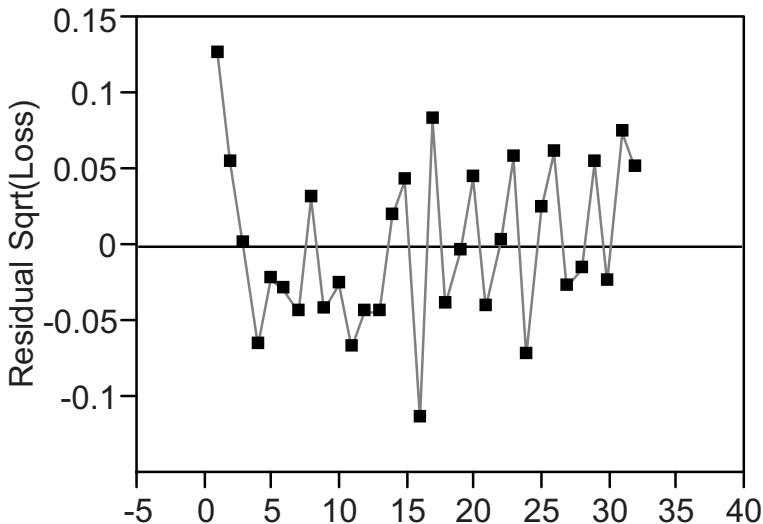


Fig. 5.9. Residuals versus predicted values for $\text{Sqrt}(\text{Loss})$

Step 5. Report results for the final model

The ANOVA and parameter estimates for our fitted model appear in Figure 5.11. This model has $R^2 = 96\%$. Because the effects of the three prominent factors are not additive but involve a three-factor interaction, their joint effect is best visualized using a cube plot that displays the predicted response at each of the eight combinations (see Figure 5.12). To assist with interpretability of this plot, natural variable names and units are used rather than coded labels.

As expected, Prepolymerization and Double polymerization both increase loss due to agglomeration, and the addition of Ammonium hydroxide to the suspension medium suppresses agglomeration (and Loss). The predicted values here represent the expected $\text{Sqrt}(\text{Loss})$, averaging over the levels of the other five factors. Several of the other five factors are deemed active, but their effects are small.

**Fig. 5.10.** Residuals versus run order for Sqrt(Loss)**Analysis of Variance**

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	12	2.1772608	0.181438	38.0350
Error	19	0.0906358	0.004770	Prob > F
C. Total	31	2.2678965		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	0.4537174	0.01221	37.16	<.0001
A	0.1159882	0.01221	9.50	<.0001
B	0.0405297	0.01221	3.32	0.0036
C	-0.030078	0.01221	-2.46	0.0235
D	-0.032934	0.01221	-2.70	0.0143
E	0.0359135	0.01221	2.94	0.0084
F	-0.129401	0.01221	-10.60	<.0001
G	0.1636349	0.01221	13.40	<.0001
H	-0.045551	0.01221	-3.73	0.0014
A*F	-0.008855	0.01221	-0.73	0.4771
A*G	0.0114207	0.01221	0.94	0.3613
F*G	-0.038692	0.01221	-3.17	0.0051
A*F*G	0.0486058	0.01221	3.98	0.0008

Fig. 5.11. Summary of reduced model for Sqrt(Loss)

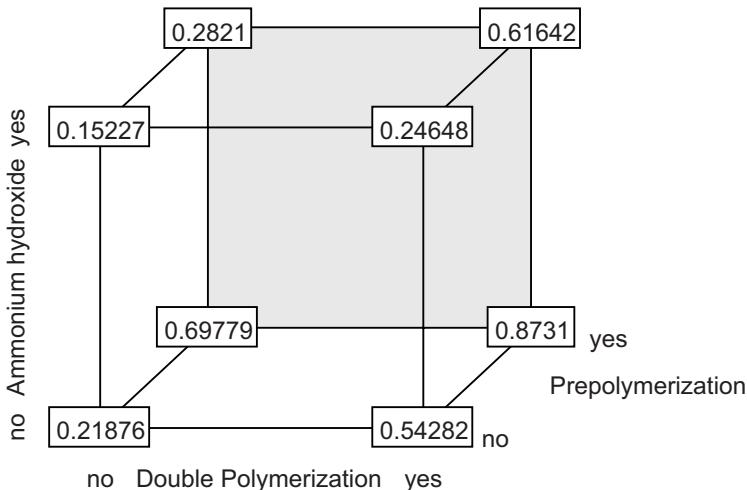


Fig. 5.12. Predicted Sqrt(Loss) based on three prominent factors

Martin and Cuellar (2001) took Yield directly as the response. They began by fitting a model that assumed no three-factor interactions, rather than the saturated model we considered, and chose the reduced model

$$\widehat{\text{Yield}\%} = 72.33 - 10.97\mathbf{A} + 13.20\mathbf{F} - 16.04\mathbf{G} + 4.45\mathbf{H} + 5.81\mathbf{FG}.$$

Their simple model for Yield has $R^2 = 85.3\%$, but it predicts yields in excess of 100% for 4 of the 32 treatment combinations in the experiment. By using a transformation in Figure 5.6 and fitting a saturated model, we arrived at a reduced model that includes the three-factor interaction **AFG** and has no predicted values outside of the feasible range.

Fractional Factorial Designs for Estimating Main Effects

This chapter focuses on efficient designs intended for estimating main effects, including regular resolution III 2^{k-f} fractional factorial designs, Plackett–Burman and other designs based on Hadamard matrices, nonorthogonal saturated main effect designs, and supersaturated designs. These designs are useful for identifying important factors when it is reasonable to expect that their effects are essentially additive. Even when the assumption of additive effects is suspect, these designs can produce useful initial experiments, provided they are augmented with additional runs. Whenever possible, we will explore evidence for two-factor interactions, even with these screening experiments. The sections are as follows:

Section 6.1. Analysis of Regular Resolution III Fractional Factorial Designs

Section 6.2. Some Theory Regarding Regular Resolution III Designs

Section 6.3. Nonregular Orthogonal Designs of Strength 2

Section 6.4. Optimal Nonorthogonal Saturated Main Effect Designs

Section 6.5. Supersaturated Designs

Section 6.6. Conclusions

Unreplicated resolution III fractional factorial designs are especially popular in the following circumstances:

- Experimentation is costly or time-consuming.
- Interactions are not expected to be important.
- The error variance is small.
- There are many factors to investigate.
- Follow-up experimentation is feasible.

Each of these characteristics contribute to the practical usefulness of resolution III designs.

This chapter includes discussion and analysis of the following nine published examples:

1. Example 6.1: Vindevogel and Sandra (1991) used an unreplicated 2^{5-2} design to improve the resolution of electrokinetic chromatography without lengthening the analysis time.
2. Example 6.2: Lai, Pan, and Tzeng (2003) sought to improve lovastatin yield using a series of replicated designs that included a 2^{7-4} and a 2^{6-3} .
3. Example 6.3: Irvine, Clark, and Recupero (1996) presented a 16-run laboratory wood chip pulping experiment involving 13 factors and several responses.
4. Example 6.4: Poorna and Kulkarni (1995) investigated 15 parameters for a fermentation process to produce inulinase, which is used in high-fructose syrups. Following preliminary work involving several single-factor experiments, the 15 parameters were examined using a 16-run, 2^{15-11} fractional factorial design.
5. Example 6.5: Bullington et al. (1993) presented results from an 11-factor, 12-run experiment to identify the causes for early thermostat failures.
6. Example 6.6: Bermejo-Barrera et al. (2001) varied seven factors in an investigation of optimizing atomic absorption spectrometry for the determination of trace elements in seafood. Their initial experiment was based on a 12-run Plackett–Burman design.
7. Example 6.7: Wu et al. (2005) illustrated the use of experimental design to optimize RoBioVision image analysis software for quantifying cDNA microarray images. They explored 19 software parameters using a 20-run Plackett–Burman design.
8. Example 6.8: Bell, Ledolter, and Swersey (2006) carefully documented a direct mail sales experiment involving 100,000 test mailings, 5000 for each of 20 different credit card offers, based on the 20-run Plackett–Burman design.
9. Example 6.9: Lin (1995) presented a supersaturated design with 138 factors in 24 runs of an AIDS model.

Examples 6.1–6.4 are regular fractional factorial designs, the type introduced in Chapter 5. Examples 6.5–6.8 are strength-2 orthogonal arrays discussed in Section 6.3, and Example 6.9, a supersaturated design, is discussed in Section 6.5.

6.1 Analysis of Regular Resolution III Fractional Factorial Designs

Example 6.1: An unreplicated 2^{5-2}

Vindevogel and Sandra (1991) provided a detailed description of their fractional factorial experiment and the underlying theory to explain the effects of

the factors being investigated. Their experiment studied five buffer composition factors, with the levels listed in Table 6.1. The experiment involved use of electrokinetic chromatography to resolve four different testosterone esters, two of which were difficult to distinguish. Regarding the choice of factor levels, Vindevogel and Sandra stated, “No fixed rules exist for the selection of low and high levels. Level selection was based on experience from preliminary experiments. Some constraints can be used by testing the buffer composition that is assumed to cause the highest current or an excessively long analysis time. In this way, a planned setup with a higher surfactant concentration (50 vs. 60 mM) was eliminated” (p. 1532).

Table 6.1. Factors and levels for Vindevogel and Sandra’s (1991) chromatography experiment

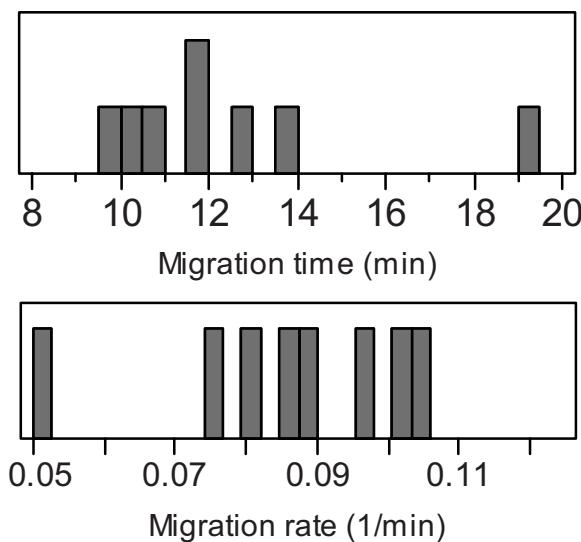
Factors	Levels	
	-1	1
A pH of pure buffer	8	9
B Surfactant % (sodium heptyl sulfate)	0	10
C Acetonitrile (%)	40	50
D Surfactant concentration (mM)	40	50
E Buffer concentration (mM)	20	40

The authors opted for a small experiment that could be run in one day, in order to keep the error variance small. They stated, “It is our experience that even when run-to-run variability is acceptable, day-to-day reproducibility is less reliable.... Collecting the data in as short a time as possible promotes the internal coherence of these data and allows us to draw conclusions that are not influenced by long term drift.” Thus, the authors opted for an eight-run experiment. The experimental runs, in the order performed, and three response variables are presented in Table 6.2. The first two responses (t_0 , t_4) present the elution time window in minutes. R_S is a measure of resolution for the most difficult separation; for the four runs with the poorest resolution, R_S is approximate, since providing any number was difficult. Vindevogel and Sandra (1991) discussed additional responses (noise, current, and efficiency) that are not included here.

Table 6.2. Vindevogel and Sandra's (1991) chromatography experiment

Buffer	A	B	C	D	E	t_0	t_4	R_S	
1	1	1	1	-1	1	-1	6.17	13.67	1.44
2	1	1	1	-1	-1		7.38	10.60	0.2
3	-1	1	1	1	1		7.42	11.78	0.5
4	-1	-1	1	1	-1		7.27	11.76	0.6
5	1	-1	-1	1	1		6.52	19.37	1.96
6	-1	1	-1	-1	1		5.55	9.79	0.73
7	1	-1	1	-1	1		7.75	12.51	0.6
8	-1	-1	-1	-1	-1		5.61	10.01	0.84

The authors assume no interaction effects exist and use the 2 df for two-factor interactions as error. We illustrate their analysis below, and then comment about the alternative approach of using Lenth's PSE. In our analysis of t_4 , we use the reciprocal, modeling migration rate for the fourth eluting compound rather than migration time t_4 . Figure 6.1 shows how this alters the variation to be explained. A model for $1/t_4$ is expected to be more successful at explaining differences between the lower t_4 values, since the reciprocal accentuates this variation.

**Fig. 6.1.** Histogram for migration time t_4 and rate $1/t_4$

Fitting a main effects model for each response supports the assumption of no interactions, in that each first-order model explains more than 99% of the variation. The fitted models and their root mean square errors are

$$\begin{aligned}\hat{t}_0 &= 6.71 + \underline{0.25\mathbf{A}} - 0.08\mathbf{B} + \underline{0.75\mathbf{C}} + \underline{0.14\mathbf{D}} + 0.10\mathbf{E}; \hat{\sigma} = .075, \\ 100(\widehat{1/t_4}) &= 8.39 - \underline{0.91\mathbf{A}} + \underline{0.48\mathbf{B}} + 0.22\mathbf{C} - \underline{1.02\mathbf{D}} - \underline{0.42\mathbf{E}}; \hat{\sigma} = .266, \\ \hat{R}_S &= 0.86 + \underline{0.19\mathbf{A}} - \underline{0.14\mathbf{B}} - \underline{0.38\mathbf{C}} + \underline{0.27\mathbf{D}} + \underline{0.09\mathbf{E}}; \hat{\sigma} = .043.\end{aligned}$$

Using the critical value of 4.303 from the Student's t distribution with 2 df, all underlined coefficients are statistically significant at $\alpha = .05$.

Note that the coefficients for the second and third models are opposite in sign. This means that factor levels that improve the resolution also decrease the migration rate and, hence, lengthen the time required to obtain the result. If increasing the migration rate is critical, then some compromise is required. Acetonitrile $\leq 40\%$ ($\mathbf{C} \leq -1$) is preferred, since this has the largest influence on resolution, but it is not statistically significant for $1/t_4$. Since pH (**A**) and Buffer Concentration (**E**) are the two factors with the largest ratio of coefficients for migration rate versus resolution, one or both of these might be set near to the low level, in order to increase the migration rate, and hence shorten the time required. Based on further testing that is not described in detail, Vindevogel and Sandra (1991) proposed the treatment combination (**A**, **B**, **C**, **D**, **E**) = $(-0.7, -1, -2, 1, -1)$ as an optimal compromise and verified that it produced superior results. At this proposed treatment combination, the estimated migration rate is $0.0752 = 1/(13.3 \text{ min})$, with an estimated resolution of 1.81. Note how this combination has a migration rate similar to buffer 1 in Table 6.2, but with resolution closer to that for buffer 5.

With only $N-1 = 7$ df, Lenth's approach cannot show so many main effects to be statistically significant. In fact, only t_0 's $b_C = 0.75$ has a Lenth's t that is statistically significant at $\alpha = .05$. Recall that Lenth's method for estimating the standard error of effects is based on an assumption that relatively few effects are active. With only 7 df for a saturated model, Lenth's procedure is suitable if two or fewer effects are important, but it is inadequate when all the main effects are active.

Even though this experiment was quite successful for estimating the factor effects, having only 2 df for error is a weakness of this design and a potential limitation. Adding replication at the design center or having prior knowledge about the magnitude of σ is particularly beneficial for such cases.

One final comment is warranted for this example. To this point we have ignored the aliasing for this design. The 2^{5-2} design in Table 6.2 was obtained using the generators $\mathbf{D} = -\mathbf{AC}$ and $\mathbf{E} = -\mathbf{AB}$, so that the defining relation is $\mathbf{I} = -\mathbf{ACD} = -\mathbf{ABE} = \mathbf{BCDE}$. Aliasing of main effects and two-factor interactions is

$$\mathbf{A} = -\mathbf{BE} = -\mathbf{CD}$$

$$\mathbf{B} = -\mathbf{AE}$$

$$\mathbf{C} = -\mathbf{AD}$$

$$\mathbf{D} = -\mathbf{AC}$$

$$\mathbf{E} = -\mathbf{AB}$$

$$\mathbf{BC} = \mathbf{DE}$$

$$\mathbf{BD} = \mathbf{CE}$$

Thus, the success of the main effects models suggests that neither factor **B** or **E** is involved in a two-factor interaction with factor **C** or **D**. It says nothing about other two-factor interactions, since the other six two-factor interactions are aliased with main effects and their presence would bias main effect estimates but would not show up in error for the first-order model. If follow-up runs at new treatment combinations did not agree with the results predicted by the first-order models, one likely suspect would be the presence of an interaction aliased with a main effect. For instance, the defining relation word **ACD** implies that an interaction between two of these factors would bias the main effect for the third.

Unreplicated fractional factorial experiments are suitable when the error variation is small relative to the magnitude of effects. Further, the success of resolution III fractions is contingent on the simplicity of the model. We now consider another example in which replication was employed, but a simple model was not adequate.

Example 6.2: A pair of less successful replicated experiments

Lai, Pan, and Tzeng (2003) reported a series of experiments conducted to enhance lovastatin production. Their experimentation began with seven ingredient factors - see Table 6.3. Eight treatment combinations were investigated, each replicated three times to increase the precision of the results. The authors stated that the order for the 24 runs was completely randomized, but the particular order is not reported. Their 2^{7-4} fraction and the lovastatin production values are reported in Table 6.4.

Table 6.3. Factors and levels for Lai et al.'s first lovastatin experiment

Factors	Levels	
	-1	1
A Lactose (g/L)	10	50
B Glucose (g/L)	10	50
C Peptone (g/L)	0	20
D Corn steep liquor (g/L)	0	20
E Soybean meal (g/L)	0	6
F Yeast extract (g/L)	0	40
G Ammonium sulfate (g/L)	0	10

Table 6.4. Lai et al.'s first lovastatin production experiment

t.c.	A	B	C	D	E	F	G	Lovastatin (g/L)
1	-1	-1	-1	1	1	1	-1	0.4032, 0.4256, 0.4494
2	1	-1	-1	-1	-1	1	1	0.0924, 0.1140, 0.1002
3	-1	1	-1	-1	1	-1	1	0.2898, 0.2794, 0.2229
4	1	1	-1	1	-1	-1	-1	0.2264, 0.2498, 0.2033
5	-1	-1	1	1	-1	-1	1	0.1622, 0.1456, 0.1433
6	1	-1	1	-1	1	-1	-1	0.3341, 0.3532, 0.2949
7	-1	1	1	-1	-1	1	-1	0.2472, 0.2396, 0.2116
8	1	1	1	1	1	1	1	0.2745, 0.2656, 0.2929

Analysis of these data is straightforward. A plot of the 24 lovastatin assay values by treatment combination demonstrates that the error variance is small (see Figure 6.2). The main effects model (1.1) is a saturated model; it explains 96.3% of the variation in lovastatin yield, with $MSE = 0.0005$. The estimated coefficients, with their t statistics and p -values, are given in Table 6.5. There is no reason to fit a reduced model here. The replication provides more than enough degrees of freedom for estimating σ^2 with the pure error mean square.

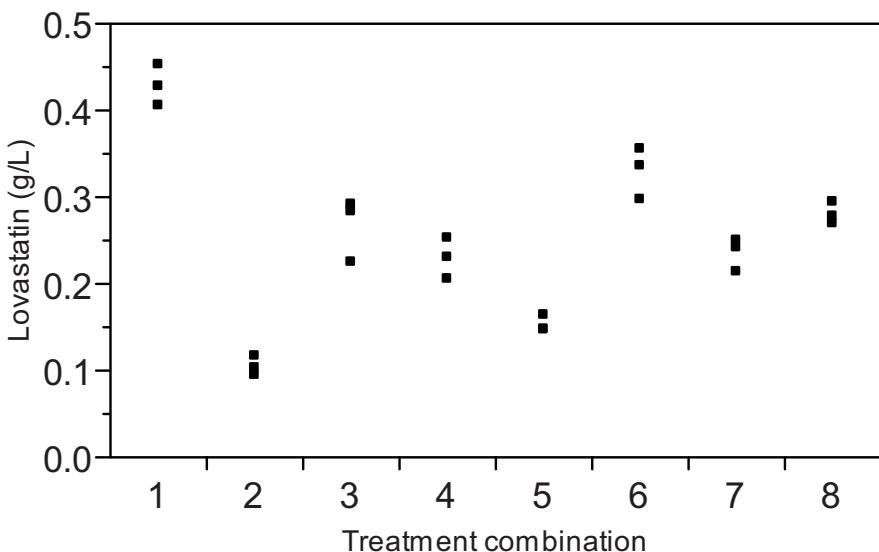
**Fig. 6.2.** Lovastatin production by treatment combination

Table 6.5. Saturated model for first replicated lovastatin experiment

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	0.251	0.0046	54.79	.000
A	-0.017	0.0046	-3.81	.002
B	-0.001	0.0046	-0.14	.892
C	-0.004	0.0046	-0.83	.416
D	0.019	0.0046	4.21	.001
E	0.073	0.0046	15.92	.000
F	0.009	0.0046	1.92	.072
G	-0.052	0.0046	-11.43	.000

Four of the estimates are statistically significant at $\alpha = .05$. The coefficients are estimated with good precision; 95% confidence intervals are constructed as $b_i \pm 2.12(0.0046)$. The largest estimates are $b_E = 0.073$ and $b_G = -0.052$, indicating that adding Soybean meal enhances production, while adding Ammonium sulfate is detrimental. Adding Corn steep liquor (**D**), and perhaps Yeast extract (**F**), is helpful also.

At this point, the authors might have attempted to use steepest ascent (see Section 9.3) to explore outside the original experimental region for a location with substantially improved lovastatin yield. Instead, they tried another fractional factorial design. Lai et al. (2003) report that in their next experiment they fixed the levels for factors **B** and **G** and added one new factor, Glycerol. The levels chosen for this second experiment are shown in Table 6.6. Why factor and level changes were made is not the focus here but rather the extent to which the effect estimates change. Data for the second experiment, a replicated 2^{6-3} fraction, are given in Table 6.7. Again, complete randomization of run order was followed. Factor labels with a “†” indicate that levels coded -1 and/or +1 have changed since the first experiment.

Table 6.6. Comparison of factor levels for two lovastatin experiments

Factors	First Experiment	Second Experiment
	Levels	Levels
A Lactose	(10, 50)	(20, 40)
B Glucose	(10, 50)	(20)
C Peptone	(0, 20)	(0, 20)
D Corn steep liquor	(0, 20)	(0, 20)
E Soybean meal	(0, 6)	(2, 10)
F Yeast extract	(0, 40)	(0, 20)
G Ammonium sulfate	(0, 10)	(0)
H Glycerol	(0)	(10, 30)

Table 6.7. Lai et al.'s second replicated lovastatin experiment

t.c.	\mathbf{A}^\dagger	\mathbf{H}	\mathbf{C}	\mathbf{D}	\mathbf{E}^\dagger	\mathbf{F}^\dagger	Lovastatin (g/L)
9	-1	-1	-1	1	1	1	0.0906, 0.0845, 0.0626
10	1	-1	-1	-1	-1	1	0.4160, 0.3781, 0.3521
11	-1	1	-1	-1	1	-1	0.1843, 0.2038, 0.1985
12	1	1	-1	1	-1	-1	0.5803, 0.6050, 0.5562
13	-1	-1	1	1	-1	-1	0.2017, 0.2734, 0.2560
14	1	-1	1	-1	1	-1	0.4405, 0.4290, 0.4205
15	-1	1	1	-1	-1	1	0.0995, 0.0612, 0.0879
16	1	1	1	1	1	1	0.0744, 0.0669, 0.0746

This 2^{6-3} design has 1 df for estimating two-factor interactions. Table 6.7's design generators are $\mathbf{E}^\dagger = \mathbf{A}^\dagger\mathbf{C}$, $\mathbf{F}^\dagger = \mathbf{A}^\dagger\mathbf{CD}$, and $\mathbf{H} = \mathbf{A}^\dagger\mathbf{D}$. The two-factor interaction contrast $\mathbf{CD} = \mathbf{A}^\dagger\mathbf{F}^\dagger = \mathbf{E}^\dagger\mathbf{H}$ is not aliased with any main effects. Fitting a saturated model produces the estimates in Table 6.8. Every estimate is highly significant, including the combination of two-factor interactions. Lai et al. fitted just a first-order model and so missed evidence for the interaction(s). If they had examined a lack-of-fit test, it would have been $F_{\text{laf}} = 60.79$, which is the square of the t statistic for the interaction they omitted. By using the MSE for this model with lack-of-fit, they also failed to recognize the statistical significance for \mathbf{D} .

Table 6.8. Saturated model for second lovastatin experiment

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	0.258	0.0045	57.66	<.0001
\mathbf{A}^\dagger	0.108	0.0045	24.09	<.0001
\mathbf{C}	-0.051	0.0045	-11.41	<.0001
\mathbf{D}	-0.014	0.0045	-3.21	0.0054
\mathbf{E}^\dagger	-0.064	0.0045	-14.30	<.0001
\mathbf{F}^\dagger	-0.104	0.0045	-23.27	<.0001
\mathbf{H}	-0.026	0.0045	-5.70	<.0001
$\mathbf{CD} = \mathbf{A}^\dagger\mathbf{F}^\dagger = \mathbf{E}^\dagger\mathbf{H}$	-0.035	0.0045	-7.80	<.0001

In the initial 2^{7-4} experiment, every interaction is aliased with a main effect. In the second experiment, only the aliases of \mathbf{CD} are not aliased with a main effect. If all these interactions were negligible, then the coefficients for the five factors explored in both experiments should agree. If they do not agree, this is evidence for interaction effects (or nonlinearity for \mathbf{E} and \mathbf{F}). To assess this problem, we combine the data for the two experiments and fit a first-order model. Because the levels for some factors were changed from

the first experiment to the second, we switch to uncoded units for all eight factors. Fitting a model with linear main effects, plus a nominal variable to account for a possible block effect between the two experiments, the results are alarming; Table 6.9 contains the analysis of variance for this model. Instead of main effects explaining nearly all of the variation, most of the variation is in lack-of-fit. A first-order model is of no use for describing how these factors affect lovastatin production. This lack-of-fit could be explained by adding some interaction terms, but the combined designs are poor for distinguishing interaction effects.

Table 6.9. Analysis of variance for main effects model for combined data

Source	df	SS	%	MS
First-order model	9	0.2546	26.0	0.0283
Lack-of-fit	6	0.7087	72.4	0.1181
Pure error	32	0.0158	1.6	0.0005
Total (corrected)	47	0.9791	100.0	0.1892

This example illustrates the following principles:

- Performing replicate runs at factorial treatment combinations increases precision of effect estimates and provides information about the error variance, but it does not increase the number of effects that can be estimated.
- The success of resolution III designs depends on main effects dominating all other effects. For these lovastatin experiments, main effects were large relative to random error, but interaction effects were too large to ignore.
- It is better to run an unreplicated fractional factorial design of resolution IV than to replicate a resolution III design. Rather than three replicates of a 2^{7-4} , the same number of runs could have been used to perform a design for seven or eight factors that permits estimating all two-factor interactions (see Sections 8.3.2 and 8.3.3).
- If the error variation is small, it may be better to choose narrow spacing of levels for the factors when conducting a resolution III design. In general, the smaller the experimental region, the simpler the model one may use.
- Following a resolution III experiment, it is advisable to check the adequacy of your model by collecting new data at treatment combinations of interest not included in the original experiment. Here, steepest ascent exploration might have been tried after the first experiment. If the new data contradict the model, then either the precision of the fitted model is poor or the model is missing important effects.

Example 6.3: A 2^{13-9} experiment with several responses

Irvine, Clark, and Recupero (1996) carefully documented a 2^{13-9} experiment investigating the best method to remove lignin during the pulping stage without negatively impacting strength and yield. The 13 factors are listed in Table 6.10, together with the levels used. Irvine et al.'s article explains the logic behind the choice of each factor and its levels.

The run order for the 16 treatment combinations of the 2^{13-9} is given in Table 6.11, together with data for three response variables. The Kappa number is proportional to the percent of lignin in the pulp; Kappa = 10 corresponds to about 1.5% lignin. Small Kappa values are preferred, but these tend to coincide with larger Tear index and lower Yield, both of which are undesirable. The design generators for this fraction are $\mathbf{E} = -\mathbf{AD}$, $\mathbf{F} = -\mathbf{ABCD}$, $\mathbf{G} = \mathbf{ABC}$, $\mathbf{H} = \mathbf{BCD}$, $\mathbf{J} = -\mathbf{AC}$, $\mathbf{K} = -\mathbf{BD}$, $\mathbf{L} = \mathbf{ACD}$, $\mathbf{M} = -\mathbf{AB}$, and $\mathbf{N} = -\mathbf{BC}$. The two interactions of the basic columns not used as generators are \mathbf{ABD} and \mathbf{CD} . Although different generators are used, this design is equivalent to the minimum aberration design presented in Appendix G.

To analyze these data, it is preferable to fit a saturated model and use Lenth's PSE to estimate the standard error for the regression coefficients. The alternative is to fit a main effects model and use $\text{RMSE}/16^{1/2}$. However, with only 2 df for error, the RMSE estimate is less precise than Lenth's PSE based on 15 contrasts. The results of fitting a saturated model for the three responses in Table 6.10 are presented in Table 6.12. The authors used normal effects plots to display the estimates. Since the coefficients are predominantly negative, half-normal plots are more effective for displaying the outcome for each model (see Figure 6.3).

Table 6.10. Factors and levels for Irvine et al.'s (1996) pulping experiment

Factors	Levels	
	-1	1
A Wood chips presoaked	No	Yes
B Chips pre-steamed for 10 min @ 110°C	No	Yes
C Initial effective alkali level (%)	6	12
D Sulfide level in impregnation (%)	3	10
E Liquor	Black	White
F Liquor/wood ratio	3.5:1	6:1
G Impregnation temperature (°C)	110	150
H Impregnation pressure (kPa)	190	1140
J Impregnation time (min)	10	40
K Anthraquinone (%)	0.00	0.05
L Cook temperature (°C)	165	170
M Water quench	No	Yes
N Extended alkali wash for 1 hour	No	Yes

Table 6.11. Irvine et al.'s (1996) 2^{13-9} experiment in run order performed

A	B	C	D	E	F	G	H	J	K	L	M	N	Kappa	Tear	Yield
1	1	-1	1	-1	1	-1	-1	1	-1	-1	-1	1	12.1	9.24	40.53
-1	1	1	1	1	1	-1	1	1	-1	-1	1	-1	12.3	8.98	41.01
-1	-1	-1	1	1	1	-1	1	-1	1	1	-1	-1	13.5	9.07	41.64
-1	-1	1	1	1	-1	1	-1	1	1	-1	-1	1	9.6	9.63	40.05
1	-1	-1	1	-1	-1	1	1	1	1	-1	1	-1	12.0	9.05	41.53
-1	1	1	-1	-1	-1	-1	-1	1	1	1	1	-1	13.6	9.24	41.19
-1	-1	1	-1	-1	1	1	1	1	-1	1	-1	1	12.6	9.77	40.54
1	-1	1	1	-1	1	-1	-1	-1	1	1	1	1	10.9	10.07	40.66
1	-1	-1	-1	1	1	1	-1	1	-1	1	1	-1	11.9	9.45	40.16
1	1	1	-1	1	1	1	-1	-1	1	-1	-1	-1	11.6	9.64	40.51
1	1	-1	-1	1	-1	-1	1	1	1	-1	1	-1	10.2	9.37	40.55
1	-1	1	-1	1	-1	-1	1	-1	-1	-1	1	1	11.6	9.86	41.01
-1	1	-1	-1	-1	1	1	1	-1	1	-1	1	1	10.6	9.39	41.10
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	14.8	9.53	41.63
1	1	1	1	-1	-1	1	1	-1	-1	1	-1	-1	13.0	9.69	40.70
-1	1	-1	1	1	-1	1	-1	-1	-1	1	1	1	11.2	9.27	39.71

Table 6.12. Estimated coefficients for Kappa number, tear index, and yield, with Lenth t statistics

Term	Kappa		Tear		Yield	
	Est.	t	Est.	t	Est.	t
Intercept	11.97		9.453		40.783	
A	-0.31	-1.26	0.093	1.12	-0.076	-0.58
B	-0.14	-0.59	-0.101	-1.21	-0.120	-0.91
C	-0.07	-0.28	0.157	1.88	-0.074	-0.56
D	-0.14	-0.59	-0.078	-0.94	-0.054	-0.41
E	-0.48	-1.97	-0.044	-0.53	-0.203	-1.54
F	-0.03	-0.13	-0.002	-0.02	-0.014	-0.10
G	-0.41	-1.67	0.033	0.40	-0.245	-1.87
H	0.01	0.03	-0.056	-0.67	0.227	1.73
J	-0.18	-0.74	-0.112	-1.34	-0.087	-0.67
K	-0.47	-1.92	-0.021	-0.25	0.121	0.92
L	0.14	0.59	0.038	0.46	-0.139	-1.06
M	-0.21	-0.85	-0.039	-0.47	0.014	0.10
N	-0.87	-3.56	0.122	1.46	-0.264	-2.01
AK = BE = CF = ...	-0.02	-0.08	0.007	0.08	-0.015	-0.11
AL = BH = CD = ...	-0.31	-1.26	0.061	0.73	-0.050	-0.38
Std Error						
PSE	0.24		0.083		0.131	
RMSE/ $16^{1/2}$	0.22		0.043		0.037	

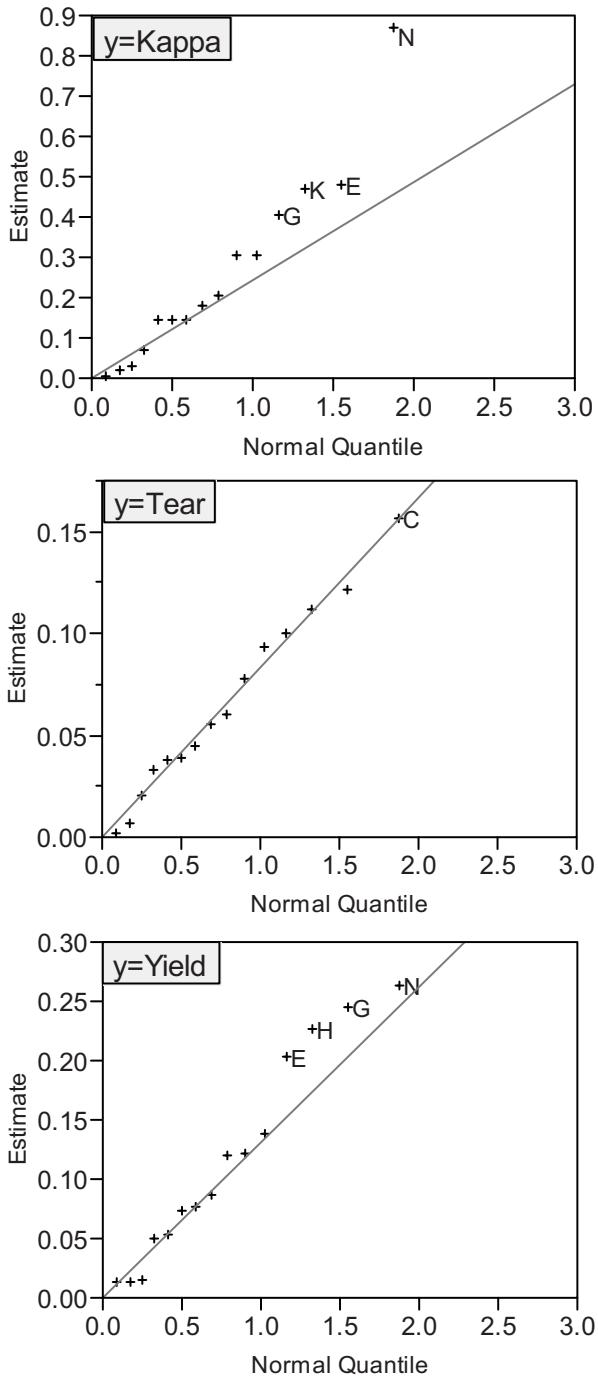


Fig. 6.3. Half-normal plot of effects for $Y = \text{Kappa}$ number, tear index, and yield

The differences between Lenth's PSE and the standard error RMSE/ $16^{1/2}$ reflect the inherent variation in estimates based on so little information. The PSE based on 15 contrasts is more precise than the RMSE with 2 df, so we base our conclusions on Lenth t statistics and critical values from Appendix C. For tear, the largest of 15 t statistics is only 1.88, and the half-normal plot matches what one would expect if all factors were inactive. For Kappa, the Extended alkali wash ($\mathbf{N} = +1$) certainly lowers the response. There is some evidence that the high levels of \mathbf{E} , \mathbf{K} , and \mathbf{G} also decrease Kappa. High \mathbf{N} and \mathbf{G} also seem to have the disadvantage of lowering yield. The coefficient for yield $b_H = 0.227$ is also statistically significant at $\alpha = .10$. (In Section 14.2, we consider an alternative to Lenth's method, which is slightly more powerful when there are only one or two active effects.)

In summary, high \mathbf{H} appears to increase yield without raising Kappa, and high \mathbf{K} appears to lower Kappa without lowering yield. Because these estimates are only marginally significant, subsequent experimentation is needed to confirm the presence of these effects. One option would be a foldover of this design, which would increase the precision and remove the aliasing of main effects with two-factor interactions (see Section 9.4). If performing 16 more runs is impractical, at least confirmation runs with $\mathbf{H} = \mathbf{K} = +1$ should be explored.

Example 6.4: A 2^{15-11} experiment

Poorna and Kulkarni (1995) investigated four carbon sources and eight nitrogen sources, together with three other parameters thought to affect inulinase production. The 15 factors and their levels are given in Table 6.13. A regular 2^{15-11} fractional factorial design was used that included the treatment combination with all 16 factors at the low level. The 11 generators were $\mathbf{E} = -\mathbf{ABCD}$, $\mathbf{F} = \mathbf{BCD}$, $\mathbf{G} = \mathbf{ABC}$, $\mathbf{H} = -\mathbf{CD}$, $\mathbf{J} = -\mathbf{BD}$, $\mathbf{K} = \mathbf{ABD}$, $\mathbf{L} = \mathbf{ACD}$, $\mathbf{M} = -\mathbf{AC}$, $\mathbf{N} = -\mathbf{AD}$, $\mathbf{O} = -\mathbf{AB}$, and $\mathbf{P} = -\mathbf{BC}$, producing the fraction given in Table 6.14. The order shown presumes that the authors' trial code indicates actual run order. The two responses are inulinase activity (units/mL) at 60 hours (denoted y_A) and dry weight biomass (mg/mL) after 96 hours of incubation (denoted y_B). Note that the final treatment combination, being at the low level for all ingredients, produced nothing measurable even after 96 hours. The scatterplot in Figure 6.4 shows how different this run is.

Table 6.13. Factors and levels for Poorna and Kulkarni's experiment

Factors	Levels	
	-1	1
Carbon source		
A: Inulin (%)	0	2
B: Fructose (%)	0	2
C: Glucose (%)	0	2
D: Sucrose (%)	0	2
Organic nitrogen source		
E: Corn steep liquor (%)	0	1
F: Peptone (%)	0	1
G: Urea (%)	0	1
H: Yeast extract (%)	0	1
Inorganic nitrogen source		
J: Corn steep liquor (%)	0	1
K: Peptone (%)	0	1
L: Urea (%)	0	1
M: Yeast extract (%)	0	1
Other		
N: Trace elements solution (mL)	0.5	1.5
O: Inoculum level (10^6 spores/mL)	0.25	25
P: pH	5	6

Table 6.14. Poorna and Kulkarni's 2^{15-11} experiment

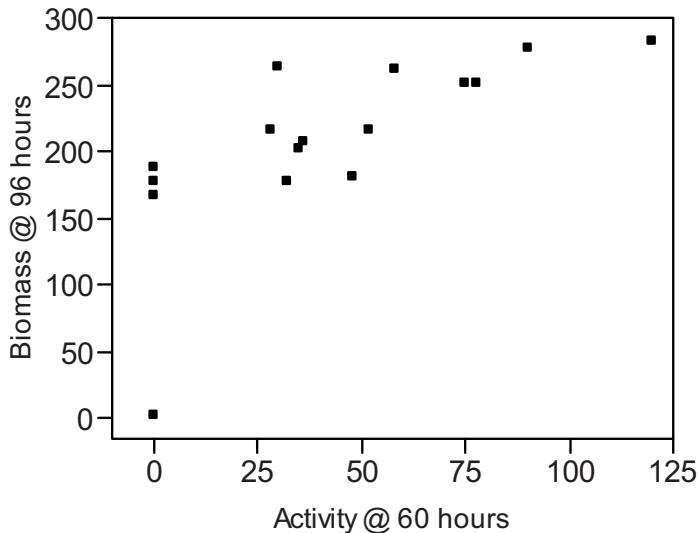


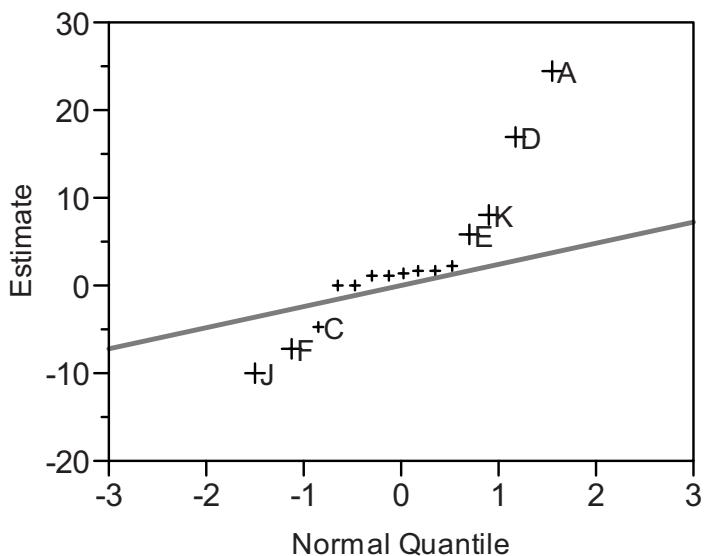
Fig. 6.4. Biomass weight y_B versus Activity y_A for inulinase experiment

The main effects model (1.1) is a saturated model here. With no replication, Lenth's PSE is used to estimate the standard error for the regression coefficients. The results of this model fit for y_A are given in Table 6.15 and Figure 6.5. Six of the 15 coefficients have Lenth t statistics $> c_{.05}^{\text{IER}} = 2.156$ and so are statistically significant at $\alpha = .05$. The p -values reported are based on simulation, as described in Appendix C. Of the six statistically significant effects, four are positive and two are negative. The authors expected Inulin to have a large positive effect, while the positive estimate for Sucrose was a surprise. The other two positive estimates represent two nitrogen sources: one organic and one inorganic. Hence, the four ingredients Inulin, Sucrose, Corn steep liquor, and Ammonium sulfate show the most benefit.

In spite of the large positive estimate $b_D = 16.9$, Poorna and Kulkarni were reluctant to accept that adding sucrose increases y_A . They stated that "to some extent sucrose repressed the enzyme production" (p. 319). This comment is based on previous literature and an inspection of individual treatment combinations, in contradiction to the fitted model. Further investigation seems warranted to examine the aliasing of two-factor interactions with these main effects. For Sucrose, $-\mathbf{D} = \mathbf{AN} = \mathbf{BJ} = \mathbf{CH} = \mathbf{EG} = \mathbf{FP} = \mathbf{KO} = \mathbf{LM}$. Since none of these interactions appears likely to account for the large positive estimate for \mathbf{D} and the data show no evidence for an \mathbf{AD} interaction, this experiment tentatively supports the usefulness of Sucrose to inulinase production, whether Inulin is present or not.

Table 6.15. Saturated model for y_A in inulinase experiment

Term	Estimate	PSE	Lenth t	p-Value
Intercept	42.625	2.4375	17.49	<.0001
A (Inulin)	24.500	2.4375	10.05	<.0001
B (Fructose)	0.125	2.4375	0.05	0.9598
C (Glucose)	-4.750	2.4375	-1.95	0.0703
D (Sucrose)	16.875	2.4375	6.92	<.0001
E (CSL)	5.875	2.4375	2.41	0.0292
F (Peptone)	-7.250	2.4375	-2.97	0.0095
G (Urea)	-0.125	2.4375	-0.05	0.9598
H (Yeast extract)	1.000	2.4375	0.41	0.6874
J (NH_4Cl)	-9.875	2.4375	-4.05	0.0010
K ($(\text{NH}_4)_2\text{SO}_4$)	8.000	2.4375	3.28	0.0050
L ($\text{NH}_4\text{H}_2\text{PO}_4$)	1.625	2.4375	0.67	0.5151
M (NaNO_3)	2.125	2.4375	0.87	0.3971
N (Trace elements)	1.250	2.4375	0.51	0.6155
O (Inoculum level)	1.750	2.4375	0.72	0.4838
P (pH)	1.500	2.4375	0.62	0.5475

**Fig. 6.5.** Normal plot of effects for y_A

6.2 Some Theory Regarding Resolution III Designs

Regular resolution III 2^{k-f} fractional factorial designs with $k = N - 1$ are very simple to construct. Define $\mathbf{S}_1 = \mathbf{1}$ and

$$\mathbf{S}_N = \begin{bmatrix} \mathbf{S}_{N/2} & \mathbf{S}_{N/2} \\ \mathbf{S}_{N/2} & -\mathbf{S}_{N/2} \end{bmatrix} \quad (6.1)$$

for $N = 2, 4, 8, \dots$ (i.e., for N any power of 2). Each \mathbf{S}_N in this series is a square matrix of -1 's and $+1$'s with orthogonal columns, which implies

$$\mathbf{S}'_N \mathbf{S}_N = N \mathbf{I}_N.$$

Any matrix with these properties is called a Hadamard matrix; see Hedayat, Sloane and Stufken (1999, Ch. 7) for additional details. Equation (6.1) is the Sylvester-type Hadamard matrix construction, hence the designation \mathbf{S}_N . Other Hadamard matrices are discussed in Section 6.3. It is easy to verify that \mathbf{S}_N defined by (6.1) is a symmetric matrix with first column (and row) being a constant vector of 1 's. Thus, \mathbf{S}_N represents the model matrix for an orthogonal two-level design, where the first column of \mathbf{S}_N is the intercept column and the remaining $N - 1$ columns form the design. For example,

$$\mathbf{S}_4 = \begin{bmatrix} \mathbf{S}_2 & \mathbf{S}_2 \\ \mathbf{S}_2 & -\mathbf{S}_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix},$$

$$\mathbf{S}_8 = \begin{bmatrix} \mathbf{S}_4 & \mathbf{S}_4 \\ \mathbf{S}_4 & -\mathbf{S}_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix}$$

are first-order model matrices for 2^{3-1} and 2^{7-4} designs, respectively. In addition, if we number the columns of \mathbf{S}_N 0, 1, 2, ..., $N-1$, then columns 1, 2, 4, 8, ..., $N/2$ form a full factorial and the remaining columns are interactions of these. In fact, the numbering matches the column numbering in Appendix F, which is used to present the minimum aberration designs in Appendix G. For instance, column 3 is the product of columns 1 and 2, column 5 is the product of columns 1 and 4, etc. Thus, one may construct the minimum aberration fractional factorial designs with $k = N - 1$ using (6.1) or by using the generators in Appendix G.

The saturated main effect designs defined by (6.1) are symmetrical in their aliasing of effects. Consider the 2^{7-4} design with generators $\mathbf{D} = \mathbf{AB}$, $\mathbf{E} = \mathbf{AC}$, $\mathbf{F} = \mathbf{BC}$, and $\mathbf{G} = \mathbf{ABC}$. The defining relation for this design is

$$\begin{aligned} \mathbf{I} = \mathbf{ABD} &= \mathbf{ACE} = \mathbf{BCDE} = \mathbf{BCF} = \mathbf{ACDF} = \mathbf{ABEF} = \mathbf{DEF} = \mathbf{ABCG} \\ &= \mathbf{CDG} = \mathbf{BEG} = \mathbf{ADEG} = \mathbf{AFG} = \mathbf{BDFG} = \mathbf{CEFG} = \mathbf{ABCDEFG}. \end{aligned}$$

This 1/16th fraction aliases 15 interactions with each of the main effects; 3 of these are two-factor interactions, since each factor appears in 3 of the 7 length-3 words of the defining relation. In particular,

$$\begin{aligned} \mathbf{A} &= \mathbf{BD} = \mathbf{CE} = \mathbf{FG} = \dots = \mathbf{BCDEF} \\ \mathbf{B} &= \mathbf{AD} = \mathbf{CF} = \mathbf{EG} = \dots = \mathbf{ACDEF} \\ \mathbf{C} &= \mathbf{AE} = \mathbf{BF} = \mathbf{DG} = \dots = \mathbf{ABDEF} \\ \mathbf{D} &= \mathbf{AB} = \mathbf{CG} = \mathbf{EF} = \dots = \mathbf{ABCEF} \\ \mathbf{E} &= \mathbf{AC} = \mathbf{BG} = \mathbf{DF} = \dots = \mathbf{ABCD} \\ \mathbf{F} &= \mathbf{AG} = \mathbf{BC} = \mathbf{DE} = \dots = \mathbf{ABCDE} \\ \mathbf{G} &= \mathbf{AF} = \mathbf{BE} = \mathbf{CD} = \dots = \mathbf{ABCDEF} \end{aligned}$$

In general, a saturated main effect design in N runs aliases $N/2 - 1$ two-factor interactions with each main effect.

Every eight-run fractional factorial design with orthogonal main effects is equivalent (i.e., isomorphic) to a projection of this seven-factor design; that is, for six factors in eight runs, we simply drop one of the seven factors. Because of the symmetry of the aliasing in the seven-factor design, it does not matter which factor we drop. Dropping factor **G**, the defining relation for the resulting 2^{6-2} design is

$$\mathbf{I} = \mathbf{ABD} = \mathbf{ACE} = \mathbf{BCDE} = \mathbf{BCF} = \mathbf{ACDF} = \mathbf{ABEF} = \mathbf{DEF}$$

and the aliasing reduces to

$$\begin{aligned} \mathbf{A} &= \mathbf{BD} = \mathbf{CE} = \dots \\ \mathbf{B} &= \mathbf{AD} = \mathbf{CF} = \dots \\ \mathbf{C} &= \mathbf{AE} = \mathbf{BF} = \dots \\ \mathbf{D} &= \mathbf{AB} = \mathbf{EF} = \dots \\ \mathbf{E} &= \mathbf{AC} = \mathbf{DF} = \dots \\ \mathbf{F} &= \mathbf{BC} = \mathbf{DE} = \dots \\ \mathbf{AF} &= \mathbf{BE} = \mathbf{CD} = \dots = \mathbf{ABCDEF} \end{aligned}$$

For five factors, it does not matter which factor is dropped from the six-factor design; the resulting defining relation will still contain two length-3 words and one length-4 word. Dropping **F** one obtains the 2^{5-2} with defining relation

$$\mathbf{I} = \mathbf{ABD} = \mathbf{ACE} = \mathbf{BCDE},$$

and the aliasing reduces to

$$\begin{aligned} \mathbf{A} &= \mathbf{BD} = \mathbf{CE} = \dots \\ \mathbf{B} &= \mathbf{AD} = \dots \\ \mathbf{C} &= \mathbf{AE} = \dots \\ \mathbf{D} &= \mathbf{AB} = \dots \\ \mathbf{E} &= \mathbf{AC} = \dots \\ \mathbf{BC} &= \mathbf{DE} = \dots \\ \mathbf{BE} &= \mathbf{CD} = \dots \end{aligned}$$

Since the aliasing for the 2^{6-3} and 2^{5-2} designs is not symmetric, it does matter how the factors are assigned to the columns. If a particular two-factor interaction is considered more likely, one should avoid aliasing it with any main effects.

Consider now dropping a third column to obtain a 2^{4-1} fraction. If one drops **A**, the resulting design is resolution IV, and if one drops any other column, the design remains resolution III. By dropping yet another column, one obtains either the full 2^3 or a replicated 2^{3-1} .

These projections of the 2^{7-4} design illustrate four general results regarding regular fractional factorial designs:

- The regular resolution III 2^{k-f} design with $k = N - 1$ is unique (in the sense of isomorphism defined in Section 5.2.4). All regular N -run designs of resolution III or higher are projections of this design.
- For $k = N - 2$ and $k = N - 3$, it does not matter which columns are deleted from the saturated main effect design. Equivalent designs are obtained.
- For $k \leq N - 4$, it matters which columns are dropped and which are retained as factors, as different designs are possible.
- For $k \leq N/2$, designs of resolution IV or higher are available.

6.2.1 Criterion for ranking regular 2^{k-f} designs of resolution III

For $k > N/2$, the maximum resolution is resolution III. In cases where non-isomorphic designs exist, we prefer the design with the fewest length-3 words. For instance, with $k = 9$ and $N = 16$, there are five non-isomorphic regular resolution III designs. The five designs have defining relations with word length patterns $wlp = (A_3, \dots, A_9)$ as follows:

Design	Word Length Pattern
9-5.1	(4, 14, 8, 0, 4, 1, 0)
9-5.2	(6, 9, 9, 6, 0, 0, 1)
9-5.3	(6, 10, 8, 4, 2, 1, 0)
9-5.4	(7, 9, 6, 6, 3, 0, 0)
9-5.5	(8, 10, 4, 4, 4, 1, 0)

One way to construct the minimum aberration design is to use columns 7 and 11–14 as generators. Assigning the letters **A–H** and **J** as factor labels, our generators are **E** = **ABC**, **F** = **ABD**, **G** = **CD**, **H** = **ACD**, and **J** = **BCD**. (As detailed in Appendix F, for a 16-run design, the first four factors **A–D** correspond to the basic column numbers {1, 2, 4, 8}, and column 7 (which equals $1 + 2 + 4$) corresponds to the interaction **ABC**, column 11 corresponds to the interaction **ABD**, etc. The aliasing, up to two-factor interactions, for this design is

$$\begin{aligned}
\mathbf{A} &= \mathbf{GH} \\
\mathbf{B} &= \mathbf{GJ} \\
\mathbf{C} &= \mathbf{DG} \\
\mathbf{D} &= \mathbf{CG} \\
\mathbf{E} &= \mathbf{FG} \\
\mathbf{F} &= \mathbf{EG} \\
\mathbf{G} &= \mathbf{AH} = \mathbf{BJ} = \mathbf{CD} = \mathbf{EF} \\
\mathbf{H} &= \mathbf{AG} \\
\mathbf{J} &= \mathbf{BG} \\
\mathbf{AB} &= \mathbf{CE} = \mathbf{DF} = \mathbf{HJ} \\
\mathbf{AC} &= \mathbf{BE} = \mathbf{DH} = \mathbf{FJ} \\
\mathbf{AD} &= \mathbf{BF} = \mathbf{CH} = \mathbf{EJ} \\
\mathbf{AE} &= \mathbf{BC} = \mathbf{DJ} = \mathbf{FH} \\
\mathbf{AF} &= \mathbf{BD} = \mathbf{CJ} = \mathbf{EH} \\
\mathbf{AJ} &= \mathbf{BH} = \mathbf{CF} = \mathbf{DE}
\end{aligned}$$

The four length-3 words in the defining relation are **AGH**, **BGJ**, **CDG**, and **EFG**. Each length-3 word produces three aliases between main effects and two-factor interactions. Thus, with $A_3 = 4$, the design aliases 12 two-factor interactions with main effects. The other resolution III designs will alias 18 or more two-factor interactions with the 9 main effects, since these designs have $A_3 \geq 6$. Because the number of two-factor interactions aliased with main effects is proportional to A_3 , minimizing the number of length-3 words is the primary criterion for ranking resolution III designs. Appendix G provides minimum aberration resolution III designs of size $N = 8, 16, 32, 64$, and 128; $N/2 < k < N$. For $N = 16$ and 32, these designs appeared in Chen, Sun, and Wu (1993). For larger N , they were obtained using the following complementary design construction.

6.2.2 Constructing resolution III designs via complementary design

The saturated resolution III 2^{k-f} design uses all f interactions among the basic factors as generators. As mentioned earlier, regular resolution III designs with $k \geq N - 3$ are isomorphic; that is, for $k = N - 3$, it does not matter which two interactions are not used as generators, since the resulting designs are equivalent and can be made identical by swapping of columns and rows and, if necessary, reversing the levels of some columns. The same is true for $k = N - 2$. When $N - k$ is much smaller than f , it is easier to specify the $N - k - 1$ columns not used as generators than it is to specify the f generators. Tang and Wu (1996) showed how to search for the best resolution III designs by considering the set of omitted columns. We illustrate the connection for the case $N = 16$ and $k = 11$, where there are three nonisomorphic designs. From Chen, Sun, and Wu (1993), the three designs can be constructed by omitting the following sets of $N - 1 - k = 4$ columns (see Appendix F for identification of interactions by column numbers):

- Design 11-7.1 ($A_3 = 12, A_4 = 26$) by omitting columns $\{7, 11, 12, 15\}$. These four omitted columns form a replicated 2^{4-1} of resolution III, since $\mathbf{7} = \mathbf{11}\cdot\mathbf{12}$.
- Design 11-7.2 ($A_3 = 13, A_4 = 25$) by omitting columns $\{11, 13, 14, 15\}$. These four omitted columns form a full 2^4 .
- Design 11-7.3 ($A_3 = 13, A_4 = 26$) by omitting columns $\{12, 13, 14, 15\}$. These four omitted columns form a replicated 2^{4-1} of resolution IV, since $\mathbf{15} = \mathbf{12}\cdot\mathbf{13}\cdot\mathbf{14}$.

The design formed by the omitted columns is labeled the *complementary design*. Let \bar{A}_j denote the number of length- j words for the complementary design with $N - k - 1$ factors, just as A_j denotes the number of length- j words for the 2^{k-f} design. Chen and Hedayat (1996) and Tang and Wu (1996) showed that for a given N and k ,

$$\begin{aligned} A_3 &= C_3(N, k) - \bar{A}_3, \\ A_4 &= C_4(N, k) + \bar{A}_3 + \bar{A}_4, \end{aligned}$$

where $C_3(N, k)$ and $C_4(N, k)$ are constants that depend only on N and k . In particular, $C_3(16, 11) = 13$ and $C_4(16, 11) = 25$, as can be verified for each of the 2^{11-7} designs and their complement above. Tang and Wu further showed that the minimum aberration 2^{k-f} design may be found by sequentially maximizing \bar{A}_3 , minimizing \bar{A}_4 , maximizing \bar{A}_5 , etc. Thus, for the minimum aberration 2^{11-7} design, one omits 4 columns which form a (replicated) resolution III 16-run fraction. In general, the minimum aberration design is obtained by deleting $N - k - 1$ columns which together have the maximum aliasing of two-factor interactions with main effects.

In some instances, there exist several resolution III designs with the same A_3 as the minimum aberration design. Such designs are said to have weak minimum aberration (Chen and Hedayat 1996). These designs alias the same number of two-factor interactions with main effects as the minimum aberration design.

6.3 Nonregular Orthogonal Designs of Strength 2

All fractional factorial designs presented in Chapter 5 and in Sections 6.1 and 6.2 can be constructed using design generators that define f additional columns in terms of interactions of the $k - f$ basic columns. This method produces regular $(1/2)^f$ fractions of a 2^k factorial. For these fractions, each factorial effect column is indistinguishable from its aliases but orthogonal to all other main effect and interaction columns.

Now we consider orthogonal main effect designs that do not have defining relations. These nonregular designs are constructed without design generators and have different aliasing and projection properties. Initially, two such designs are shown to highlight the differences between nonregular and regular

fractional factorial designs. Properties such as aliasing, resolution, word length pattern, and aberration are generalized to provide a means of characterizing nonregular orthogonal designs. Then, in the following subsections, we present the recommended nonregular designs of sizes 12, 16, 20, 24, and larger and reanalyze four published examples.

Tables 6.16 and 6.17 show nonregular fractional factorial designs. Each design is a strength-2 orthogonal array, which means every pair of columns

Table 6.16. OA(12, 2¹¹, 2) design

1	2	3	4	5	6	7	8	9	10	11
1	1	1	1	1	1	1	1	1	1	1
-1	1	-1	1	1	1	-1	-1	-1	1	-1
-1	-1	1	-1	1	1	1	-1	-1	-1	1
1	-1	-1	1	-1	1	1	1	-1	-1	-1
-1	1	-1	-1	1	-1	1	1	1	-1	-1
-1	-1	1	-1	-1	1	-1	1	1	1	-1
-1	-1	-1	1	-1	-1	1	-1	1	1	1
1	-1	-1	-1	1	-1	-1	1	-1	1	1
1	1	-1	-1	-1	1	-1	-1	1	-1	1
1	1	1	-1	-1	-1	1	-1	-1	1	-1
-1	1	1	1	-1	-1	-1	1	-1	-1	1
1	-1	1	1	1	-1	-1	-1	1	-1	-1

Table 6.17. OA(16, 2¹⁵, 2) design (Hall Type V)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
-1	1	-1	1	-1	1	-1	1	1	-1	-1	-1	-1	1	1
1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1
-1	-1	1	1	-1	-1	1	1	-1	1	-1	-1	1	-1	1
1	1	1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	1	1	-1	-1	1	-1	1	1	-1
1	-1	-1	-1	1	1	1	-1	1	-1	1	-1	1	-1	-1
-1	-1	-1	-1	1	1	1	-1	1	-1	1	-1	1	-1	-1
-1	-1	1	-1	1	1	-1	1	-1	1	1	-1	-1	1	1
1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	1	-1	1	-1	1	-1	-1	-1	1	1	1	1	-1	-1
1	-1	-1	1	1	-1	-1	-1	1	1	-1	-1	1	1	1
-1	-1	1	1	-1	-1	1	-1	1	1	1	-1	-1	1	-1
1	1	1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1
-1	1	-1	-1	1	-1	1	-1	1	-1	1	-1	-1	1	1
1	-1	-1	-1	1	1	-1	1	-1	1	-1	1	-1	1	1
-1	-1	1	1	-1	-1	-1	1	1	-1	-1	1	-1	1	-1

forms an equally replicated 2^2 factorial. In general, a *strength-t* orthogonal array projects into an equally replicated full factorial in every subset of t factors. We denote an N -run, strength- t orthogonal array with k two-level factors by $\text{OA}(N, 2^k, t)$. An $\text{OA}(N, 2^k, t)$ requires N to be divisible by 2^t . While regular unreplicated 2^{k-f} designs are restricted to N a power of 2, strength two nonregular designs are more flexible in size.

6.3.1 Properties of strength-2 orthogonal arrays

We now discuss five properties of orthogonal arrays: (i) bias from omitted interactions, (ii) projection properties, (iii) estimability, (iv) efficiency of estimation, and (v) existence and construction.

(i) Bias due to omission of active interactions

Although aliasing for regular designs is apparent from the defining relation, here one must use the alias matrix (see Appendix I). The alias matrix reveals the bias to least squares estimators from a model that omits active terms. For strength-2 orthogonal arrays, it is relevant to consider the impact of omitted two-factor interactions on estimates for a first-order model. The potential bias to each main effect estimate from omitted two-factor interactions is characterized by rows of this alias matrix. In particular,

$$E(b_r) = \beta_r + \sum_{i=1}^{k-1} \sum_{j=i+1}^k \rho_{r,i,j} \beta_{i,j},$$

where $\rho_{r,i,j}$ is the correlation between x_r and $x_i x_j$. For the design in Table 6.16, $|\rho_{r,i,j}| = 1/3$ for all $r \neq i, j$, something that does not arise for regular designs. For the design in Table 6.17, these correlations are 0, $1/2$, and 1. When the correlation is not 0 or ± 1 , we say that the effects are partially aliased. Obviously, the stronger the correlation, the greater the potential bias to the main effect.

(ii) Projections

Regular 2^{k-f} fractional factorials with resolution R project into an equally replicated full factorial in any set of $R - 1$ factors. Hence, $\text{OA}(N, 2^k, t)$ designs include the regular 2^{k-f} fractions of resolution $t + 1$. Whereas $\text{OA}(N, 2^k, 2)$ and regular resolution III designs appear similar when projecting into just two columns, projections into three or more columns will highlight their differences. Every nonregular strength-2 design in this section projects into an unequally replicated 2^3 in some sets of three columns. By contrast, a regular resolution III 2^{k-f} design projects either into an equally replicated 2^{3-1} or an equally replicated 2^3 , depending on whether the three factors appear together

in a length-3 word of the defining relation or not. From Table 6.16, one may verify that every set of three columns projects into a full 2^3 , with each row appearing once or twice. This is ideal for 12 runs, since the frequencies are as nearly equal as possible. From Table 6.17, three projections occur:

- Four replicates of a 2^{3-1} ; e.g., columns **1, 2, 3**.
- Two replicates of a 2^3 ; e.g., columns **1, 2, 4**.
- One replicate of a 2^3 , plus two replicates of a 2^{3-1} ; e.g., columns **1, 2, 8**.

Hence, the **1·2** interaction column is completely aliased with the factor **3** main effect and partially aliased with the factor **8** main effect.

(iii) Estimability

Exploring all possible projections of a design reveals what models may be fit in subsets of the factors. Since the design in Table 6.16 projects into a full 2^3 in every subset of three factors, one can estimate the full factorial model (1.4) in every subset of three factors. One can also estimate the two-factor interaction model (1.3) in every subset of four factors for this design. Table 6.16 is our first example of a general result due to Cheng (1995): that every $\text{OA}(N, 2^k, 2)$ with $k \geq 4$ and N not a multiple of 8 supports estimation of the two-factor interaction model for every set of four factors. Bulutoglu and Cheng (2003) extended this result to apply to some $\text{OA}(N, 2^k, 2)$ with N a multiple of 8, but this does not include the design in Table 6.17.

Qu (2006) proposed the *maxest* criterion for ranking designs based on sequentially maximizing components of an estimability vector

$$\text{EV} = (e_{11}, e_{12}, e_{22}, e_{13}, e_{23}, e_{33}, e_{14}, \dots),$$

where e_{1s} (e_{rs}) is the proportion of main effects (r -factor interactions) that are estimable in a hierarchical model containing all s -factor interactions. For regular 2^{k-f} designs, resolution III guarantees $e_{11} = 1$, and, additionally, resolution IV guarantees $e_{12} = 1$ and resolution V guarantees $e_{22} = e_{13} = 1$.

For all $\text{OA}(N, 2^k, 2)$ satisfying the conditions of Cheng's (1995) and Bulutoglu and Cheng's (2003) theorems, $e_{11} = e_{12} = e_{22} = 1$ for all four-factor projections. Note, however, that estimability criteria do not take into account the model matrix's lack of orthogonality; precision of estimates must be addressed with an additional criterion.

(iv) Efficiency

For a regular design, the orthogonality of columns from different alias sets implies that any model that can be fit will be estimated with full precision; that is, if the model matrix \mathbf{X} for any regular design is not singular, then $\mathbf{X}'\mathbf{X}$ will be a diagonal matrix. By contrast, although nonregular designs permit estimation of more models, these estimable models do not necessarily

have diagonal $\mathbf{X}'\mathbf{X}$. Thus, for nonregular designs, one is interested in both the proportion of models with interactions that are estimable for a particular OA($N, 2^k, 2$) as well as the precision of estimated coefficients. Efficiency may be measured by variance inflation factors or by A- and D-efficiency (all of which are described at the beginning of Section 6.4).

(v) Existence and construction of OA($N, 2^k, 2$)

The designs in Tables 6.16 and 6.17 are called Hadamard designs, because if one adds a column of +1's, the resulting square matrices H_{12} and H_{16} are Hadamard matrices satisfying the property $\mathbf{H}'_N \mathbf{H}_N = N\mathbf{I}_N$. A common conjecture is that Hadamard matrices \mathbf{H}_N exist for every N that is a multiple of 4. Plackett and Burman (1946) listed a Hadamard design for every order up to 100, except for $N = 92$ (a case that was not solved until 1962). The website <http://www.research.att.com/~njas/hadamard/> contains Hadamard matrices \mathbf{H}_{4u} for every case up to \mathbf{H}_{256} . For an even larger collection, see <http://www.math.ntua.gr/people/ckoukouv/>. Hedayat, Sloane, and Stufken (1999, Ch. 7) provide an excellent summary of how these matrices are constructed; see also Seberry, Wysocki, and Wysocki (2005). Many people refer to these as Plackett and Burman designs. However, this book prefers the term *Hadamard design*, since we include designs such as Table 6.17 not considered by Plackett and Burman.

OA($N, 2^k, 2$) with $k < N - 1$ may be constructed in two ways. First, one may search to find the best k -factor projection of known Hadamard designs. Second, one may use algorithms to search for all orthogonal arrays of a given strength and size, sequentially adding additional columns. For $k > N/2$, the best orthogonal designs appear to be projections of Hadamard designs. However, this is not the case for smaller k . For example, Xu and Deng's (2005, p. 130) design 20.7.1 is an OA(20, 2⁷, 2) not obtainable as a 7-factor projection from any 20-run Hadamard design. Those who work to construct these orthogonal arrays must use either rigorous checks for isomorphism (see Clark and Dean 2001) or some heuristic for distinguishing designs that is quicker but not guaranteed to correctly distinguish all designs (Katsaounis and Dean 2008).

6.3.2 Criteria for ranking nonregular orthogonal designs

Just as word length pattern is used to rank regular fractional factorial designs with the same resolution, we need criteria to distinguish OA($N, 2^k, t$) of the same strength. Deng and Tang (1999) introduced the concepts of generalized word length pattern and generalized resolution. Following their notation, let $s = \{d_{j_1}, \dots, d_{j_r}\}$ denote a subset of r of the k factors for a two-level design \mathbf{D} and define

$$J_r(s) = J_r(d_{j_1}, \dots, d_{j_r}) = \left| \sum_{i=1}^N d_{ij_1} \cdots d_{ij_r} \right|, \quad (6.2)$$

where d_j denotes the j^{th} column of \mathbf{D} , with elements d_{ij} for $i = 1, \dots, N$.

For regular fractional factorial designs, $J_r(s)$ takes on the values 0 and N , and A_r equals the number of subsets of size r for which $J_r(s) = N$. Deng and Tang (1999) showed that for any OA($N, 2^k, t$) of strength $t \geq 2$, $J_r(s)$ is a multiple of 4. Tang and Deng (1999) defined the normalized J-characteristics by dividing by N . Hence, while $J_r(s)$ is restricted to the values 0, 4, 8, ..., N , the normalized J-characteristics values are 0, $4/N$, $8/N$, ..., 1.

Deng and Tang (1999) defined the *confounding frequency vector*

$$\text{cfv} = [F_3, \dots, F_k] = [(f_{31}, \dots, f_{3u})_3, \dots, (f_{k1}, \dots, f_{ku})_k], \quad (6.3)$$

where $u = N/4$ and f_{rj} is the frequency of r column subsets that give $J_r(s) = 4(u + 1 - j)$. For instance, the confounding frequency vector for the design in Table 6.16 is

$$\text{cfv} = [(0, 0, 165)_3, (0, 0, 330)_4, (0, 66, 0)_5, \dots, (1, 0, 0)_{11}]. \quad (6.4)$$

This cfv reveals that all $\binom{11}{3} = 165$ subsets of 3 columns and all $\binom{11}{4} = 330$ subsets of 4 columns have normalized J-characteristics of $4/12 = 0.3$. Of the $\binom{11}{5} = 462$ subsets of 5 columns, 66 have normalized J-characteristic of $8/12 = 0.6$, while the remaining 396 have J-characteristic of 0, and so do not appear in the cfv. Finally, the product of all 11 columns sums to N , and so has normalized J-characteristic of 1.

The elements of the cfv are sorted from high to low J-characteristics within each subset size. Hence, for regular fractional factorial designs, $f_{r1} = A_r$, since it is the number of size- r subsets with $J_r(s) = N$. In general, at least half of the elements of each F_r are zero. To emphasize how the cfv is an extension of the word length pattern and to abbreviate cfv and its subvectors F_r , we omit zero frequencies and rewrite (6.4) as

$$[A_3(1/3) = 165, A_4(1/3) = 330, A_5(2/3) = 66, \dots, A_{11}(1) = 1]. \quad (6.5)$$

Thus, in general, $A_r(\rho)$ will denote the number of subsets of size r for which $J_r(s) = \rho N$ and $A_r(\rho_1, \rho_2, \dots)$ will denote a vector of such frequencies.

For regular designs, the resolution of the design is the length of the shortest word in the defining relation; that is, the resolution equals the smallest r for which $A_r(1) > 0$. Deng and Tang (1999) defined generalized resolution for orthogonal arrays in terms of the first non-zero element of the cfv. Suppose $A_r(\rho)$ is the first nonzero element of the cfv. Then the generalized resolution is

$$R = r + (1 - \rho). \quad (6.6)$$

For the Table 6.16 design, $R = 3 + (1 - 1/3) = 3.6$.

One useful criterion for ranking orthogonal arrays is minimum generalized aberration (minimum G-aberration), which ranks designs based on the confounding frequency vector, just as minimum aberration ranks designs based

on the word length pattern. For instance, consider the design in Table 6.17. The confounding frequency vector for this OA(16, 2¹⁵, 2) is

$$\text{cfv} = [(7, 0, 112, 0)_3, (21, 0, 336, 0)_4, (0, 0, 672, 0)_5, \dots, (1, 0, 0, 0)_15], \quad (6.7)$$

or in terms of our abbreviated vector of nonzero elements,

$$[A_3(1, 0.5) = (7, 112), A_4(1, 0.5) = (21, 336), A_5(0.5) = 672, \dots, A_{15}(1) = 1].$$

Although the regular 2¹⁵⁻¹¹, with $A_3(1) = 35$, and the Table 6.17 design both have generalized resolution of 3, the nonregular design has less G-aberration, since the first nonzero element of the confounding frequency vector is $A_3(1) = 7 < 35$.

A second criterion proposed in Tang and Deng (1999) is minimum G₂-aberration, which converts each vector $(f_{r1}, \dots, f_{ru})_r$ from (6.3) into the scalar

$$B_r = f_{r1} + (1 - 4/N)^2 f_{r2} + \dots + (4/N)^2 f_{ru}. \quad (6.8)$$

Tang and Deng proposed the vector (B_3, B_4, \dots, B_k) as a “generalization of word length pattern,” since it is directly comparable to the word length pattern (A_3, A_4, \dots, A_k) for regular fractions. Thus, we refer to the vector of B_r values as the generalized word length pattern (gwlp). See Ma and Fang (2001) for an alternative development of the same vector. The design in Table 6.17 with cfv (6.7) has generalized word length pattern $(B_3, B_4, \dots, B_{15})$ equal to

$$(7 + (0.5)^2 112, 21 + (0.5)^2 336, (0.5)^2 672, 1) = (35, 105, 168, \dots, 1),$$

which is identical to the word length pattern for the regular 2¹⁵⁻¹¹. That this is necessarily the case is proven by the following argument.

Since $\mathbf{S_N S'_N} = \mathbf{H_N H'_N} = N\mathbf{I}_N$, both regular and nonregular Hadamard designs with $k = N - 1$ have identical row coincidence matrices \mathbf{DD}' , with diagonal elements $N - 1$ and off-diagonal elements of -1 . Since gwlp can be computed from \mathbf{DD}' , identical \mathbf{DD}' implies identical gwlp. Thus, when N is a power of 2, the gwlp of any OA($N, 2^{N-1}, 2$) design necessarily equals the wlp of the regular saturated fractional factorial design of the same size. Thus, saturated nonregular OAs can be preferred to the regular saturated 2 ^{$k-f$} in terms of generalized aberration and sometimes with respect to generalized resolution—both of which are based on the confounding frequency vector (6.3). However, G_2 -aberration will not distinguish the designs. Still, Tang and Deng (1999) justified using G_2 -aberration to rank designs, arguing that gwlp corresponds to the expected bias from omitted higher-order terms. Using Butler's (2003b) formula relating the r^{th} moment of $\mathbf{T} = \mathbf{DD}'$ to B_r , we have that when $k = N - 1$, every orthogonal design has $B_3 = k(k - 1)/6$, $B_4 = (k - 3)B_3/4$, $B_5 = (k - 7)B_4/5$, and $B_6 = (k - 5)B_5/6$.

In the next five subsections we focus on nonregular, strength-2 designs of sizes 12, 16, 20, 24, and larger. In each subsection we discuss the projection

properties of the designs as well as the potential bias to main effect estimates from omitted two-factor interactions. Since the designs of size 24 and smaller are the most commonly used, we give them particular attention and highlight their differences. We illustrate the successful use and analysis of such designs via Examples 6.5–6.8.

6.3.3 The 12-run design with generalized resolution $3\bar{6}$

All $\text{OA}(12, 2^{11}, 2)$ are isomorphic to the design given earlier in Table 6.16, with cfv (6.5). Note that the largest correlation between a main effect column and a two- or three-factor interaction (and between two two-factor interaction columns) is 1/3.

The first three columns of Table 6.16 project to

1	2	3	Frequency
-1	-1	-1	2
-1	-1	1	1
-1	1	-1	1
-1	1	1	2
1	-1	1	2
1	-1	-1	1
1	1	1	1
1	1	-1	2

The frequency is 1 when $\mathbf{123} = +1$, and 2 when $\mathbf{123} = -1$. Lin and Draper (1992) enumerated all possible projections of the 12-run design into three, four, or five columns. They showed that:

- Every three columns project to a 2^3 plus a resolution III 2^{3-1} .
- For every set of 4 columns, the 12-run design has 11 distinct treatment combinations.

Since N is not a multiple of 8, Cheng (1995) guaranteed that one can estimate the two-factor interaction model in any subset of four factors. This projection property is quite useful, assuming that no more than four factors are active and that one can identify the relevant subset.

For $7 \leq k \leq 10$ factors, it does not matter which of the columns of the 12-run design are used. For a comparison of the choices when $k = 5$ or 6, see Miller and Sitter (2004).

We now analyze two examples: one saturated with factors and the second with $k = 7$. In each case, we consider the possibility of active two-factor interactions, in addition to main effects.

Example 6.5: Eleven factors in 12 runs

Bullington, Lovin, Miller, and Woodall (1993) reported an 11-factor experiment in 12 runs conducted to identify causes for early failures in thermostats manufactured by the Eaton Corporation. A team narrowed down a list of over

50 potential factors to the 11 factors in Table 6.18 and chose extreme levels for each factor to magnify their effects. This choice seemed reasonable because historical evidence indicated that effects were monotonic.

Table 6.18. Factors and levels for Bullington et al. thermostat experiment

Factors	Levels	
	-1	1
A Diaphragm plating rinse	Clean	Dirty
B Current density (min @ amps)	5 @ 60	10 @ 15
C Sulfuric acid cleaning (seconds)	3	30
D Diaphragm electroclean (min)	2	12
E Beryllium copper grain size (in)	0.008	0.018
F Stress orientation (to seam weld)	Perpendicular	Parallel
G Diaphragm condition after brazing	Wet	Air-dried
H Heat treatment (hours @ 600°F)	0.75	4
J Brazing machine water and flux	None	Extra
K Power element electroclean time	Short	Long
L Power element plating rinse	Clean	Dirty

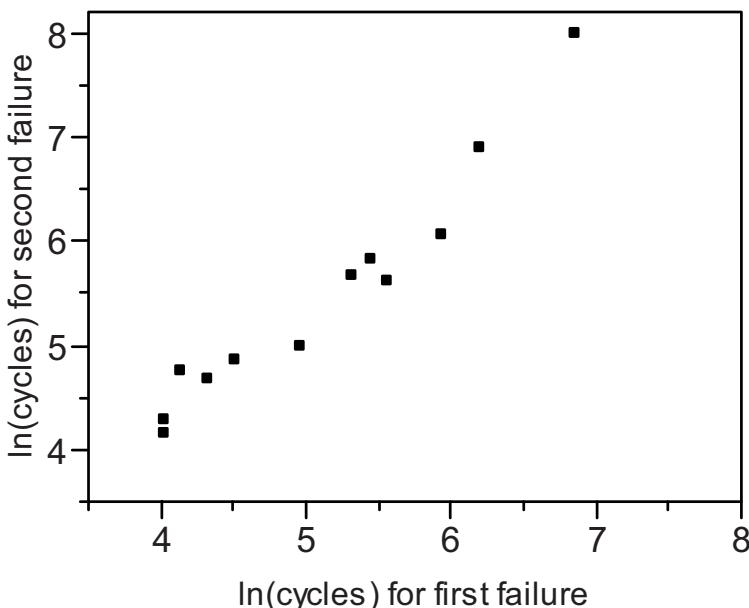
Ten thermostats were manufactured corresponding to each of the 12 treatment combinations in Table 6.19. These 120 thermostats were then tested for 7342K cycles, or until failure, whichever came first. Twenty-two of the 120 thermostats did not fail, and so their responses are right censored. Bullington et al. (1993) took as their response the average $\log(\text{cycles to failure})$ —or, for treatment combinations with censored values, the maximum likelihood estimate for the mean. To simplify the analysis here, we use only the first two failures from each set of 10 thermostats, since 2 treatment combinations only had 2 failures. These early failure data are provided in Table 6.19, where $y_{(1)}$ and $y_{(2)}$ denote the number of cycles (in thousands) corresponding to the first and second failures, respectively.

Due to severe skewness in $y_{(1)}$ and $y_{(2)}$, some transformation is needed. We elect to analyze the \log of $y_{(1)}$ and $y_{(2)}$. This transformation is consistent with Bullington et al.'s assumption of a lognormal distribution for the failure distribution. Figure 6.6 reveals the high correlation (.97) between the transformed first and second failure times, so we expect similar results for the two models.

Fitting a saturated model for $\ln[y_{(1)}]$ and $\ln[y_{(2)}]$, we find that only factor **E** shows any systematic effect. Table 6.20 shows the estimated coefficients and Lenth t values for each model, and Figure 6.7 shows a half-normal plot of effects for $\ln[y_{(2)}]$. Factor **E** is statistically significant in each case, and the next largest Lenth t statistic is -1.44 . Figure 6.8 plots $\ln[y_{(2)}]$ versus **E**, with the simple regression model $\widehat{\ln[y_{(2)}]} = 5.45 - 0.86\mathbf{E}$.

Table 6.19. Number of cycles (in thousands) until failure of first two thermostats for each treatment combination in Bullington et al's design

A	B	C	D	E	F	G	H	J	K	L	$y_{(1)}$	$y_{(2)}$
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	957	2846
-1	-1	-1	-1	-1	1	1	1	1	1	1	206	284
-1	-1	1	1	1	-1	-1	-1	1	1	1	63	113
-1	1	-1	1	1	-1	1	1	-1	-1	1	76	104
-1	1	1	-1	1	1	-1	1	-1	1	-1	92	126
-1	1	1	1	-1	1	1	-1	1	-1	-1	490	971
1	-1	1	1	-1	-1	1	1	-1	1	-1	232	326
1	-1	1	-1	1	1	1	-1	-1	-1	1	56	71
1	-1	-1	1	1	1	-1	1	1	-1	-1	142	142
1	1	1	-1	-1	-1	-1	1	1	-1	1	259	266
1	1	-1	1	-1	1	-1	-1	-1	1	1	381	420
1	1	-1	-1	1	-1	1	-1	1	1	-1	56	62

**Fig. 6.6.** Scatterplot of $\ln(\text{cycles})$ for first two failures from each group of 10 thermostats

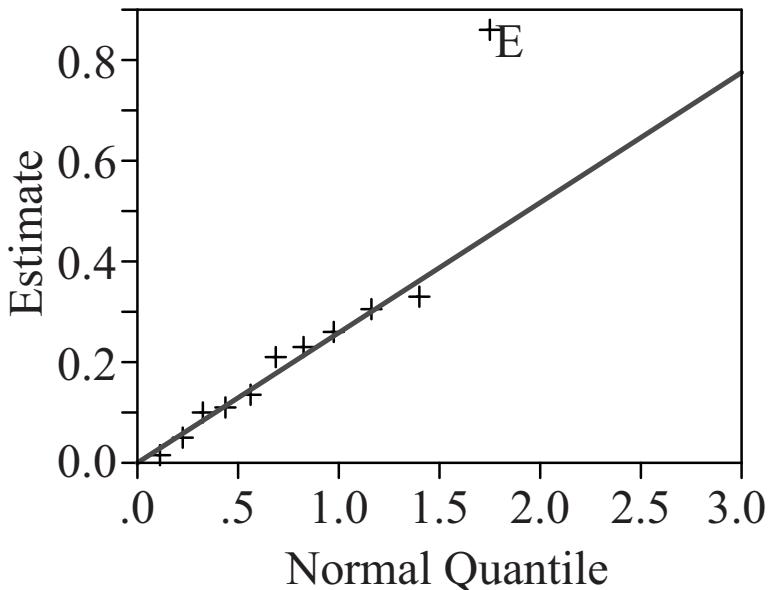


Fig. 6.7. Half-normal plot of effects from saturated model for $\ln(\text{cycles})$ for second failure from each group of 10 thermostats

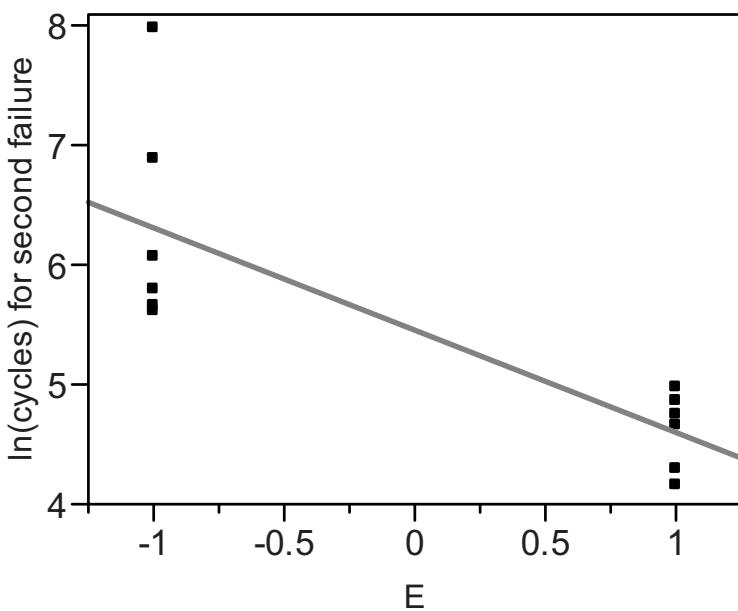


Fig. 6.8. $\ln(\text{cycles})$ for second failure versus E

Table 6.20. Saturated models for $\ln[y_{(1)}]$ and $\ln[y_{(2)}]$ in thermostat experiment

Term	$\ln[y_{(1)}]$		$\ln[y_{(2)}]$	
	Estimate	Lenth t	Estimate	Lenth t
Intercept	5.111	32.92	5.454	21.09
A	-0.119	-0.77	-0.328	-1.27
B	-0.016	-0.10	-0.102	-0.40
C	-0.130	-0.84	-0.108	-0.42
D	0.058	0.37	0.052	0.20
E	-0.778	-5.01	-0.861	-3.33
F	0.050	0.32	-0.017	-0.06
G	-0.219	-1.41	-0.229	-0.89
H	-0.088	-0.57	-0.211	-0.82
J	-0.077	-0.50	-0.134	-0.52
K	-0.210	-1.35	-0.259	-1.00
L	-0.223	-1.44	-0.303	-1.17

From Table 6.20, the preference for $\mathbf{E} = -1$ (i.e., small grain size) is obvious. Although a main effects model for the 12 runs did not show any other significant factor, we consider whether there is any useful information in the $\mathbf{E} = -1$ data regarding the effects of the remaining factors? Consider analyzing just the six observations with $\mathbf{E} = -1$. If this were a regular resolution III fractional factorial design with \mathbf{E} appearing in any length-3 words, then some main effect columns would have correlations of -1 or 1 after splitting the data in half. Instead, for Table 6.19, after eliminating the rows with $\mathbf{E} = +1$, we find that every column has a correlation of $1/3$ or $-1/3$ with the others. Of the 10 factors besides \mathbf{E} , factor \mathbf{H} stands out. Although its effect is not statistically significant for the response $\ln[y_{(2)}]$, the fact that the largest three values all occur at $\mathbf{E} = \mathbf{H} = -1$ is worthy of further investigation. Looking at the failure data for all 120 thermostats, the 3 treatment combinations $(1, 6, 11)$ with $\mathbf{E} = \mathbf{H} = -1$ only had 8 of the 30 thermostats fail before the test was terminated at 7342K cycles (while all 90 thermostats at other treatment combinations had already failed by 732K cycles). Follow-up tests at this treatment combination are recommended.

Example 6.6: Seven factors in 12 runs

Bermejo-Barrera et al. (2001) conducted an experiment to optimize the acid leaching step for determining trace amounts of elements in seafood products via atomic absorption spectrometry. The factors included three different reagents and four other parameters shown in Table 6.21. Twelve combinations of levels for the seven factors were investigated, as shown in Table 6.22. Although recovery was measured for 13 elements, we show results only for arsenic, calcium, cadmium, cobalt, mercury, manganese, and zinc.

Table 6.21. Factors and levels for Bermejo-Barrera et al. (2001) experiment

Factors	Levels	
	-1	1
A Nitric acid concentration (M)	0	2.4
B Hydrochloric acid concentration (M)	0	2.4
C Hydrogen peroxide concentration (M)	0	1.2
D Acid solvent volume (mL)	3	7
E Ultrasonic water-bath temperature (°C)	15	60
F Ultrasound exposure time (min)	10	120
G Mussel particle size (μm)	30	300

Table 6.22. Bermejo-Barrera et al. (2001) experiment

A	B	C	D	E	F	G	As	Ca	Cd	Co	Hg	Mn	Zn
1	-1	1	-1	-1	-1	1	77.0	82.6	95.2	93.8	8.8	100.4	85.7
1	1	-1	1	-1	-1	-1	87.3	86.6	75.6	75.4	66.5	86.9	92.6
-1	1	1	-1	1	-1	-1	92.9	107.1	82.5	84.1	77.2	105.0	78.9
1	-1	1	1	-1	1	-1	56.4	62.7	50.0	68.3	3.2	59.8	51.1
1	1	-1	1	1	-1	1	100.0	89.9	83.3	92.5	92.5	107.2	96.6
1	1	1	-1	1	1	-1	68.5	70.9	63.1	64.1	59.4	87.8	81.0
-1	1	1	1	-1	1	1	100.0	105.7	73.3	79.8	94.1	88.0	80.1
-1	-1	1	1	1	-1	1	81.2	73.1	56.7	42.5	4.4	67.5	74.1
-1	-1	-1	1	1	1	-1	62.2	39.6	30.0	26.9	8.8	34.6	31.4
1	-1	-1	-1	1	1	1	75.2	83.3	60.0	68.6	11.0	106.6	84.9
-1	1	-1	-1	-1	1	1	92.4	96.2	75.9	81.2	95.6	104.2	83.0
-1	-1	-1	-1	-1	-1	-1	55.2	31.0	15.0	36.2	13.2	46.6	27.1

For this design, a main effects model leaves 4 df for error. Table 6.23 provides a summary of this model for Co, Hg, Mn, and Zn. With 4 df, an individual t statistic must exceed 2.776 to be statistically significant at $\alpha = .05$. By this requirement, Co has no significant effects, Ca has one, Mn has two, and Zn has three. (Note: If one had 9–10 factors, and so only 1–2 df, rather than use $\text{Std Error} = \text{RMSE}/(12)^{1/2}$, one should fit a model with 11 orthogonal columns and use Lenth's PSE for testing effects.)

Table 6.23. Main effects models for Co, Ca, Mn, and Zn

Term	Co		Ca		Mn		Zn	
	Est.	t	Est.	t	Est.	t	Est.	t
Intercept	67.78	15.41	77.39	14.57	82.88	21.70	72.21	21.32
A	9.33	2.12	1.94	0.37	8.57	2.24	9.78	2.89
B	11.73	2.67	15.34	2.89	13.63	3.57	13.16	3.89
C	4.32	0.98	6.29	1.18	1.87	0.49	2.94	0.87
D	-3.55	-0.81	-1.12	-0.21	-8.88	-2.33	-1.22	-0.36
E	-4.67	-1.06	-0.07	-0.01	1.90	0.50	2.28	0.67
F	-2.97	-0.67	-0.99	-0.19	-2.72	-0.71	-3.62	-1.07
G	8.62	1.96	11.08	2.09	12.77	3.34	11.86	3.50
R^2 (%)	82.34		78.15		89.83		90.51	
RMSE	15.24		18.40		13.23		11.73	
Std Error	4.40		5.31		3.82		3.87	

For all seven elements listed in Table 6.22, no main effect besides **A**, **B**, and **G** has a *t* statistic exceeding 2.776 in magnitude. Thus, it appears that the two acid concentrations and particle size are the influential factors. With this in mind, a second analysis that is useful here is to fit a two-factor interaction model in these three factors, to investigate the possibility of interaction effects. The results are given in Table 6.24. Nominally, we have 5 df for error, although it must be remembered that these regression models were fit conditional on the results of a preliminary analysis. Given the likely downward bias in the MSE, the *t* statistics should be viewed as approximate.

Table 6.24. Two-factor interaction models involving three factors for Co, Ca, Mn and Zn

Term	Co		Ca		Mn		Zn	
	Est.	t	Est.	t	Est.	t	Est.	t
Intercept	67.78	22.92	77.39	29.68	82.88	40.88	72.21	48.34
A	9.14	2.91	-0.27	-0.10	6.08	2.83	7.00	4.42
B	12.91	4.12	15.33	5.54	15.53	7.22	13.01	8.21
G	5.38	1.71	7.88	2.85	10.19	4.74	11.44	7.22
AB	-9.72	-3.10	-9.58	-3.46	-7.72	-3.59	-1.26	-0.80
AG	3.54	1.13	-0.03	-0.01	5.69	2.65	-0.44	-0.28
BG	-0.59	-0.19	-6.63	-2.40	-7.46	-3.47	-8.32	-5.25
R^2 (%)	90.02		93.42		96.42		97.69	
RMSE	10.25		9.03		7.02		5.17	
Std Error	3.14		2.77		2.15		1.58	

For the 12-run design, each two-factor interaction column has a correlation of $+1/3$ or $-1/3$ with factors not appearing in the interaction. Because \mathbf{ABG} sums to -4 , the correlations between \mathbf{A} and \mathbf{BG} , \mathbf{B} and \mathbf{AG} , and \mathbf{G} and \mathbf{AB} are $-1/3$. This has the following consequences:

- In Table 6.24, main effect estimates are each correlated with one interaction; this makes their standard error $\text{RMSE}/(10\bar{6})^{1/2}$ rather than $\text{RMSE}/(12)^{1/2}$. This is a small (12.5%) increase in the variance, since $1/10\bar{6} = 1.125/N$.
- Estimates for main effects change when interactions are added. Compare estimates in the main effects model (Table 6.23) with estimates in Table 6.24. If the true model contains no interactions besides \mathbf{AB} , then for the main effects model each effect besides \mathbf{A} and \mathbf{B} is biased by an amount $\pm\beta_{AB}/3$. Since \mathbf{ABG} sums to -4 , $E(b_G) = \beta_G - (1/3)\beta_{AB}$. Note this relationship in the estimates themselves; the main effects model estimate b_G equals $5.38 - (-9.72/3)$. Since β_{AB} appears to be negative and β_G positive, the bias from omitting the \mathbf{AB} term in Table 6.23 makes \mathbf{G} 's effect appear more prominent.

The 12-run design is very well suited for fitting the two-factor interaction model in three factors. The two-factor interaction model for four factors can also be estimated but with much poorer precision. The additional correlations makes the standard error for each coefficient increase to $\text{RMSE}/(7.38)^{1/2}$. This is a substantial (62.5%) increase in the variance, since $1/7.38 = 1.625/N$.

6.3.4 Nonregular 16-run designs with generalized resolution 3.0

There exist four nonisomorphic OA($16, 2^{15}, 2$) in addition to the regular 2^{15-11} resolution III fraction. The best characterizations of these designs are due to Sun and Wu (1993) and Evangelaras, Georgiou, and Koukouvinos (2003). Table 6.17 is the choice recommended by Sun and Wu. The first seven factors of Table 6.17 (label them $\mathbf{A-G}$) form a replicated 2^{7-4} with generators $\mathbf{C} = \mathbf{AB}$, $\mathbf{E} = \mathbf{AD}$, $\mathbf{F} = \mathbf{BD}$, and $\mathbf{G} = \mathbf{ABD}$, and the last eight factors (label them \mathbf{H} and $\mathbf{J-P}$) form a 2^{8-4} with generators $\mathbf{L} = \mathbf{HJK}$, $\mathbf{N} = \mathbf{HJM}$, $\mathbf{O} = \mathbf{HKM}$, and $\mathbf{P} = \mathbf{JKM}$. This structure is identical to that for the regular 2^{15-11} obtained via the Sylvester construction described in Section 6.2. The difference between the regular 2^{15-11} fraction (S_{16}) and Table 6.17's design is that the 16 treatment combinations for the last 8 factors are reordered so that there is no complete aliasing of effects involving both groups except very high-order interactions such as $\mathbf{ABC} = \mathbf{HJKL}$. In particular, the Sylvester Hadamard matrix S_{16} is of the form

$$\begin{bmatrix} \mathbf{S}_8 & \mathbf{S}_8 \\ \mathbf{S}_8 & -\mathbf{S}_8 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 \end{bmatrix},$$

whereas the Table 6.17 design is taken from the Hall Type V Hadamard matrix

$$\mathbf{H}_{16}^V = \begin{bmatrix} \mathbf{S}_8 & \mathbf{H}_8 \\ \mathbf{S}_8 & -\mathbf{H}_8 \end{bmatrix},$$

where \mathbf{H}_8 is the following reordering of the rows of \mathbf{S}_8 :

$$\mathbf{H}_8 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \end{bmatrix}.$$

This reordering keeps the main effects for \mathbf{A} – \mathbf{G} orthogonal to main effects for \mathbf{H} – \mathbf{P} while avoiding any complete aliasing between main effects of one set and two-factor interactions of the other.

The confounding frequency vector for this design is given in (6.7). Sun and Wu (1993) summarized the aliasing of two-factor interactions as follows:

- The 21 two-factor interactions for factors \mathbf{A} – \mathbf{G} are fully aliased together with each other (in sets of 3) and the main effects. This accounts for $A_3(1) = 7$ and one-third of $A_4(1)$.
- The 28 two-factor interactions for factors \mathbf{H} – \mathbf{P} are fully aliased with each other in sets of 4, but are not fully aliased with any main effects. This accounts for the rest of $A_4(1)$.

- The 56 two-factor interactions involving 1 factor from **A–G** and 1 from **H–P** only have partial aliasing with other lower-order effects. This accounts for $A_3(1/2) = 112$ and $A_4(1/2) = 336$.

By contrast, the regular 2^{15-11} fully aliases each two-factor interaction with one main effect and six other two-factor interactions. One has much greater flexibility in fitting models with some two-factor interactions by using the design in Table 6.17.

The transpose of \mathbf{H}_{16}^V is Hall's Type IV Hadamard matrix. Since \mathbf{S}_8 is symmetric,

$$\mathbf{H}_{16}^{IV} = \begin{bmatrix} \mathbf{S}_8 & \mathbf{S}_8 \\ \mathbf{H}'_8 & -\mathbf{H}'_8 \end{bmatrix}.$$

The corresponding design appears in Table 6.25. Table 6.25's design has the same confounding frequency vector (6.7). Its seven fully aliased length-3 words are **{AHJ, BHK, CHL, DHM, EHN, FHO, GHP}**, and the 21 fully aliased length-4 words are obtained as the generalized interaction of pairs of these length-3 words. The 15 main effects have 21 two-factor interactions as full aliases. The remaining 84 two-factor interactions are fully aliased with one another in pairs.

Table 6.25. Hall Type IV OA($16, 2^{15}, 2$)

A	B	C	D	E	F	G	H	J	K	L	M	N	O	P
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1
1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1
-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1
1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1
1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1
-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1	1	-1
1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	1	-1	1	-1	-1	-1	-1	-1	1	-1	1	1	1	1
-1	-1	1	1	-1	1	-1	-1	1	1	-1	-1	1	-1	1
-1	-1	-1	1	1	-1	1	-1	1	1	1	-1	-1	1	-1
-1	1	-1	-1	-1	1	1	-1	1	-1	1	1	1	-1	-1
-1	1	1	-1	1	-1	-1	1	-1	-1	-1	1	-1	1	1
1	-1	-1	-1	1	1	-1	-1	-1	1	1	1	-1	-1	1
1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	1	1	-1

Evangelaras, Georgiou, and Koukouvinos (2003) showed that the designs in Tables 6.17 and 6.25 have similar projection properties when projecting into five or fewer factors. Both designs permit estimation of the following:

- a full factorial model in all but 7 of the 455 projections into 3 factors;

- the two-factor interaction model (1.3) in over 92% of the possible 1365 subsets of 4 factors.

What 16-run design is recommended for fewer than 15 factors? Deng and Tang (2002) identified the projections of the five OA(16, 2^{15} , 2) designs with the minimum generalized aberration. For 14 factors, eliminate factor **H** from the design in Table 6.25; this is the only OA(16, 2^{14} , 2) that permits estimation of the full factorial model in any subset of 3 factors. For 13 factors, eliminate **H** and **P** from Table 6.25. For 12 factors, eliminate factors **C**, **E**, and **F** from the Table 6.17 design; for 11 factors, also drop **P**. For 9 or 10 factors, the minimum generalized aberration projections come from Hall's Type III Hadamard matrix; see Deng and Tang (2002) for details.

Although these nonregular 16-run designs have been recommended since Sun and Wu (1993), the author is not aware of any published examples.

6.3.5 20-Run designs with generalized resolution 3.4

There are three nonisomorphic OA(20, 2^{19} , 2); see, for example, Hedayat, Sloane, and Stufken (1999, pp. 155 and 158) or Deng and Tang (2002). The simplest of these designs to construct is the design proposed by Plackett and Burman (1946), which is obtained by cycling

$$(1, -1, 1, 1, -1, -1, -1, 1, -11, -1, 1, 1, 1, 1, -1, -1, 1)$$

and then appending the treatment combination with all -1 's. Table 6.26 displays this design in full. Evangelaras, Georgiou, and Koukouvinos (2003) found that the three OA(20, 2^{19} , 2) have the same projection properties into three or four factors but differed on five-factor projections. For over 94% of the three-factor projections, the frequencies are as even as possible: a replicated 2^3 plus an additional 2^{3-1} fraction. However, for the remaining subsets of size 3, the frequencies are very uneven: 1 for four treatment combinations and 4 for the others, producing a large correlation ($\pm .6$) between some main effects and two-factor interactions. The impact of such uneven frequencies will become apparent in the following example. Because the design in Table 6.26 was generated by cycling of rows, the 57 projections with such disparate frequencies can be obtained by cycling 3 sets as follows (note that 1 follows 19):

(1, 2, 13), (2, 3, 14), (3, 4, 15), (4, 5, 16), (5, 6, 17), (6, 7, 18), (7, 8, 19), (1, 8, 9), (2, 9, 10), (3, 10, 11), (4, 11, 12), (5, 12, 13), (6, 13, 14), (7, 14, 15), (8, 15, 16), (9, 16, 17), (10, 17, 18), (11, 18, 19), (1, 12, 19);

(1, 3, 17), (2, 4, 18), (3, 5, 19), (1, 4, 6), (2, 5, 7), (3, 6, 8), (4, 7, 9), (5, 8, 10), (6, 9, 11), (7, 10, 12), (8, 11, 13), (9, 12, 14), (10, 13, 15), (11, 14, 16), (12, 15, 17), (13, 16, 18), (14, 17, 19), (1, 15, 18), (2, 16, 19);

(1, 5, 11), (2, 6, 12), (3, 7, 13), (4, 8, 14), (5, 9, 15), (6, 10, 16), (7, 11, 17), (8, 12, 18), (9, 13, 19), (1, 10, 14), (2, 11, 15), (3, 12, 16), (4,

(13, 17), (5, 14, 18), (6, 15, 19), (1, 7, 16), (2, 8, 17), (3, 9, 18), (4, 10, 19).

Example 6.7: 20-run design with 19 factors

Wu et al. (2005) reported an experiment involving image analysis for processing cDNA microarrays using RoBioVision software. Five slide images from one rat (treated with acetaminophen) were produced. Each slide is two-sided, with 1248 spots on each side. RoBioVision considers the pixels for each spot and computes numerous measures of spot intensity and quality. Median intensity is just 1 of 27 summary measures. The correlation for median intensity for one side versus the other side for each slide provides a measure of reproducibility. The authors used the 20-run Hadamard design in Table 6.26 to investigate the effects of 19 image analysis parameters. Table 6.27 provides the factor names and levels (using x_1 – x_{19} rather than letters as factor labels, to more readily identify the three-factor projections with uneven frequencies.) At each of the 20 treatment combinations, 5 correlations for median intensity were computed, 1 for each slide. Thus, the raw data consisted of 100 correlations. Since Wu et al.’s article did not include this detailed data, Table 6.26 simply shows the average of the five correlations at each treatment combination.

Table 6.26. Wu et al's (2005) 20-run experiment

Table 6.27. Factors and levels for Wu et al. (2005) image processing experiment

Factors	Levels	
	-1	1
x_1 Spot minimum diameter filter (pixels)	5	12
x_2 Spot minimum diameter sort (pixels)	13	20
x_3 Spot maximum diameter filter (pixels)	40	60
x_4 Spot maximum diameter sort (pixels)	25	35
x_5 Spot minimum volume (pixels ³)	10^3	10^6
x_6 Spot minimum mean (pixels)	25	65
x_7 Spot minimum solidity filter (units)	0.40	0.58
x_8 Spot minimum solidity sort (units)	0.65	0.90
x_9 Spot minimum roundness filter (units)	0.40	0.58
x_{10} Spot minimum roundness sort (units)	0.65	0.90
x_{11} Spot maximum aspect ratio filter (units)	1.8	3.0
x_{12} Spot maximum aspect ratio sort (units)	1.1	1.6
x_{13} Spot maximum off-center (pixels)	10	20
x_{14} Dust minimum diameter filter (pixels)	2	5
x_{15} Dust minimum diameter sort (pixels)	6	18
x_{16} Dust threshold (units)	150	500
x_{17} Dust minimum mean (pixels)	66	85
x_{18} Dust minimum solidity (pixels)	20	40
x_{19} Dust minimum roundness (pixels)	25	45

Figure 6.9 shows a Pareto plot of the 19 estimated regression coefficients. Wu et al. (2005) computed a mean square error of 0.030933, and used this to compute a standard error of 0.01759 for each coefficient, based on $N = 100$. Their Figure 3a is equivalent to Figure 6.9 here, except they plot t statistics. With six t statistics larger than 2, Wu et al. proceeded to investigate these factors further in a resolution VI 2^{6-1} fractional factorial design. Without knowledge of their RMSE, we would need to use Lenth's PSE, which is 0.02478. Even though larger than the standard error based on replication, the largest four estimates still have Lenth t -ratios exceeding the critical value for 19 estimates, $c_{.05}^{\text{IER}} = 2.120$ (from Appendix C), and Lenth t for x_{11} is only slightly smaller at 2.113.

Before proceeding further in our analysis, consider the plot of the average correlation versus x_1 in Figure 6.10. Two facts are obvious from this plot. First, $x_1=+1$ increases repeatability, producing mean correlations between .89 and .91 for all 10 treatment combinations. These are the 10 highest values. Only 3 of the 10 treatment combinations with $x_1 = -1$ produce a similarly high correlation. Second, if any other effect is active, it must have an interaction with x_1 , since that factor's effect at $x_1=+1$ will be weaker than its effect at $x_1 = -1$. From the Pareto plot, it appears that x_4 might have an effect. If one splits the data by x_1 , the resulting models for x_4 are

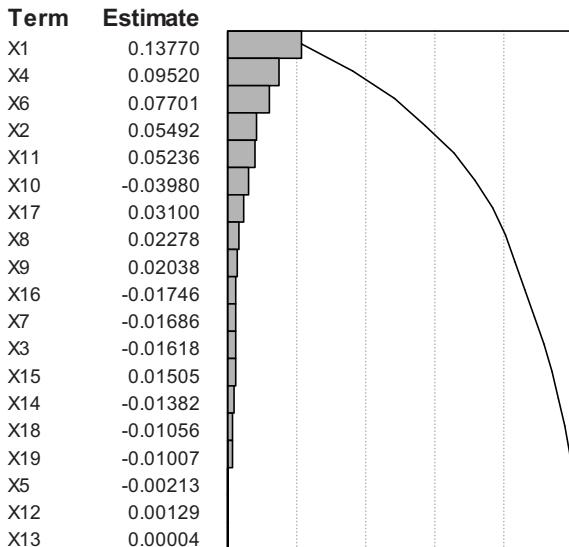


Fig. 6.9. Pareto plot of estimates from saturated model for mean correlation

- At $x_1 = -1$: $\hat{\rho} = 0.627 + 0.184x_4$ with RMSE = 0.148; $t = 3.93$;
- At $x_1 = +1$: $\hat{\rho} = 0.90225 + 0.00639x_4$ with RMSE = 0.003; $t = 6.41$.

These two models are plotted in Figure 6.11.

Because the x_1x_4 interaction column is correlated with every other main effect besides x_1 and x_4 , if $\beta_{1.4} \neq 0$, its absence when fitting the main effects model in Figure 6.9 means that every other estimate is biased. We consider the bias first for b_6 and then for b_2 , since these are the next largest estimates in the Pareto plot. The projection into columns (1, 4, 6) is 1 of the 57 problem subsets noted before discussion of this example. The four treatment combinations with $x_1x_4x_6 = +1$ each appear only once, whereas the four treatment combinations with $x_1x_4x_6 = -1$ each appear four times. Thus, the $x_1x_4x_6$ column sums to $4 - 16 = -12$, producing a correlation of $-.6$ ($= -12/20$) between the x_6 and x_1x_4 columns. Assuming no interactions besides the x_1x_4 interaction exist,

$$E(b_6) = \beta_6 - 0.6\beta_{1.4},$$

if we fit a model without the x_1x_4 term. If one includes the x_1x_4 interaction in the model for our example, the estimate for β_6 is greatly diminished and is no longer statistically significant, whereas $\beta_{1.4}$ remains statistically significant.

Consider now the bias in b_2 in our original analysis based on a main effects model. The correlation between x_2 and x_1x_4 is not as great. The four treatment combinations with $x_1x_2x_4 = +1$ each appear twice, whereas the four treatment combinations with $x_1x_2x_4 = -1$ each appear three times. Thus, the $x_1x_2x_4$ column sums to $8 - 12 = -4$, producing a correlation of $-4/N = -.2$

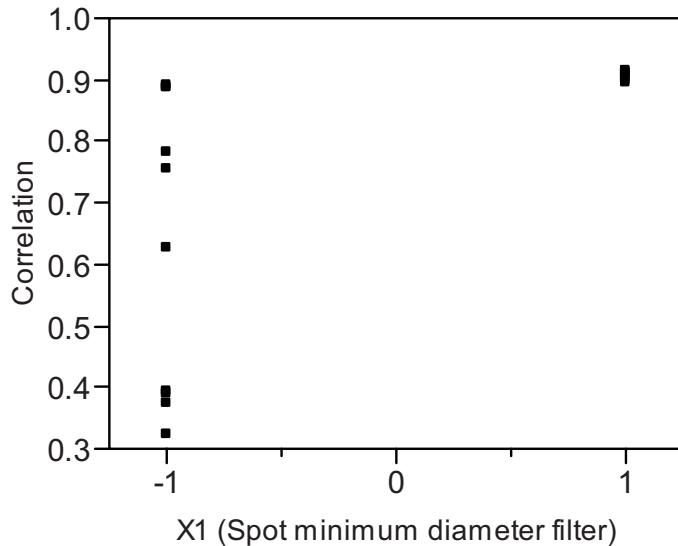


Fig. 6.10. Mean correlations versus x_1 for Wu et al. (2005)

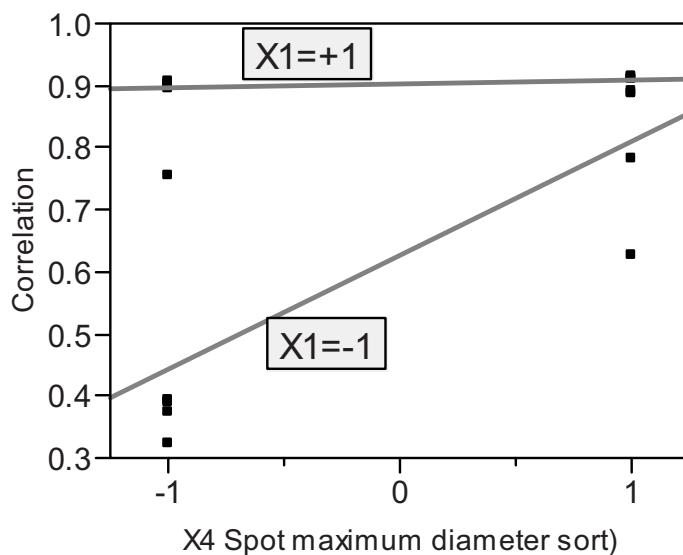


Fig. 6.11. Mean correlations versus x_4 for Wu et al. (2005); fitted models are for different values of x_1

between the x_2 and x_1x_4 columns. Assuming no interactions besides the x_1x_4 interaction exist,

$$E(b_2) = \beta_2 - 0.2\beta_{1.4},$$

if we fit a model without the x_1x_4 term. If one includes the x_1x_4 interaction in the model, the estimate for β_2 is diminished, but the impact is less than for b_6 , since the correlation is only one-third as large.

One cannot fit a model with the x_1x_4 interaction and all 19 main effects, since this would produce a singular model matrix. However, one can list all of these effects and use the forward selection method for selecting a regression model. After including x_1 , x_4 , and x_1x_4 , no main effects appear to be useful; the smallest p -value for adding another factor is .107 for x_2 .

Wu et al. (2005) continued to collect data, varying factors 1, 2, 4, 6, 10, and 11. Adding data where $x_1x_4x_6 = +1$ decreased the correlation between x_6 and x_1x_4 . The additional data confirmed what we suspected from the initial 20-run design: that x_1x_4 is an active effect while x_6 is not. In fact, after more than doubling the number of treatment combinations, no other main effects appeared important, although some minor two-factor interactions did.

For the 12-run design, all nonzero correlations between main effect and two-factor interaction columns were of the same magnitude (1/3). However, for any OA($20, 2^{19}, 2$), each two-factor interaction has a correlation of $\pm .6$ with one main effect and $\pm .2$ with 16 other factors. These larger coefficients mean than the potential for bias to main effects from omitted two-factor interactions should be routinely considered in any thorough analysis of 20-run designs.

We have performed this analysis using the correlation as a response rather than the recommended transformation $z = .5[\ln(1+r) - \ln(1-r)]$ (see Section 2.8.4); the results are essentially the same either way. If we had available the 100 individual correlations, the variance-stabilizing transformation might have made some difference in the analysis.

Example 6.8: Another 20-run design with 19 factors

Bell, Ledolter, and Swersey (2006) reviewed the literature for marketing experiments based on fractional factorial designs and then carefully described a mass mailing experiment based on the 20-run Plackett–Burman design in Table 6.26. The 19 factors and their levels are listed in Table 6.28. For each treatment combination, 5000 credit card offers were sent. The number of positive responses for rows 1–20 from Table 6.26 were 52, 47, 43, 86, 99, 37, 49, 40, 39, 108, 30, 57, 68, 61, 60, 104, 134, 42, 38, and 104, respectively. Thus, the response rates ranged from a low of $30/5000 = 0.6\%$ to a high of 2.68%. The mean response rate was 1.3%. If the factors had no effect, the observed proportions should exhibit a standard deviation of approximately $[(0.013)(0.987)/5000]^{1/2} = 0.0016$; since the data show much more variability than this, at least one of the factors must impact response rate.

Table 6.28. Factors and levels for Bell et al. (2006) credit card offer experiment

Factors	Levels	
	-1 (Control)	1 (New Idea)
x_1 Envelope teaser	General	Product-specific
x_2 Interest rate	Low	High
x_3 Second buckslip	No	Yes
x_4 Information on buckslip	Product info	Free gift
x_5 Reply envelope	Control	New style
x_6 Free gift value	High	Low
x_7 Product selection	Many	Few
x_8 Signature	Manager	Senior executive
x_9 Postscript on letter	Standard	New
x_{10} List of benefits	Standard layout	Creative layout
x_{11} Letter headline	Headline 1	Headline 2
x_{12} Copy message	Targeted	Generic
x_{13} Personalize letter copy	No	Yes
x_{14} Sticker	Yes	No
x_{15} Price graphic	Small	Large
x_{16} Additional graphic on envelope	Yes	No
x_{17} Postage	Pre-printed	Stamp
x_{18} “Official” ink stamp on envelope	Yes	No
x_{19} Return address	Blind	Add company name

We now must choose a variance-stabilizing transformation, since if the true response rates varied from 0.008 to 0.025, the largest variance for \hat{p} would be three times the smallest variance. Since the Freeman–Tukey transformation (2.12) stabilizes the variance so effectively, we begin by fitting a model to this response. For instance, the first response of 52/5000 becomes

$$f_{\text{FT}}(\hat{p}) = \arcsin[\sqrt{52/5001}] + \arcsin[\sqrt{(53)/5001}] = 0.2053.$$

Fitting a saturated main effects model to the transformed response $f_{\text{FT}}(\hat{p})$, we obtain the sorted estimates shown in Table 6.29. For binomial data, the Freeman–Tukey transformation has a standard deviation of approximately $[1/(n + 0.5)]^{1/2} = 0.01414$. Thus, the standard error for the least squares estimates is $0.01414/N^{1/2} = 0.00316$. We use this known standard error to construct test statistics in Table 6.29, rather than using Lenth’s PSE. Five main effects have test statistics that exceed the .05 critical value for a standard normal distribution. The largest effect was anticipated; a higher interest rate should decrease the response rate. Furthermore, the absence of a sticker ($x_{14} = 1$) lowers the response rate. The next two largest estimates were surprising in that $b_{12} = 0.0122$ implies that the generic message was more effective than a targeted message, and $b_3 = -0.0118$ implies that including the second buckslip

made responses less likely. Perhaps one or both of these are the result of bias from two-factor interactions.

There are additional clues that there might be at least one active interaction. First, there are no p -values in Table 6.29 greater than .75, which is equivalent to saying that there is no clump of estimates near zero. This could be the consequence of a large interaction, since its presence would bias 16 of the main effect estimates by $\pm 0.2\beta_{i,j}$. Second, this bias would tend to inflate the PSE. The PSE calculated from the 19 estimates in Table 6.29 is 0.00513, making it larger than the correct standard error of 0.00316 based on the Binomial distribution.

Table 6.29. Sorted estimates from main effects model for $f_{FT}(\hat{p})$

Term	Estimate	z	p-Value
x_2	-0.0375	-11.88	.0000
x_{14}	-0.0228	-7.21	.0000
x_{12}	0.0122	3.87	.0001
x_3	-0.0118	-3.73	.0002
x_{11}	-0.0093	-2.96	.0031
x_5	-0.0044	-1.39	.1644
x_{16}	0.0042	1.31	.1886
x_9	-0.0036	-1.14	.2549
x_6	0.0035	1.11	.2673
x_{15}	-0.0034	-1.08	.2784
x_{13}	0.0034	1.08	.2796
x_8	-0.0028	-0.88	.3815
x_4	-0.0027	-0.86	.3921
x_{10}	0.0025	0.79	.4270
x_{17}	0.0023	0.74	.4610
x_1	0.0021	0.68	.4965
x_{19}	0.0015	0.47	.6389
x_{18}	0.0014	0.45	.6535
x_7	-0.0010	-0.33	.7411

The 20-run Plackett–Burman design will permit estimation of the two-factor interaction model in any projection of four factors, and in some but not all projections of five factors. For the five factors 2, 3, 11, 12, and 14, the two-factor interaction model is not estimable. However, if one specifies this model with 15 terms and uses forward selection regression to identify important effects, the third term to enter is the $x_2 * x_{14}$ interaction, and the model with x_2, x_{14} and this interaction explains 87% of the variation in $f_{FT}(\hat{p})$. The $x_2 * x_{14}$ interaction column is most correlated with x_3 . If the $x_2 * x_{14}$ interaction is included, the x_3 main effect estimate is greatly reduced and is no longer statistically significant. Since the $x_2 * x_{14}$ interaction is reasonable

to explain, Bell et al. (2006) concluded that the second buckslip's significant estimate in Table 6.29 ($b_3 = -0.0118$) was due to bias from omitting this interaction. We consider adding additional main effects and interactions to the three-term model. Only x_{12} is found to be statistically significant. The reduced model is displayed in Figure 6.12. The largest predicted response is 0.30566; this corresponds to a predicted proportion of

$$0.5\{1 - [1 - (\sin 0.30566 + (\sin 0.30566 - 1/\sin 0.30566)/5000)^2]^{1/2}\} = 0.0231,$$

using the transformation by Miller (1978) discussed prior to Figure 2.12. Actually, here the Freeman–Tukey variance-stabilizing transformation provided assurance that the constant variance assumption was not violated. However, Bell et al. (2006) reached the same conclusions simply using the proportion as the response.

Projections of the 19-factor, 20-run orthogonal design

What about using a 20-run design with fewer than 19 factors? Although the 20-run design suggested by Plackett and Burman (1946) and shown in Table 6.26 is the most convenient and commonly used 20-run design, it is not the best choice for several cases where $k < 19$. Deng and Tang (2002) searched for the minimum G-aberration projection from all three nonisomorphic OA(20, 2^{19} , 2). To reduce the computational burden, the cfv (6.3) was only calculated up to five-factor interactions; their MA-5 classifier ranked designs based on (F_3, F_4, F_5) . Xu and Deng (2005) improved that search in two ways. They considered all 20-run strength-2 orthogonal arrays in fewer than 19 factors, not just projections of the three OA(20, 2^{19} , 2). In addition, Xu and Deng (2005) used the moment aberration projection (MAP) criterion, which is more effective than the cfv at discriminating between designs. For $N = 20$, the best MAP design was always best in terms of minimum G-aberration; the converse is not true. Furthermore, Xu and Deng illustrated, for 20-run designs with seven factors, that the subtle differences MAP detects and minimum G-aberration misses can affect the design's estimation capacity.

Although not evident from the tables in either of the above-cited articles, the best projections for 9–18 factors in terms of minimum G-aberration can all be obtained from the single OA(20, 2^{19} , 2) in Table 6.30, which is equivalent to the design H20-P in Deng and Tang (2002) and Xu and Deng (2005) or Hedayat, Sloane, and Stufken's (1999, Table 7.23). Table 6.31 shows the poor three-factor projections for the design in Table 6.30, and Table 6.32 lists the minimum G-aberration projections for 9–18 factors, together with the number of poor projections into three columns. Equivalent designs in terms of minimum G-aberration are available as projections from the Table 6.26 (Plackett–Burman) design for 14–18 factors, but not for 13 or fewer.

Summary of Fit for Reduced Model

RSquare	0.908178
Root Mean Square Error	0.017233
Mean of Response	0.223963
Observations	20

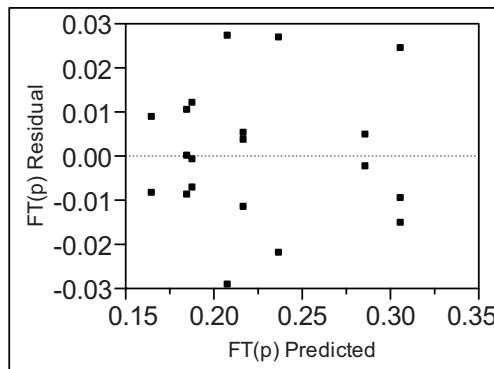
Analysis of Variance

Source	DF	Sum of Sq.	Mean Square	F Ratio
Model	4	0.0440609	0.011015	37.09
Error	15	0.0044548	0.000297	
C. Total	19	0.0485157		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	0.22396	0.00385	58.12	<.0001
x2	-0.03754	0.00385	9.74	<.0001
x12	0.00996	0.00393	-2.53	0.0230
x14	-0.02277	0.00385	5.91	<.0001
x2*x14	0.01142	0.00393	2.90	0.0109

Residual by Predicted Plot



Prediction Profiler

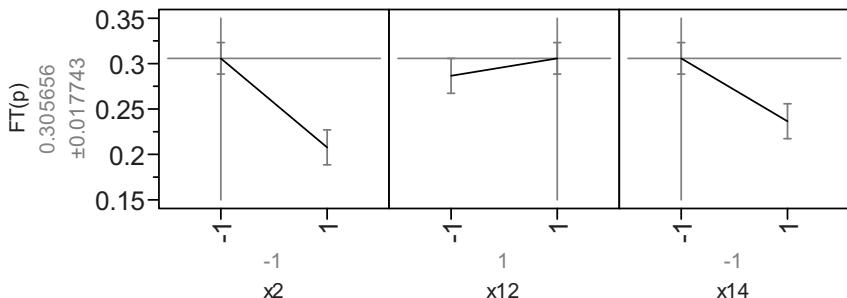


Fig. 6.12. Reduced model for Bell et al.'s (2006) mail experiment data

Table 6.30. 20-Run design based on Williamson's Hadamard matrix

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	1	1	1	-1	-1	-1	-1	1	1	1	1	1	-1	-1	-1	-1	-1	-1
1	1	1	-1	1	-1	-1	-1	-1	1	-1	-1	-1	1	1	1	1	-1	-1
1	1	-1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1
1	-1	1	-1	-1	1	1	-1	-1	-1	-1	1	1	-1	-1	1	1	1	1
1	-1	-1	1	1	-1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	-1	1
1	-1	-1	1	1	-1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	1
1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	1	-1	1	1	-1	-1
-1	1	1	-1	-1	-1	1	1	-1	1	1	-1	-1	1	-1	-1	1	1	1
-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	1
-1	1	-1	1	-1	-1	1	-1	1	1	-1	-1	1	1	-1	1	-1	1	1
-1	1	-1	-1	1	1	-1	-1	1	-1	1	-1	1	1	-1	-1	1	-1	-1
-1	1	-1	-1	1	1	-1	-1	1	-1	1	-1	1	1	-1	1	-1	-1	-1
-1	-1	1	1	-1	-1	-1	-1	1	-1	-1	-1	1	1	-1	1	1	-1	-1
-1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	1	1	-1	1	1	-1	-1
-1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	1	1	-1	1	1	-1	-1
-1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	1	1	-1	1	1	-1	-1
-1	-1	-1	1	1	1	-1	-1	1	-1	-1	-1	1	1	-1	1	1	-1	-1

Zhao, et al. (2005) used $k = 13$ adjacent columns of the cyclic 20-run Hadamard design. That design has $A_3(0.6, 0.2) = (17, 269)$, which is higher G-aberration (and G₂-aberration) than the subset of columns for $k = 13$ suggested in Table 6.32. By using the columns recommended in Table 6.32, one minimizes the number of large correlations between main effects and two-factor interaction columns. Kalil, Maugeri, and Rodrigues (2000) used 10 adjacent columns of the cyclic 20-run Hadamard design, which has $A_3(0.6, 0.2) = (7, 113)$. Note that by the more careful choice of the 10 columns as recommended in Table 6.32, they could utilize an OA(20, 2^{10} , 2) with generalized resolution 3.8 rather than 3.4. With $k = N/2$, they also could have used a nonorthogonal resolution IV design (see Section 7.4).

Table 6.31. Three-factor projections from the OA($20, 2^{19}, 2$) in Table 6.30 corresponding to $A_3(0.6) = 57$

(1, 2, 3)	(1, 4, 5)	(1, 6, 7)
(1, 8, 9)	(1, 10, 19)	(1, 11, 18)
(1, 12, 17)	(1, 13, 16)	(1, 14, 15)
(2, 4, 10)	(2, 5, 19)	(2, 6, 17)
(2, 7, 18)	(2, 8, 15)	(2, 9, 16)
(2, 11, 12)	(2, 13, 14)	(3, 4, 19)
(3, 5, 10)	(3, 6, 11)	(3, 7, 12)
(3, 8, 13)	(3, 9, 14)	(3, 15, 16)
(3, 17, 18)	(4, 6, 14)	(4, 7, 13)
(4, 8, 12)	(4, 9, 11)	(4, 15, 17)
(4, 16, 18)	(5, 6, 16)	(5, 7, 15)
(5, 8, 18)	(5, 9, 17)	(5, 11, 13)
(5, 12, 14)	(6, 8, 10)	(6, 9, 19)
(6, 12, 15)	(6, 13, 18)	(7, 8, 19)
(7, 9, 10)	(7, 11, 16)	(7, 14, 17)
(8, 11, 17)	(8, 14, 16)	(9, 12, 18)
(9, 13, 15)	(10, 11, 15)	(10, 12, 16)
(10, 13, 17)	(10, 14, 18)	(11, 14, 19)
(12, 13, 19)	(15, 18, 19)	(16, 17, 19)

Table 6.32. Minimum G-aberration projections from the OA($20, 2^{19}, 2$) in Table 6.30

k	Columns from Table 6.30	$A_3(0.6, 0.2)$	$A_4(0.6, 0.2)$
9	4–7, 10–12, 17, 19	0, 84	18, 108
10	4–7, 10–12, 17–19	0, 120	30, 180
11	4–7, 9–12, 17–19	5, 160	30, 300
12	1–6, 9–10, 12–13, 15, 18	8, 212	39, 456
13	1–7, 9–10, 12–13, 15, 18	14, 272	47, 668
14	1–4, 6–7, 10, 12–16, 18–19	20, 344	60, 941
15	1–7, 10, 12–16, 18–19	26, 429	81, 1284
16	1–7, 9–10, 12–16, 18–19	32, 528	108, 1712
17	Any 17 columns	40, 640	140, 2240
18	Any 18 columns	48, 768	180, 2880
19	All 19 columns	57, 912	228, 3648

6.3.6 24-Run designs with generalized resolution 3. $\bar{6}$

Plackett and Burman (1946) proposed the OA(24, 2³, 2) obtained by cycling the row “+++++----+---++--++-+” and then appending the treatment combination with all 23 factors at the low level. The design is shown in Table 6.33. Of the 60 OA(24, 2³, 2) examined by Evangelaras, Georgiou, and Koukouvinos (2004), this design has the best three-factor projections, providing three replicates of a 2³ for four-sevenths of the three-factor subsets and two replicates of a 2³ plus two replicates of a 2³⁻¹ for three-sevenths of the subsets. Each two-factor interaction column has a correlation of $\pm 1/3$ with 9 main effects and is orthogonal to the other 14. It projects into a partially replicated 2⁴ for three-sevenths of the four-factor projections and into 14 of the 16 treatment combinations for the remaining four-factor projections. As proven by Bulutoglu and Cheng (2003), this design (and any other orthogonal array obtained by the first Paley construction) supports estimation of the two-factor interaction model in every subset of four factors.

Table 6.33. 24-Run design based on Paley's Hadamard matrix

Strength-3 orthogonal arrays with 24 runs exist for up to 12 factors (see Section 7.3), so we only consider designs for $k \geq 13$ here. Ingram and Tang (2005) examined projections of 60 (of the 130 possible) Hadamard designs. The designs mentioned below are the best known at this time. For 22 (or 21) factors, the best design is a projection of Paley's design in Table 6.33; simply drop the last column (or two). For $k = 19$ and 20, Ingram and Tang found that the best design was a projection of a different Hadamard design. Table 6.34 is the best design (currently known) for 20 factors in 24 runs with respect to G-aberration and G_2 -aberration. For 19 factors, drop the last column. For smaller k , refer to Ingram and Tang (2005).

Dürig and Fasshi (1993) utilized 13 adjacent columns from the 24-run design in Table 6.33 for their 13-factor experiment; this choice produces $A_3(1/3) = 122$, 32 more than $A_3(1/3)$ for the best OA(24, 2^{13} , 2) reported by Ingram and Tang (2005), obtained as a projection of another Hadamard matrix.

Table 6.34. Best 20-factor design, a projection of Sloane's Had.24.59

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	-1	-1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	1	1	1	-1	1	-1	-1	-1	1	1	1	1	1	-1	-1	-1	-1	-1	-1
1	1	1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1
1	1	-1	-1	-1	-1	-1	1	1	-1	1	-1	1	-1	1	1	-1	1	-1	1
1	-1	1	-1	-1	-1	-1	1	1	1	1	-1	1	-1	1	1	-1	1	1	-1
-1	1	-1	-1	-1	-1	-1	1	-1	-1	-1	1	1	1	-1	1	1	-1	1	-1
1	-1	-1	1	1	1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	-1	-1
-1	1	-1	1	1	1	1	-1	1	-1	1	-1	-1	1	1	-1	-1	1	1	-1
1	1	-1	-1	-1	1	1	1	-1	1	-1	-1	1	1	-1	1	1	-1	-1	-1
1	-1	-1	-1	-1	1	1	-1	1	1	-1	-1	1	-1	1	1	-1	-1	1	1
-1	-1	1	-1	-1	1	1	1	1	-1	-1	1	1	-1	1	-1	1	1	-1	1
-1	-1	1	-1	-1	1	1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1
1	-1	-1	-1	-1	1	1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1
-1	-1	1	-1	-1	1	1	1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	-1
1	-1	-1	1	1	-1	1	-1	1	1	-1	1	1	-1	1	1	-1	1	1	-1
-1	1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	1	-1	-1	1	1
-1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	1	-1	1	1	-1
-1	-1	1	1	-1	1	-1	1	-1	1	-1	1	-1	1	1	-1	1	1	-1	1
-1	-1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1
1	-1	1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1
-1	1	1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1
-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1
-1	1	-1	-1	1	1	-1	-1	1	1	1	-1	-1	1	-1	-1	1	-1	-1	1
-1	-1	1	-1	-1	1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1

6.3.7 Nonregular strength-2 designs of size 28 or more

As mentioned earlier, Plackett and Burman (1946) proposed OA($N, 2^{N-1}, 2$) designs up to $N = 100$, and even larger designs are available from the website <http://www.research.att.com/~njas/hadamard/>. See also Hedayat, Sloane, and Stufken (1999, Ch. 7). The 28-run design proposed by Plackett and Burman is one of literally hundreds of OA(28, 2^{27} , 2). Plackett and Burman's design is constructed as

$$\mathbf{D} = \begin{bmatrix} \mathbf{X} & \mathbf{Y} & \mathbf{Z} \\ \mathbf{Z} & \mathbf{X} & \mathbf{Y} \\ \mathbf{Y} & \mathbf{Z} & \mathbf{X} \\ -1 & \cdots & 1 \end{bmatrix},$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 \\ -1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix},$$

$$\mathbf{Y} = \begin{bmatrix} -1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 \\ -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 \\ -1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 \\ -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 \\ -1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 \\ -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 \end{bmatrix},$$

$$\mathbf{Z} = \begin{bmatrix} 1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 \\ -1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 \\ -1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 \\ -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & 1 \end{bmatrix}.$$

Its correlations between main effects and two-factor interactions are $\pm 1/7$ and $-3/7$, with frequencies 2574 and 351, respectively. Commercial statistical software sometimes furnishes other OA(28, 2^{27} , 2) with correlations between main

effects and interactions as high as 5/7. Belcher-Novosad and Ingram (2003) have searched for the best projections, although their excursion algorithm does not guarantee optimality.

There exists an OA(32, 2^{31} , 2) obtained by the first Paley method. Cheng (1998) showed that it permits estimation of the two-factor interaction model in every set of six factors. That design is obtained by cycling the row $[+ - - + - + - - + + + - + + - + + +]$ and then appending a row of -1 's. The nonzero elements of the cfv are $[A_3(1/4) = 2480, A_4(1/4) = 17360, A_5(1/2, 1/4) = (3720, 68448), \dots, A_{31}(1) = 1]$, with gwlp $(155, 1085, 5208, \dots, 1)$. Whereas the regular 2^{31-26} has $A_3(1) = 155$ (i.e., 155 subsets of 3 factors for which we have 8 replicates of a 2^{3-1}), the worst three-factor projections for this Paley nonregular OA(32, 2^{31} , 2) have frequencies of 3 and 5 for the 8 treatment combinations of the 2^3 . For projections into three or four columns, this nonregular design is much preferred. In fact, of the 14 nonregular OA(32, 2^{31} , 2) studied by Evangelaras, Kolaiti, and Koukouvinos (2006), this cyclic design by Paley had by far the best projectivity into four factors.

For the OA(36, 2^{35} , 2), Plackett and Burman (1946) proposed the design obtained by cycling the row $[- + - + + + - - + + + + + + - + - - + - + - + + - + -]$ and then appending a row of -1 's. The relevant portion of its cfv is

$$A_3(1/3, 1/9) = (1190, 5355),$$

$$A_4(0.\bar{6}, 0.\bar{4}, 0.\bar{3}, 0.\bar{2}, 0.\bar{1}) = (11, 1260, 1088, 20565, 4896).$$

Hence, the correlations between main effects and two-factor interaction columns are small, but 11 poor projections into 4 factors means that there exist 33 pairs of highly correlated two-factor interaction columns.

For larger designs, refer to Hedayat, Sloane, and Stufken (1999), Plackett and Burman (1946), or the websites mentioned earlier.

6.4 Optimal Nonorthogonal Saturated Main Effect Designs

There are two basic reasons for considering nonorthogonal designs to estimate main effects

1. Economy of run size is important, and one would prefer a nonorthogonal saturated design of size $N = k + 1$ rather than increasing N to the next multiple of 4. Budgetary constraints are so severe that one can only afford enough runs to estimate each main effect, and nothing more.
2. The number of factors to be investigated is large, but only a very few are expected to dominate the rest. The purpose of the experiment is to identify these few factors as economically as possible. This is the situation for considering a supersaturated design, where $k \geq N$. Since the number

of factors exceeds the degrees of freedom, these designs cannot have all main effect columns pairwise orthogonal.

This section will address saturated main effect designs when N is not a multiple of 4. Supersaturated designs are discussed in the following section.

Given the availability of $\text{OA}(N, 2^{N-1}, 2)$ for N a multiple of 4, we need only consider nonorthogonal saturated main effect designs where $N = k + 1$ is not a multiple of 4. Here, optimal design criteria will be invoked to select designs. This is the first of several occasions where optimal design concepts will prove useful for creating designs. Therefore, before listing specific saturated main effect designs, we provide a brief introduction to optimal design. For a book-length treatment, see Atkinson and Donev (1992) or Atkinson, Donev, and Tobias (2007).

For some two-level N -run design \mathbf{D} , let \mathbf{X} denote its model matrix for a particular model. For the main effects model (1.1), \mathbf{X} has $r = 1 + k$ columns, whereas for models with interactions, the number of columns r will be larger. For a full factorial design with coding ± 1 , $\mathbf{X}'\mathbf{X} = N\mathbf{I}_r$,

$$\det[(\mathbf{X}'\mathbf{X})^{-1}] = 1/|\mathbf{X}'\mathbf{X}| = N^{-r},$$

and $\text{trace}[(\mathbf{X}'\mathbf{X})^{-1}] = r/N$. If the design is not orthogonal so that $\mathbf{X}'\mathbf{X}$ is not a diagonal matrix, then

$$\det[(\mathbf{X}'\mathbf{X})^{-1}] > N^{-r}$$

and

$$\text{trace}[(\mathbf{X}'\mathbf{X})^{-1}] > r/N.$$

Recall that the variance–covariance matrix for the least squares estimator is $\sigma^2[\mathbf{X}'\mathbf{X}]^{-1}$. Thus, the variance efficiency of a design may be judged by computing

$$\text{D-eff} = \{N^{-r}/|[\mathbf{X}'\mathbf{X}]^{-1}|\}^{1/r} = |\mathbf{X}'\mathbf{X}|^{1/r}/N \quad (6.9)$$

or

$$\text{A-eff} = (r/N)/\text{trace}[(\mathbf{X}'\mathbf{X})^{-1}]. \quad (6.10)$$

A two-level design of size N that maximizes (6.9) is labeled a D-optimal design for this particular model, and the design that maximizes (6.10) is the A-optimal design. Other functions of $(\mathbf{X}'\mathbf{X})^{-1}$ are used to establish optimal design criteria. For example, E-optimality is based on the maximum eigenvalue of $(\mathbf{X}'\mathbf{X})^{-1}$. We now turn to the problem at hand: that of choosing the best possible main effects design \mathbf{D} when $N = k + 1$ is not a multiple of 4.

Useful surveys of nonorthogonal saturated designs are provided by Crosier (2000) and Evangelaras, Koukouvinos, and Stylianou (2005). These surveys evaluate the efficiency of alternative designs in terms of A-, D-, and E-efficiency and in terms of variance inflation factors. Of these, E-efficiency is the least relevant, given that the purpose of these designs is to estimate

individual factor effects. Variance inflation factors (VIFs) are easiest to interpret since, for the i^{th} factor of a two-level designs, $\text{Var}(b_i) = \text{VIF}_i \sigma^2 / N$. For the construction of these designs, the cases $N(\text{mod } 4) = 1, 2$, and 3 are often discussed separately. However, for practitioners, it makes sense to simply present the recommended designs in a single sequence. For $k = 4, 5$, and 6 , the saturated designs are 37.5%, 25%, and 12.5%, respectively, smaller than the orthogonal eight-run design. For $k = 8\text{--}10$, the savings are 25%–8.3%, while for $k = 16\text{--}18$, the savings are only 15%–5%, relative to the next largest orthogonal design. Here, we consider only cases with $k \leq 18$. For larger k , the reader is referred to the journal articles cited above.

Table 6.35. Summary for D-optimal saturated main effect designs

N	D-effA-e	ffVIF	$\sigma^2/\text{Var}(b_i)$
3	0.840	0.667	2@1.5
4	1	1	3@1.0
5	0.941	0.900	4@1.1
6	0.905	0.833	5@1.2
7	0.878	0.783	5@1.296, 1@1.16
8	1	1	7@1.0
9	0.932	0.871	1@1.469, 7@1.056
10	0.941	0.900	9@1.1
11	0.915	0.858	4@1.21, 6@1.1275
12	1	1	11@1.0
13	0.977	0.962	12@1.04
14	0.957	0.929	13@1.08
15	0.941	0.900	11@1.114, 3@1.1
16	1	1	15@1.0
17	0.966	0.943	16@1.04
18	0.967	0.944	17@1.06
19	0.953	0.913	6@1.105, 12@1.086
			17.199, 17.496

Table 6.35 provides a summary of recommended designs; we list the D-optimal design (according to current literature) with the highest A-efficiency. The most convenient means for constructing these designs for many users will be an optimal design search method. For example, for $N \leq 14$, JMP's "Custom Design" D-optimal search is able to quickly find a D-optimal design, whereas for $N = 15, 17$, and 19 , one must increase the number of starts to find the D-optimal design. By comparing D-eff and A-eff for the design produced by such an algorithm with the summary in Table 6.35, one may verify that the optimal design has actually been obtained. These calculations are illustrated below for the 6-run and 9-run saturated designs. The designs in Table 6.35 with even N are not equal-occurrence designs. The listed designs are given because they provide better precision for the factor effects. As a

typical example, consider the D-optimal design for $N = 6$ and $k = 5$:

$$\mathbf{D} = \begin{bmatrix} -1 & -1 & 1 & -1 & -1 \\ -1 & 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 \\ -1 & -1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 \\ -1 & 1 & -1 & -1 & -1 \end{bmatrix}.$$

Although columns 1 and 4 sum to -2 rather than to zero, variances for estimates of the first-order model coefficients are equal, since

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.20 & 0.05 & 0.00 & 0.00 & 0.05 & 0.00 \\ 0.05 & 0.20 & 0.00 & 0.00 & -0.05 & 0.00 \\ 0.00 & 0.00 & 0.20 & -0.05 & 0.00 & -0.05 \\ 0.00 & 0.00 & -0.05 & 0.20 & 0.00 & -0.05 \\ 0.05 & -0.05 & 0.00 & 0.00 & 0.20 & 0.00 \\ 0.00 & 0.00 & -0.05 & -0.05 & 0.00 & 0.20 \end{bmatrix},$$

where $\mathbf{X} = [\mathbf{1} \ \mathbf{D}]$. Each variance is $0.2\sigma^2$, so $VIF = 0.2N=1.2$. In addition, A-efficiency and D-efficiency for this saturated two-level designs are calculated using (6.9) and (6.10):

$$A\text{-eff} = (6/6)/\text{trace}[(\mathbf{X}'\mathbf{X})^{-1}] = 1/1.2 = 0.8\bar{3},$$

$$D\text{-eff} = |\mathbf{X}'\mathbf{X}|^{1/6}/6 = 25600^{1/6}/6 = 0.905.$$

The alias matrix for this design, reflecting the potential bias from omitting active two-factor interactions, is

Effect	x_1x_2	x_1x_3	x_1x_4	x_1x_5	x_2x_3	x_2x_4	x_2x_5	x_3x_4	x_3x_5	x_4x_5
Intercept	-0.2	-0.2	0.2	-0.2	0.6	-0.2	0.6	-0.2	0.6	-0.2
x_1	0.2	0.2	-0.2	0.2	0.4	-0.8	0.4	-0.8	0.4	-0.8
x_2	-0.6	0.4	-0.4	0.4	-0.2	-0.6	-0.2	0.4	0.8	0.4
x_3	0.4	-0.6	-0.4	0.4	-0.2	0.4	0.8	-0.6	-0.2	0.4
x_4	-0.8	-0.8	-0.2	-0.8	0.4	0.2	0.4	0.2	0.4	0.2
x_5	0.4	0.4	-0.4	-0.6	0.8	0.4	-0.2	0.4	-0.2	-0.6

These coefficients range in magnitude from 0.2 to 0.8, so every two-factor interaction is partially aliased with each first-order model coefficient.

If one restricts attention to equal-occurrence designs, the best design is the cyclic design

$$\mathbf{D} = \begin{bmatrix} -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 & -1 \\ -1 & 1 & 1 & 1 & -1 \\ -1 & -1 & -1 & -1 & -1 \end{bmatrix},$$

with

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.1667 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.3333 & -0.1667 & 0.0833 & 0.0833 & -0.1667 \\ 0.0000 & -0.1667 & 0.3333 & -0.1667 & 0.0833 & 0.0833 \\ 0.0000 & 0.0833 & -0.1667 & 0.3333 & -0.1667 & 0.0833 \\ 0.0000 & 0.0833 & 0.0833 & -0.1667 & 0.3333 & -0.1667 \\ 0.0000 & -0.1667 & 0.0833 & 0.0833 & -0.1667 & 0.3333 \end{bmatrix}.$$

Clearly, the equal-occurrence design yields less information regarding the five main effects. In addition, its alias matrix is worse, in that the largest coefficients are 4/3.

Results are similar for other cases with even N . Thus, for saturated main effect designs, the D-optimal designs are recommended, even though they lack equal occurrence. If the experimenter has more interest in one level over the other a priori for some factor, it seems reasonable to assign that factor to an unbalanced column and the level of interest to the level that occurs for more than half the runs.

For cases where the diagonals of $(\mathbf{X}'\mathbf{X})^{-1}$ are not equal, the largest diagonal is usually shifted to the intercept so that the variances for the main effect estimates are as small as possible. We now illustrate this calculation for the $N = 9$ design. The following design found by JMP has D-eff = 0.932 and A-eff = 0.871:

$$\mathbf{D} = \begin{bmatrix} -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 \\ -1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 \\ -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 \end{bmatrix}.$$

Appending an intercept column and computing $\text{diag}(\mathbf{X}'\mathbf{X})^{-1}$, one finds that the largest diagonals (0.1633) are associated with factors 2 and 8, and the other seven diagonals equal 0.1173. Multiplying the i^{th} entry in the other columns of \mathbf{D} by $d_{i,8}$ produces the design matrix

$$\begin{bmatrix} -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 \\ -1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 \\ -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ -1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 \end{bmatrix},$$

which is the D-optimal design with only one large VIF, as reported in Table 6.35.

The D-optimal designs with $N = 5$ and 9 are the most likely to be useful. Designs of sizes 3, 7, 11, 15, ... should rarely be used, since adding a single run will increase the efficiency markedly, and provide 1 df for error.

6.5 Supersaturated Designs

The first published papers discussing supersaturated designs appeared in *Technometrics* in 1959. Satterthwaite (1959) proposed the use of a small number of randomly selected treatment combinations from the 2^k . Such designs are named random balance designs and were espoused as a simple, general solution. In the same issue, five prominent statisticians Youden, Kempthorne, Tukey, Box, and Hunter (1959) expressed their reservations about using random balance designs. George Box (Youden et al. 1959, p. 180) argued that systematically determined supersaturated designs would outperform random balance designs of the same size and stated that systematic supersaturated designs were “not yet available only because they have not been looked for.” John Tukey sought to highlight the potential good from discussion of random balance designs. He, too, predicted that in place of random balance designs, “constant near balance patterns (designs) are going to appear, and are going to be used.”

In response to these suggestions, Booth and Cox (1962) proposed seven supersaturated designs found by computer search, with sizes $N = 12, 18$, and 24 and k as large as $2N$. Booth and Cox (p. 489) expressed reservations about the use of these new designs:

We have no experience of practical problems where such designs are likely to be useful; the conditions that the interactions should be unimportant and that there should be a few dominant main effects seem very severe.

Given such cautions, it is not surprising that the designs resulted in little use in practice. The alternative design methodology of multistage group-screening experiments (Watson 1961) proved to be more popular than supersaturated designs. However, group screening methods typically assume that one knows the direction of each effect. (Group screening experiments are discussed again in the context of a Chapter 11 case study.)

Although not necessarily a reflection of their nonuse, papers regarding the construction and analysis of supersaturated designs were absent for 30 years. Ending this period of inactivity, four construction methods were published between 1993 and 1996:

- Using half of a Plackett–Burman design (Lin 1993).

- Using a Plackett–Burman design, augmented with partially-aliased two-factor interactions as generators for additional factors (Wu 1993). Example 6.9 analyzed later is of this type.
- Construction based on balanced incomplete block designs (Nguyen 1996).
- Computer search for designs (Lin 1995, Nguyen 1996).

Numerous subsequent papers further developed the theory, showed the connections between various construction methods and emphasized the difficulties of analyzing such experiments. However, few published examples have appeared. We discuss this work in the subsequent subsections.

6.5.1 Optimality criteria for supersaturated designs

Let \mathbf{D} denote the $N \times k$ matrix for a supersaturated design with the usual ± 1 coding, and let $\mathbf{S} = \mathbf{D}'\mathbf{D}$ and $\mathbf{X} = [\mathbf{1} \ \mathbf{D}]$. Since $N < k + 1$ for supersaturated designs, $|\mathbf{X}'\mathbf{X}| = 0$, so standard optimality criteria are not useful. Booth and Cox (1962) proposed two criteria based on the off-diagonal elements s_{ij} ($1 \leq i < j \leq k$) of the matrix \mathbf{S} :

1. Minimize $\text{Max}_{i < j} |s_{ij}|$.
2. Minimize $E_{i < j}(s_{ij}^2)$, which for designs with equal occurrence of -1 and $+1$ is equivalent to minimizing the variance of the s_{ij} 's.

For brevity, we simply write $\text{Max } |s|$ and $\text{Min } E(s^2)$ to denote these criteria. As an example, consider the supersaturated design with $N = 6$ and $k = 10$ proposed by Lin (1993) and obtained by taking the six rows of Table 6.15 for which the 11th column is $+1$; see Table 6.36. Since the design is only strength 1, having correlated equal-occurrence columns, it is referred to as a near-orthogonal array (NOA) rather than an OA. In Example 6.5, when we considered evidence for factor effects conditional on $\mathbf{E} = -1$, we were actually analyzing data for six runs equivalent to Table 6.36.

Table 6.36. NOA(12, 2^{10} , 1) Design

1	2	3	4	5	6	7	8	9	10	11
1	1	1	1	1	1	1	1	1	1	1
-1	-1	1	-1	1	1	1	-1	-1	-1	1
-1	-1	-1	1	-1	-1	1	-1	1	1	1
1	-1	-1	-1	1	-1	-1	1	-1	1	1
1	1	-1	-1	-1	1	-1	-1	1	-1	1
-1	1	1	1	-1	-1	-1	1	-1	-1	1

Taking the first 10 columns as \mathbf{D} ,

$$\mathbf{S} = \mathbf{D}'\mathbf{D} = \begin{bmatrix} 6 & -2 & 2 & 2 & -2 & -2 & 2 & -2 & -2 & -2 \\ -2 & 6 & -2 & -2 & 2 & -2 & 2 & -2 & -2 & 2 \\ 2 & -2 & 6 & -2 & -2 & -2 & -2 & -2 & 2 & 2 \\ 2 & -2 & -2 & 6 & 2 & 2 & -2 & -2 & -2 & -2 \\ -2 & 2 & -2 & 2 & 6 & -2 & -2 & -2 & 2 & -2 \\ -2 & -2 & -2 & 2 & -2 & 6 & -2 & 2 & -2 & 2 \\ 2 & 2 & -2 & -2 & -2 & -2 & 6 & 2 & -2 & -2 \\ -2 & -2 & -2 & -2 & -2 & 2 & 2 & 6 & 2 & -2 \\ -2 & -2 & 2 & -2 & 2 & -2 & -2 & 2 & 6 & -2 \\ -2 & 2 & 2 & -2 & -2 & 2 & -2 & -2 & -2 & 6 \end{bmatrix}.$$

Since all of the off-diagonal elements $s_{ij} = \pm 2$, $E(s^2) = 4$. This design is both Max $|s|$ and $E(s^2)$ optimal. Table 6.36 is a special case, where $k = 2(N - 1)$; that is, this design has twice the number of effects as a saturated main effect design. Nguyen (1996) was the first to show that when $k = q(N - 1)$,

$$E(s^2) \geq N^2(k - N + 1)/[(N - 1)(k - 1)] = N^2(q - 1)/(k - 1). \quad (6.11)$$

For columns with equal occurrence of -1 and 1 , dividing s_{ij}^2 by N^2 produces the correlation between columns i and j . Thus, $(q - 1)/(k - 1)$ is a lower bound for the average squared correlation. For small N , the Max $|s|$ and Min $E(s^2)$ optimality criteria coincide. However, this is less common for larger N .

Supersaturated designs generally can only be used to identify a small number of effects, relative to the run size N . The six-run design in Table 6.36 permits estimation of any three main effects. However, many sets of four columns are linearly dependent (e.g., columns 1–3 and 5). For this reason, design measures based on projection into a small number of dimensions d are useful. Wu (1993) considered expected D-efficiency and A-efficiency, averaging across all subsets of d factors, with $d \leq 5$. Deng, Lin, and Wang (1999) ignored efficiency, focusing only on the percentage of d -factor main effects models that are estimable.

6.5.2 Supersaturated designs with $k \approx 2(N - 1)$

Lin (1993) proposed using half of a Hadamard design ($\frac{1}{2}\text{HD}$) and produced eight supersaturated designs with $k = 2(N - 1)$; see Table 6.37. Nguyen (1996) generalized the results of Lin (1993) to produce designs with $k = 2(N - 1)$ for every even N from 6 to 30. All of these designs are $E(s^2)$ optimal. Liu and Zhang (2000) found designs that decreased the maximum correlation (at $N = 20$) or decreased the frequency of the maximum correlation (at $N = 24$). Table 6.37 presents the generators for the Nguyen designs (or Liu and Zhang's improvements for $N = 20$ and 24), and this method of construction is illustrated below. The designs for $N = 16, 22, 26$, and 30 are especially attractive, given the small maximum correlations. All of the designs in Table 6.37 achieve the theoretical lower bound for the maximum correlation, except for $N = 24$ and 28.

Table 6.37. Maximum correlation for Lin's (1993) and Nguyen's (1996)
 $E(s^2)$ optimal supersaturated designs with $k = 2(N - 1)$

N	k	Max ρ	$E(\rho^2)$	Lin's Design	Nguyen's Generating Vectors
6	10	0.333	0.111	$\frac{1}{2}$ HD	(+---+) (-++-+)
8	14	0.500	0.077		(-+----+) (-++-+--)
10	16	0.600	0.059	$\frac{1}{2}$ HD	(++-+---+) (-++++-+--)
12	22	0.333	0.048	$\frac{1}{2}$ HD	(-+++-+---+) (+----+---+--)
14	26	0.429	0.040	$\frac{1}{2}$ HD	(-+++-+---+--) (++-+---+---+--)
16	30	0.250	0.034		(+----+---+---+--) (-+---+---+---+--)
18	34	0.333	0.030	$\frac{1}{2}$ HD	(---+---+---+---+--) (+---+---+---+---+--)
20	38	0.200	0.027		(+---+---+---+---+--) (+---+---+---+---+--)
22	42	0.273	0.024	$\frac{1}{2}$ HD	(+---+---+---+---+--) (-+---+---+---+---+--)
24	46	0.333	0.022	$\frac{1}{2}$ HD	(+---+---+---+---+--) (+---+---+---+---+--)
26	50	0.230	0.020		(-+---+---+---+---+--) (-+---+---+---+---+--)
28	54	0.285	0.019		(-+---+---+---+---+--) (+---+---+---+---+--)
30	58	0.200	0.018	$\frac{1}{2}$ HD	(+---+---+---+---+--) (+---+---+---+---+--)

Note: The $N = 20$ and 24 designs are from Liu and Zhang (2000).

Lin's (1993) construction of taking half a Hadamard design as we did in Table 6.36 is the simplest construction. When this method is not available, Nguyen's cyclic construction is as follows. Each generating vector, by repeated cycling, produces an $(N - 1) \times (N - 1)$ matrix. Appending a row of +1's to each, we obtain two $N \times (N - 1)$ matrices that, when combined, form the supersaturated design with $k = 2(N - 1)$. For $N = 16$, $(+ - + + + - - + - - - + -)$ produces columns 1–15,

$$\begin{bmatrix} 1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 \\ -1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 \\ -1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 \\ -1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ -1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 & -1 \\ -1 & 1 & -1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix},$$

and $(- + - - + + - + - - + + + -)$ produces columns 16–30,

$$\begin{bmatrix} -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 \\ -1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 & -1 \\ -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 \\ -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

This design has a maximum correlation between main effect columns of .25, an average squared correlation of $1/29$, and $B_2 = 15$. A similar design may be obtained by taking half of the Paley OA($32, 2^{31}, 2$) Hadamard design mentioned near the end of Section 6.3.

Designs with $k = 2(N - 1)$ and $N = 24$ and 28 may exist with maximum correlations of $1/6$ and $1/7$, respectively, but they are not known at this time. Lin (1995, Table 5) provided a design with $N = 24$ and $k = 33$ with a maximum correlation of $1/6$.

Cheng (1997) showed that one may drop (or add) one or two columns to these designs with $k = 2(N - 1)$ and still satisfy the $E(s^2)$ optimality criterion.

When dropping two columns, they must be orthogonal. When adding one or two columns, care must be taken to avoid increasing the maximum correlation.

6.5.3 Supersaturated designs with $k > 2N$

As mentioned earlier, $E(s^2)$ optimal designs may generally be obtained by adding one or two columns to designs in the previous section. What about larger numbers of factors? Here is a brief synopsis of the literature.

Wu (1993) proposed a supersaturated design with $N = 12$ and $k = 66$ by adding to the OA(12, 2^{11} , 2) the 55 two-factor interaction columns. This design achieves the lower bound (6.11) and so is $E(s^2)$ optimal. This design also contains the maximum number of columns for an equal-occurrence design with $N = 12$ and $|s_{ij}/N| \leq 1/3$ (see Cheng and Tang 2001).

Cheng (1997) unified the previous design construction literature by Lin (1993), Nguyen (1996), and Tang and Wu (1997), providing a general theory for $E(s^2)$ optimal supersaturated designs. Cheng (1997, Sect. 4) also contains a complete characterization of the case $N = 8$ and $k \leq 35$ using a construction based on balanced incomplete block designs. However, supersaturated designs with such small run sizes are of questionable utility. Liu and Zhang (2000), using a cyclic block construction method, presented additional designs for even $N = 10, 12, \dots, 24$ and k as large as $6(N - 1)$, although some have very large maximum correlations. Table 6.38 summarizes their larger designs. Only three of the designs achieve the lower bound for the maximum correlation in Cheng and Tang (2001); see Table 6.39. For N a multiple of 4, Liu and Zhang also presented $E(s^2)$ optimal designs for $k = 3(N - 1)$ and $5(N - 1)$ with the same maximum correlations as reported in Table 6.38.

Table 6.38. Liu and Zhang's (2000) $E(s^2)$ optimal supersaturated designs with $k = 4(N - 1)$ and $6(N - 1)$

N	k	Max ρ	Max ρ	
			Optimal?	$E(\rho^2)$
16	60	0.50	Yes	0.051
	90	0.50		0.056
18	68	0.33	Yes	0.045
	102	0.55		0.050
20	76	0.40	Yes	0.040
	114	0.40		0.044
22	84	0.45		0.036
24	92	0.50		0.033
	138	0.50		0.036

Table 6.39. Upper bound on number of factors k for equal-occurrence designs constrained by maximum $|s_{ij}|/N \leq \rho$; k underlined means bound has been achieved

N	Max ρ									
	1/11	1/9	1/7	1/6	1/5	1/4	1/3.6	1/3	1/2.3	1/2
6								<u>10</u>		
8										<u>35</u>
10						<u>12</u>				
12								<u>66</u>		
14				16					<u>182</u>	
16						75				
18		20						212		
20						76				
22	24						399			
24					61					

Liu and Dean (2004) obtained an $N = 20$, $k = 38$ supersaturated design with the same maximum correlation as that found by Liu and Zhang (2000) by cycling a length-38 row vector by two each time and then appending a final row. This is equivalent to creating a $k \times k$ matrix by usual cycling, taking every other row, and then appending a final row of +1's to form a design with $k/2 + 1$ rows. If one takes every third row, one obtains a design with $k/3 + 1$ rows. In this way, Liu and Dean produced an 18-run, 51-factor design with a maximum correlation of $1/3$. By taking every fourth row, they produced a 14-run, 52-factor design with maximum correlation $3/7$, which is better than the comparable design found by Liu and Zhang. Circulant designs are easy to construct. When they also provide a small maximum correlation, they are recommended.

What if the desired k is not near a multiple of $N - 1$ for which designs have been constructed. In such cases, one should generally resort to algorithms. Nguyen (1996) proposed a useful algorithm for finding NOAs, including supersaturated designs. Using a modification of Nguyen's (1996) NOA algorithm, Ryan and Bulutoglu (2007) were able to find 12-run designs that were both Max $|s|$ and $E(s^2)$ optimal for $k = 12\text{--}43$, $45\text{--}55$, and $64\text{--}66$. [For $k = 44$ and $56\text{--}63$, the theoretical lower bound for $E(s^2)$ was not achieved, but the best resulting designs were likely still acceptable.] Additional algorithms have been proposed by Li and Wu (1997) and Xu (2002).

JMP uses an alternative algorithm for constructing supersaturated designs. By specifying each main effect as "estimate if possible," a small quantity is added to the diagonal of $\mathbf{D}'\mathbf{D}$, making the use of a D-optimal search possible. This has a Bayesian motivation, where the quantity added represents the prior information that each coefficient is close to zero (Jones, Lin, and Nachtsheim 2008). The resulting designs are typically not equal-occurrence designs

even when N is even; however, as we learned in Section 6.4, equal-occurrence is not necessarily helpful, and these designs appear to perform well when the run size is 20 or more.

6.5.4 Marginally oversaturated designs

Deng, Lin, and Wang (1996, 1999) defined the resolution rank of a design to be r if every r -factor main effects model is estimable, but some projections with $r + 1$ factors have linearly dependent columns. Most supersaturated designs considered earlier in this section have resolution rank of $N/2$. In practice, the number of effects that can be identified successfully by supersaturated design is even less; $N/4$ is a useful rule of thumb. If more than this number of effects are expected, then one should consider increasing the number of runs. Designs with k only slightly larger than $N - 1$ were labeled marginally oversaturated designs by Deng, Lin, and Wang (1996). Even these designs can suffer from linear dependencies in some low-dimension projections. For instance, Deng, Lin, and Wang (1996) found that by adding one or two columns to the 12-run Hadamard design reduced the resolution rank from 11 to 9 or 7, respectively.

Literature on marginally oversaturated designs is limited. Here are a few useful references for systematic designs. (Design construction algorithms mentioned earlier should also perform well for most cases.)

- For $N = 10$, Lin (1995) found a 12-factor design with $|s_{ij}| = 2$ for all pairs of factors. Adding any more factors in an equal-occurrence design will increase $\text{Max } |s_{ij}|$ to 6.
- For $N = 12$, add any interaction column to the OA(12, 2^{11} , 2) for $k = 12$ and any two orthogonal interactions (e.g., **1*2** and **1*3**) for $k = 13$. For $k = 14\text{--}16$, Butler et al. (2001, p. 625) presented $E(s^2)$ optimal designs.
- When $N = 16$, begin with the OA(16, 2^{15} , 2) in Table 6.17 and add two-factor interactions that are partially aliased with main effects and orthogonal to one another.

When a subset of the factors are deemed more likely than the rest, choose N to be a multiple of 4, and assign the more likely factors to a set of orthogonal columns (see Yamada and Lin 1997).

6.5.5 Analysis of supersaturated designs

Several researchers have proposed methods for analyzing supersaturated designs and/or critiqued methods proposed by others. Summarizing the literature, in time order:

- Westfall, Young, and Lin (1998) recognized the excessive Type I error rates for the ordinary forward selection procedure and recommended adjustments to control the risk of declaring inactive effects to be active. They concluded, “Identification of significant variables in supersaturated

designs is very tricky, and many Type I and Type II errors are expected using forward variable selection.”

- Abraham, Chipman, and Vijayan (1999) warned that “The correlation structure inherent in supersaturated designs can obscure real effects or promote nonreal effects. Whatever analysis is used, this problem can occur, although all-subsets regression is preferable to stepwise regression.”
- Kelly and Voelkel (2000) found that Lenth-type t -tests (constructed from simple regression estimates for each factor) and stepwise regression both performed poorly, even when there are few real effects. They recommended: “Examine all possible subsets of effects of size 1, 2, ..., m , where m is ...at least as large as the maximum number of active effects suspected... Summarize each subset with a criterion (e.g., a cross-validated R^2 value). Select subsets for which the criterion is optimized or near optimal.”
- Beattie, Fong, and Lin (2002) recommended a two-stage Bayesian procedure, especially for applications where all-subsets regression is infeasible due to the huge number of possible models with m or fewer effects.
- Holcomb, Montgomery, and Carlyle (2003) reported an extensive evaluation of 5 analysis methods in terms of both Type I and Type II errors for 15 different supersaturated designs. For stepwise regression, they began with forward selection; once three or more variables are included, backward elimination is used to check for continued significance. This stepwise procedure performed best of the five methods for many cases, in terms of maximizing the proportion of selected factors that are in fact active. Also performing well in some cases was a simplified version of all-subsets regression; their “many models method” begins by sorting all two-factor models based on R^2 and identifying the factors that show up frequently in the set of better two-factor models.

The above listing is far from exhaustive. See also Li and Lin (2002), Lu and Wu (2004), and Koukouvinos and Stylianou (2005). The choice of analysis method and the relevance of these warnings and criticism depend on the objective of the experiment. For instance, one’s objective may be to identify with minimal cost the largest effect, or largest few effects, out of many factors (e.g., see Chen and Lin 1998). The following modest goal seems more realistic. If through a supersaturated design one can identify a subset of the factors for which the proportion of active factors (and the average effect size) is much larger than was the case for the original list of factors, then the design has served as a useful screen.

6.5.6 Examples of supersaturated designs

Five examples are mentioned briefly:

- Lin (1995): An AIDS computer model with 138 variables was explored in $N = 24$ runs, leading to the conclusion that 8 factors were active. These data are reanalyzed in Section 6.5.7.

- Bandurek (1999): A 20-factor robustness study for a vending machine was conducted in just $N = 8$ runs; which factors were considered more likely differed for the various responses. Interpretation of the results relied heavily on engineering knowledge. The machine was found to be quite insensitive to variation in these factors, except perhaps for the gas pressure's effect on drink volume.
- Bandurek (1999): A cake mix project with 28 factors and $N = 8$ cakes provided no clear conclusions about the factors.
- Cela and colleagues have published several papers espousing the use of supersaturated designs for applications where water or soil samples are combined into composite samples and then tested (Martinez, Cela et al. 2002; Martinez, Landin et al. 2002; Quintana et al. 2003; Carpintero et al. 2004; Pensado et al. 2004). In each case, the composite samples are created based on a 12-run supersaturated design. In the published examples, the concentrations are known for each individual sample, and so the parameters of each model are known.
- Li (2008) reported an application involving 120 computer runs of a financial model involving nearly 500 factors.

Other authors have analyzed half-Hadamard designs by taking half of an orthogonal design to test their analysis method. Such examples are not mentioned here, except for Vander Heyden et al. (2000), who advocated the use of supersaturated designs for robustness testing in chemistry applications. They found that the variance of the response based on a Hadamard design can generally be well approximated by conducting just half of the runs.

Two-stage group screening experiments provide an alternative means of exploring a large number of factors economically, provided one can assume the sign of any active effects is known when creating the design. The literature of good group screening applications is somewhat limited. In Chapter 11, we analyze a case study based on Rooda and van der Schilden's (1982) group screening example involving 29 factors. For two other case studies, see Vine, Lewis, Dean, and Brunson (2008) and Rahni, Ramdani, Candau, and Dalicieux (1997).

6.5.7 Example 6.9: Analysis of a supersaturated design

Lin (1995) described a 24-run supersaturated design used to investigate 138 different variables in the AIDS simulation model, iwgAIDS. The response variable is the AIDS incidence rate per 100,000 persons. The supersaturated design was constructed using a 24-run Hadamard design for x_1-x_{23} (see Table 6.40), and two-factor interactions as generators for the remaining 115 factors numbered as follows:

$$\begin{aligned}
x_{24}-x_{45}: \quad & x_{22+j} = x_1 * x_j, \quad j = 2, \dots, 23 \\
x_{46}-x_{66}: \quad & x_{43+j} = x_2 * x_j, \quad j = 3, \dots, 23 \\
x_{67}-x_{86}: \quad & x_{63+j} = x_3 * x_j, \quad j = 4, \dots, 23 \\
x_{87}-x_{105}: \quad & x_{82+j} = x_4 * x_j, \quad j = 5, \dots, 23 \\
x_{106}-x_{123}: \quad & x_{100+j} = x_5 * x_j, \quad j = 6, \dots, 23 \\
x_{124}-x_{138}: \quad & x_{117+j} = x_6 * x_j, \quad j = 7, \dots, 21
\end{aligned}$$

Table 6.40. First 23 factors and response for Lin's supersaturated design

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	y
+	+	+	+	+	-	+	-	+	-	+	+	-	-	+	-	-	-	-	-	-	-	22.61	
+	+	+	+	-	+	-	+	+	-	-	+	-	+	-	-	-	-	-	-	-	+	14.26	
+	+	+	-	+	+	-	+	+	-	-	+	-	+	-	-	-	-	-	-	-	+	58.42	
+	+	-	+	+	-	+	+	-	-	+	-	+	-	-	-	-	-	-	-	-	+	24.59	
+	-	+	-	+	+	-	+	-	-	+	-	+	-	-	-	-	-	-	+	+	+	10.28	
-	+	-	+	+	-	+	-	+	-	+	-	+	-	-	-	-	-	+	+	+	+	188.46	
+	-	+	+	-	+	-	+	-	-	+	-	-	-	+	+	+	+	+	+	+	-	22.68	
-	+	+	-	+	+	-	+	-	-	-	-	-	+	+	+	+	+	-	+	-	+	22.90	
+	+	-	+	+	-	+	-	-	-	-	-	+	+	+	+	+	-	+	-	-	-	52.04	
+	-	-	+	-	+	-	-	-	-	-	-	+	+	+	+	+	-	-	-	-	-	381.61	
-	-	+	+	-	+	-	-	-	-	-	-	+	+	+	+	+	-	-	-	-	-	16.22	
-	+	-	-	+	-	-	-	-	-	-	-	+	+	+	+	+	-	-	-	-	-	108.59	
+	+	-	-	+	-	-	-	-	-	-	-	+	+	+	+	+	-	-	-	-	-	98.05	
+	-	-	+	-	-	-	-	-	-	-	-	+	-	+	-	+	-	-	-	-	-	53.13	
-	-	+	-	-	-	-	-	-	-	-	-	+	+	-	+	-	-	-	-	-	-	83.41	
-	+	-	+	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-	13.59	
+	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	-	-	-	-	-	242.96	
-	+	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-	663.93	
+	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-	57.95	
-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-	177.49	
-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-	40.22	
-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-	52.23	
-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-	53.50	
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2463.24	

Lin (1995) reported using forward selection to identify active factors. One can replicate his results for the first seven steps. However, the eighth factor to enter using stepwise regression is $x_{71} = x_3 * x_8$ rather than $x_{76} = x_3 * x_{13}$; see Table 6.41. Although this gives the appearance of a useful model, it is actually a mirage. If one were to exclude these 8 columns and repeat forward selection for the remaining 130 factors, comparable R^2 values would be obtained. In fact, for seven and eight factors, a higher R^2 is actually obtained when the factors in Table 6.41 are excluded! Using a follow-up 2^{8-4} fraction for

the seven factors in Table 6.41 plus x_{76} , only three factors were statistically significant, with Lenth t statistics of 47.5, 22.5, and 2.83, for x_{129} , x_{13} and x_{118} , respectively, but b_{13} changed sign, from negative in the forward selection model to positive in the resolution IV fraction, and b_{118} is 1/100th as large in the follow-up data as in the stepwise model. Clearly, something more than a naive application of stepwise regression is required here.

Table 6.41. Lin's forward selection of factors for y = incidence rate

Step	Factor	p -Value	R^2
1	x_{118}	.1171	0.108
2	x_{25}	.0587	0.251
3	x_{129}	.0265	0.418
4	x_{13}	.0254	0.555
5	x_{91}	.0109	0.693
6	x_{93}	.0055	0.807
7	x_{86}	.0055	0.883
8	x_{71}	.0175	0.920

The first modification to Lin's analysis is to use a transformation of incidence rate as the response; see Figure 6.13. Using either the log or reciprocal transformation would avoid having any outliers that dominate the fit of the model. When one response value is far removed from the rest, no simple model will account for that variation. Second, one should use all-subsets regression to find the best k -factor models, for $k = 1, 2, \dots, m$, where $m = N/3$. If models with $R^2 > .95$ are obtained for models with $k < N/3$, there is usually no benefit in searching for models with larger R^2 .

The best models should be evaluated in terms of a global model test. Edwards (2008) advocates calculating global model p -values using a permutation test. A permutation test for the models with k factors is performed as follows. Suppose one has a model of interest fit to y with an $R^2 = .93$, obtained using all-subsets regression. Then shuffle the y values and compute R^2 for the best k -factor model using all-subsets regression. Do this shuffling and refitting 1000 times and determine what proportion of the 1000 R^2 values exceed .93. This proportion is the global test p -value for the model in question. Only when one or more of the models obtained by all-subset regression has an unusually high R^2 , compared to models obtained by the same method to the permuted y 's, do we have any assurance that our model actually accounts for systematic variation.

Finally, for any model with a small global p -value, one must perform tests for each individual term in the model. As proposed by Westfall, Young, and Lin (1998), a simple means of doing this is to use Bonferroni adjusted p -values, which multiplies each ordinary p -value times the number of eligible factors not in the model plus 1 for the factor in question. For instance, given 138 candidate

factors, p -values for terms in a k -factor model would be multiplied by $139 - k$. This Bonferroni adjustment is justified for testing individual terms in models selected by forward selection. For models selected by all-subsets regression, it is somewhat ad hoc and is likely to provide too little adjustment (see Edwards and Mee 2009).

Here, Lin (1995) used forward selection rather than all-subsets regression to select the model. Thus, for the permutation test, we shuffle the y values and refit a model using forward selection. Each of the models selected in Table 6.41 are unexceptional, in that larger R^2 values are routinely obtained when fitted to randomly shuffled y values. This is true, whether we are fitting a model to $y = \text{incidence rate}$ or $\ln(y)$. This lack of statistical significance is further confirmed using Bonferroni adjusted p -values. For instance, even the smallest p -value in Table 6.41 (0.0055 for x_{86}), when multiplied by the number of eligible columns at that step equals $132(0.0055) = 0.73$, an insignificant outcome.

From the iwgAIDS model data reported by Lin (1995), it appears that interaction effects cannot be ignored. Thus, it is not surprising that a design intended to find only a few dominant main effects is ineffective when many main effects and interactions are important. Lewis and Dean (2001) have proposed a group screening procedure when interactions are deemed possible. Such a design might be useful for this AIDS computer model.

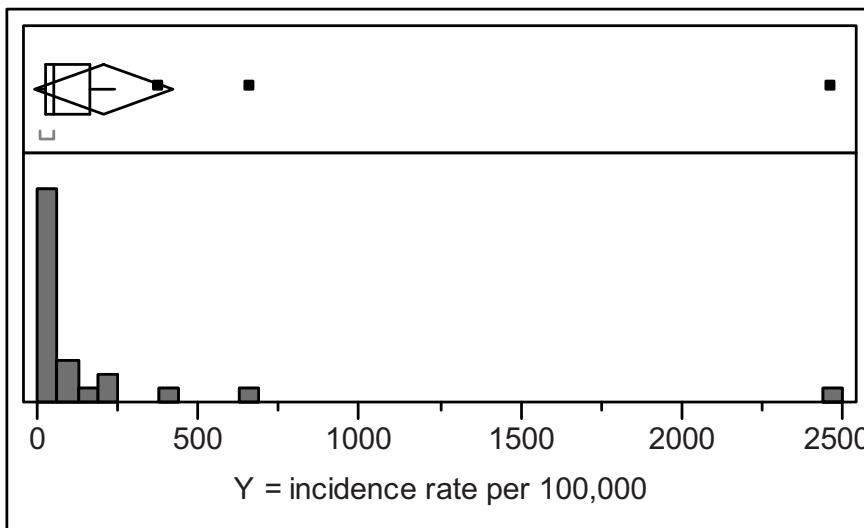


Fig. 6.13. Histogram of 24 y values for Lin's AIDS data

6.6 Conclusions

Resolution III designs are quite commonly used. Their success depends on a sparsity and simplicity of active effects, so that simple models suffice. The presence of one or two interactions can sometimes be detected, except for the case of regular fractional factorial designs with nearly as many factors as there are degrees of freedom. Because of the risk inherent in resolution III regular fractions or strength-2 nonregular designs, confirmation runs should always be performed. How best to do this is the topic of Section 9.2.

This chapter has emphasized minimum aberration 2^{k-f} fractions and Hadamard designs with minimum G-aberration. In general, these are the designs most useful. Two exceptions come to mind. First, occasionally blocking and other randomization restrictions lead to preferences for other fractions; see Chapter 10. Second, sometimes one has very detailed prior knowledge about the importance of specific terms and the objective is parameter estimation for a particular model, rather than factor screening. In such cases, designs constructed using a D-optimal algorithm are appropriate; the work of Hedayat and Pesotan (1992, 1997) is also relevant. Finally, sometimes a number of two-factor interactions are considered more likely than others, and it is desired to make the main effects orthogonal to these likely two-factor interactions. Using a regular resolution III fraction that aliases these likely two-factor interactions with one another but not with any main effect, one can create a design that provides efficient estimation of the main effects. In this case, one may create a resolution III design that is not minimum aberration but which is attractive for the specific problem at hand.

Designs for Estimating Main Effects and Some Two-Factor Interactions

This chapter presents regular resolution IV 2^{k-f} fractional factorial designs, strength-3 orthogonal arrays, and folded-over nonorthogonal designs. Provided three-factor and higher-order interactions are negligible, all of these designs provide unbiased estimates for main effects. These designs also devote at least $N/2 - 1$ degrees of freedom to estimating combinations of two-factor interactions; some designs will even have two-factor interactions clear of aliasing with other two-factor interactions. The sections of this chapter are as follows:

Section 7.1. Five Examples Analyzed

Section 7.2. Regular Resolution IV Designs

Section 7.3. Strength-3 Orthogonal Arrays

Section 7.4. Nonorthogonal Resolution IV Designs

Section 7.5. Recommendations Regarding Design Choice

To appreciate the difference between the designs discussed here and those in Chapters 6 and 8, consider the structure of the information matrix $\mathbf{X}'\mathbf{X}$. Let \mathbf{X}_1 denote the $N \times (k + 1)$ model matrix for a main effects model, and let \mathbf{X}_2 denote the $N \times k(k - 1)/2$ matrix of two-factor interaction contrasts. Then $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ is the model matrix for the two-factor interaction model (1.3). Using this partitioning we write the information matrix

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{X}'_1 \mathbf{X}_1 & \mathbf{X}'_1 \mathbf{X}_2 \\ \mathbf{X}'_2 \mathbf{X}_1 & \mathbf{X}'_2 \mathbf{X}_2 \end{bmatrix}. \quad (7.1)$$

For Chapter 6 designs, there is aliasing between main effects and two-factor interactions. In the current chapter, such aliasing is avoided by requiring the main effect columns to be orthogonal to two-factor interaction contrasts (i.e., $\mathbf{X}'_1 \mathbf{X}_2 = \mathbf{0}$). Ignoring three-factor and higher-order interactions, for designs in this chapter the only aliasing is among two-factor interactions. In order

to provide this clear estimation for main effects, for a given k , Chapter 7 designs require more runs than those in Chapter 6. (The designs in Chapter 8 will be larger still.) If one expects only a few two-factor interactions, then a Chapter 7 design will be suitable. For example, suppose one has seven factors. If one thought interactions were all negligible, one might use either the 2^{7-4} design or seven columns from the 12-run Plackett–Burman design, as we saw for Examples 6.2 and 6.6, respectively. Suppose instead that many two-factor interactions are expected to be important. To estimate main effects and all 21 two-factor interactions requires either 64 runs for an orthogonal design or 48 runs for a nonorthogonal three-eighths fraction, both of which are described in Chapter 8. Such run sizes are too large for many applications. We now consider a compromise between the frugal resolution III (and strength-2) Chapter 6 designs and the large designs of the next chapter. Chapter 7 will discuss regular resolution IV designs for 7 factors of sizes 16 and 32 runs, as well as a strength-3 24-run design. Although these designs do not permit estimation of the full two-factor interaction model, they do provide

$$\begin{bmatrix} \mathbf{X}_1' \mathbf{X}_1 & \mathbf{X}_1' \mathbf{X}_2 \\ \mathbf{X}_2' \mathbf{X}_1 & \mathbf{X}_2' \mathbf{X}_2 \end{bmatrix} = \begin{bmatrix} N\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2' \mathbf{X}_2 \end{bmatrix} \quad (7.2)$$

with $\text{rank}(\mathbf{X}_2' \mathbf{X}_2)$ between 7 and 18.

7.1 Five Examples Analyzed

We now present five examples to illustrate the analysis for various designs to be discussed in this chapter.

- Example 7.1: Su and Lua (2006) used a 2^{4-1} design to create eight batches of membranes and then tested the permeability of four gases with each batch. This is a popular resolution IV design, although we need additional information regarding the error variance, or some replication, to be able to determine statistical significance adequately.
- Example 7.2: Bafna and Beall (1997) conducted an experiment involving 6 factors in 16 runs, following an initial screening experiment that suggested the importance of these 6 factors.
- Example 7.3: Barnett, Czitrom, John, and León (1997) also conducted a 2^{6-2} , but with the inclusion of centerpoint replicates.
- Example 7.4: Choueiki, Mount-Campbell, and Ahalt (1997) utilized a resolution IV 2^{10-4} design. From this 64-run design, one may estimate most of the 45 two-factor interactions.
- Example 7.5: Yi, Lilja, and Hawkins (2005) conducted an even larger study of a computer architecture simulator, varying $k = 41$ factors in $N = 88$ runs. This design is a foldover of a 44-run orthogonal array. Analysis of such designs differs from that for regular fractions, since two-factor interactions are partially aliased.

Examples 7.1–7.3 are *even* regular resolution IV designs; this type is discussed in Section 7.2.2. Example 7.4 is an *even/odd* regular resolution IV design, and this type is discussed in Section 7.2.3. Section 7.3 presents orthogonal arrays such as the one employed in Example 7.5.

7.1.1 Example 7.1: The smallest resolution IV design: an unreplicated 2^{4-1}

Su and Lua (2006) described a carbon membrane experiment involving four factors, with the levels shown in Table 7.1. Eight batches of membranes were fabricated from Kapton®, according to the treatment combinations for a 2^{4-1} design. The eight batches of membranes were fabricated and the permeability measured in the order of the eight rows of Table 7.2 (per communication with the authors). For each treatment combination, three membranes were sampled, and the permeability of each membrane was measured for four gases: He, CO₂, O₂, and N₂. Thus, each permeation rate in Table 7.2 is a mean of three measurements, and the means in each row are obtained using the same three membranes.

Table 7.1. Factors and levels for Su and Lua experiment

Factors	Levels	
	-1	1
A Atmosphere	Nitrogen	Vacuum
B Temperature (K)	923	1073
C Heating rate (K/hr.)	30	240
D Thermal soak time (hr.)	2	5

Table 7.2. Permeability rates for four gases in 2^{4-1} experiment

Run	A	B	C	D	Permeability Rate			
					He	CO ₂	O ₂	H ₂
1	1	1	-1	-1	58.84	50.01	42.88	26.57
2	1	-1	1	-1	708.99	715.03	188.04	67.59
3	1	-1	-1	1	88.77	45.03	37.97	30.16
4	1	1	1	1	51.82	43.67	26.24	15.55
5	-1	1	1	-1	71.11	93.61	33.25	18.57
6	-1	1	-1	1	89.31	133.88	26.58	26.08
7	-1	-1	-1	-1	563.90	1056.83	304.22	111.69
8	-1	-1	1	1	571.48	684.17	157.69	21.89

Figure 7.1 plots the permeation rates on a logarithmic scale for each of the four gases. A log scale is used because the values differ by more than one

order of magnitude. The same two or three treatment combinations produced high permeability rates for all four gases. The gases are arranged in increasing molecular diameters. Except for helium, the permeability decreases as diameter increases.

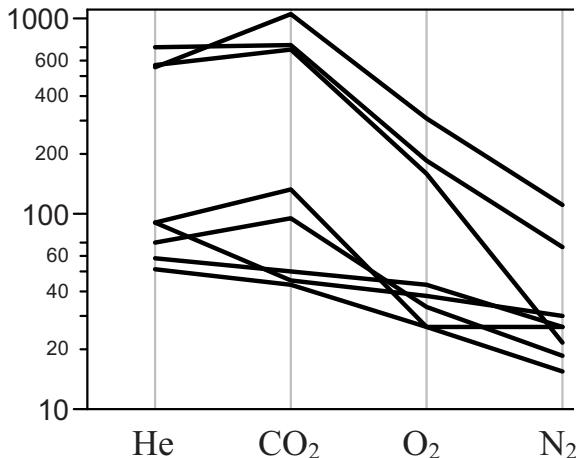


Fig. 7.1. Parallel plot of permeability rates for eight treatment combinations

We fit a saturated model and compute p -values for Lenth t statistics as described in Appendix C. For the O₂ permeability rate, the estimates are given in Table 7.3. Here, Lenth's PSE = 40.6, based on the median effect estimate $b_{AB} = 30.6$, and the largest estimate is only 1.52 times the PSE. For eight-run designs, Lenth's method has little power for detecting effects. Except when one or two terms explain most of the variation, having no significant effects is typical. Here, no simple model suffices. Without a prior estimate for the error variance σ^2 , one must interpret the estimates without being able to assess their precision. A negative estimate for **B** (Temperature) is reasonable. At 1073 K, the membrane visibly shrinks, which is expected to lower the permeability. Similarly, a negative coefficient for **D** is reasonable.

In conclusion, even though the resolution is better than for designs in the previous chapter, such a small experiment is of limited use unless one has a separate estimate of the error variance. When the choice of a model is not clear from an initial 2⁴⁻¹ design, it may be worthwhile running a second experiment containing the other half-fraction. Then the combined experiments form a full 2⁴ factorial in two blocks, as was discussed in Chapter 3.

Table 7.3. Saturated model for O₂ permeability rate

Term	Estimate	PSE	Lenth <i>t</i>	<i>p</i> -Value
Intercept	102.1			
A	-28.3	46.0	-0.62	.610
B	-69.9	46.0	-1.52	.130
C	-0.8	46.0	-0.02	.985
D	-40.0	46.0	-0.87	.327
AB = CD	30.6	46.0	0.67	.500
AC = BD	34.2	46.0	0.74	.406
AD = BC	-1.7	46.0	-0.04	.972

7.1.2 Example 7.2. A 2⁶⁻² with repeated determinations

Sixteen-run resolution IV designs for six to eight factors are very useful. Here we analyze such an experiment involving six factors. Bafna and Beall (1997) conducted a screening experiment involving many factors to assess the ruggedness of the measurement procedure for the melt index (MI). Approximately half of the factors in this preliminary experiment appeared to cause little or no measurement error variation and were dropped from further consideration. The experiment presented here is a subsequent 2⁶⁻² fraction involving the six factors that appeared to cause some variation in MI. Using a resolution IV design guarantees that two-factor interactions will not bias the estimates for main effects. Although this design provides 7 df for two-factor interactions, the chief interest is in main effects. Table 7.4 shows the names and levels for each factor, and Table 7.5 provides the run order for the 16 treatment combinations of this resolution IV fraction and the MI measurements for each. The three measurements do not represent replication of runs. Rather they are repeated measurements that we average together to obtain a mean MI that is more precise than if a single measurement had been obtained for each run.

Table 7.4. Factors and levels for Bafna and Beall experiment

Factors	Levels	
	-1	1
A Die orifice diameter (mm)	2.0930	2.1448
B Sample mass (g)	4	8
C Temperature (°C)	188.1	191.9
D Die cleanliness	Dirty	Clean
E Piston diameter (mm)	9.462	9.500
F Barrel cleanliness	Dirty	Clean

Bafna and Beall fitted a model for mean MI containing main effects and two-factor interactions, leaving 2 df for error. The MSE was 0.09785, producing a standard error for the coefficients of $[0.09785/16]^{1/2} = 0.078$. The estimates are provided in Table 7.6, where we show the aliasing among the 15 two-factor interactions. The largest five estimates are all main effects. These, plus the **BD** = **CF** interaction contrast, are statistically significant. Contrary to the initial screening experiment, the main effect for **E** is not important; that is, the measurement process is not affected by small differences in the Piston diameter. Perhaps this main effect was aliased with **BD** or **CF** in the initial experiment, and this biased the previous estimate for β_E . From this resolution IV design, we can confidently say that the other five factors do impact MI in an essentially additive manner. Reduction of measurement error will be facilitated especially by control of the Die orifice diameter (**A**) and Temperature (**C**).

Table 7.5. Treatment combinations for 2^{6-2} and MI measurements, with randomized run order

Run	A	B	C	D	E	F	MI ₁	MI ₂	MI ₃	Mean
1	-1	-1	1	1	1	-1	36.6	36.8	37.6	37.00
2	1	1	-1	-1	-1	1	38.9	39.5	39.6	39.33
3	-1	-1	-1	-1	-1	-1	31.9	32.3	31.9	32.03
4	1	1	1	1	1	1	42.5	41.9	41.6	42.00
5	1	1	-1	1	-1	-1	38.6	38.1	38.8	38.50
6	1	1	1	-1	1	-1	40.0	39.8	40.5	40.10
7	-1	-1	1	-1	1	1	34.8	35.5	34.8	35.03
8	-1	1	-1	1	1	-1	35.8	34.7	34.6	35.03
9	1	-1	-1	-1	1	-1	36.6	37.4	37.0	37.00
10	-1	-1	-1	1	-1	1	35.2	34.6	35.5	35.10
11	-1	1	1	1	-1	1	37.8	37.9	37.2	37.63
12	1	-1	1	-1	-1	1	40.7	40.9	39.0	40.20
13	1	-1	1	1	-1	-1	41.1	40.9	41.2	41.07
14	-1	1	-1	-1	1	1	36.0	36.0	35.1	35.70
15	-1	1	1	-1	-1	-1	36.6	35.4	35.3	35.77
16	1	-1	-1	1	1	1	38.9	39.4	39.5	39.27

The interaction coefficient of -0.37 is an estimate for $\beta_{BD} + \beta_{CF}$. By fitting a model first with **BD** and then with **CF**, one can view and interpret each possible interaction. Here, the conclusion is either that there is little or no Sample mass effect when the die is clean, or no Barrel cleanliness effect at high temperature. The authors do not discuss this interaction but instead devote their discussion to the each main effect estimate. They expected main effects to dominate, and chose a resolution IV design rather than resolution III to ensure the validity of the main effect estimates.

Table 7.6. Two-factor interaction model for mean MI

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	37.55	0.078	480.14	<.0001
A	2.14	0.078	27.31	.0013
B	0.46	0.078	5.89	.0277
C	1.05	0.078	13.45	.0055
D	0.65	0.078	8.34	.0141
E	0.09	0.078	1.20	.3534
F	0.49	0.078	6.21	.0250
AB = CE	-0.16	0.078	-2.05	.1767
AC = BE	0.11	0.078	1.36	.3072
AD = EF	-0.13	0.078	-1.63	.2456
AE = BC = DF	-0.19	0.078	-2.37	.1412
AF = DE	0.03	0.078	0.40	.7281
BD = CF	-0.37	0.078	-4.72	.0422
BF = CD	0.17	0.078	2.21	.1576

In Section 14.2, we analyze the data in Table 7.5 by fitting a saturated model and computing Lenth's PSE = 0.2156. Lenth's method would generally be preferred unless the assumption of effect sparsity is considered unreasonable; that is, if it is possible that a third or more of the effects are active, then Lenth's PSE can be severely biased upward and the power to detect active effects diminished. Another alternative is to modify Lenth's method to make it more robust (see Section 14.2). Here, it seems likely that the true standard error ($\sigma/4$) is between 0.078 and 0.216. The reduced model, either with or without the non-significant main effect for **E**, yields a standard error of 0.12–0.13.

The repeated determinations of MI enable us to precisely estimate the measurement error variance, which is one component of σ^2 . The measurement error variance, based on pooling the 16 sample variances of 2 df each, is 0.2548. From this, the measurement error variance for means of three values would be $0.2548/3 = 0.085$. The error variance (σ^2) for our response must contain this measurement error variance component plus any run-to-run variance. The MSE from the reduced model with six main effects and one interaction is 0.261. If this is an accurate estimate for σ^2 , then the measurement error variance accounts for about a third of our experimental error. If the researchers had not made repeated determinations of MI, the error variation would have been larger by approximately 0.17 ($= 0.2548 - 0.085$).

7.1.3 Example 7.3: Another 2^{6-2} with centerpoint replication

Uniformity is critical to profitability for manufacturing semiconductors. Commonly in that industry, experiments are run to identify factor settings that

minimize variability. Barnett, Czitrom, John, and León (1997) described such an experiment involving the etching of individual wafers. The objective of the experiment was to characterize the etching process for two etch target thicknesses, 50Å and 200Å. Table 7.7 lists this and five other factors; the actual low and high levels for the other factors were withheld for proprietary reasons. Target etch thickness is altered by varying the amount of anhydrous hydrofluoric acid flow. To achieve the required target etch amount, test runs were performed to determine the proper level of acid. It would have been simpler to use acid flow as a factor rather than target etch thickness.

Table 7.7. Factors and levels for Barnett et al. experiment

Factors	Levels	
	-1	1
A Target etch amount (Å)	50	200
B Etch N ₂ flow	Low	High
C Etch water vapor flow	Low	High
D Pre-etch N ₂ flow	Low	High
E Pre-etch water vapor flow	Low	High
F Wafer rotation speed (rpm)	Low	High

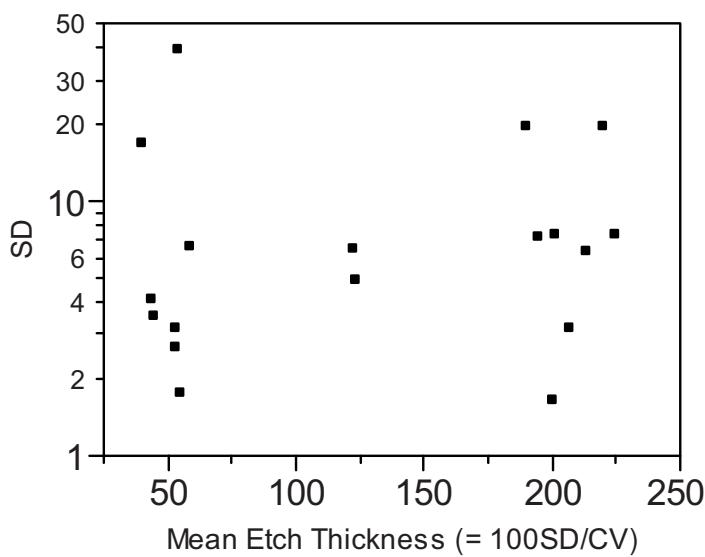
Table 7.8 lists the 18 treatment combinations for this experiment. The generators for the design are **E** = **ABD** and **F** = **ABC**. For each run, a single wafer was measured at nine fixed locations, etched, and then remeasured. The difference in thickness was determined for each location. From these nine differences, the mean, standard deviation, and coefficient of variation (CV) were calculated. The actual run order was not provided, so Table 7.8 is sorted by the standard deviation.

Barnett et al. (1997) took as their response the logarithm of the coefficient of variation. However, since CV was rounded to one decimal place, we use SD instead. Since SD (on a log scale) is uncorrelated with mean thickness (see Figure 7.2), there is no advantage to using the CV as the response.

Figure 7.3 shows the ANOVA from fitting a two-factor interaction model for ln(SD), which has $R^2 = .99$. The lack-of-fit test indicates no evidence for pure quadratic curvature (1 df) or three-factor interactions (2 df). Table 7.9 lists the estimates for this model, sorted by their *p*-value. Five factorial effects stand out, although four of these are strings of aliased interactions.

Table 7.8. Treatment combinations for Example 7.3, sorted by the within-wafer standard deviation

A	B	C	D	E	F	SD	CV
1	1	1	-1	-1	1	1.60	0.8
-1	-1	1	-1	-1	1	1.71	3.1
-1	1	1	1	-1	-1	2.61	4.9
-1	1	1	-1	1	-1	3.10	5.8
1	1	1	1	1	1	3.10	1.5
-1	-1	1	1	1	1	3.43	7.7
-1	-1	-1	1	1	-1	4.05	9.2
0	0	0	0	0	0	4.83	3.9
1	1	-1	1	1	-1	6.18	2.9
0	0	0	0	0	0	6.37	5.2
-1	-1	-1	-1	-1	-1	6.49	11.0
1	-1	-1	1	-1	1	7.00	3.6
1	1	-1	-1	-1	-1	7.20	3.2
1	-1	-1	-1	1	1	7.25	3.6
-1	1	-1	1	-1	1	16.58	41.0
1	-1	1	-1	1	-1	19.11	8.7
1	-1	1	1	-1	-1	19.20	10.1
-1	1	-1	-1	1	1	38.50	71.0

**Fig. 7.2.** Plot of the within-wafer SD versus mean etch thickness

Analysis of Variance

<u>Source</u>	<u>d.f.</u>	<u>Sum of Squares</u>	<u>Mean Square</u>	<u>F Ratio</u>
Model	13	12.57170	0.96705	32.969
Error	4	0.11733	0.02933	Prob > F
C. Total	17	12.68903		0.0020

Lack Of Fit

<u>Source</u>	<u>d.f.</u>	<u>Sum of Squares</u>	<u>Mean Square</u>	<u>F Ratio</u>
Lack Of Fit	3	0.07903	0.02634	0.688
Pure Error	1	0.03830	0.03830	Prob > F
Total Error	4	0.11733		0.6856

Fig. 7.3. ANOVA corresponding to the two-factor interaction model for $\ln(\text{SD})$ **Table 7.9.** Two-factor interaction model for $\ln(\text{SD})$, sorted by p -value

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	1.791	0.040	44.37	.000
AB = CF = DE	-0.497	0.043	-11.600	.000
AF = BC	-0.440	0.043	-10.277	.001
C	-0.380	0.043	-8.881	.001
AC = BF	0.353	0.043	8.251	.001
CD = EF	0.168	0.043	3.930	.017
E	0.111	0.043	2.595	.060
A	0.103	0.043	2.396	.075
F	-0.082	0.043	-1.904	.130
CE = DF	0.079	0.043	1.856	.137
AD = BE	0.079	0.043	1.851	.138
B	-0.058	0.043	-1.348	.249
AE = BD	-0.044	0.043	-1.022	.365
D	-0.019	0.043	-0.455	.673

In Section 9.5, we present a follow-up experiment that Barnett et al. (1997) conducted to clarify which of the aliased interactions are active. Here, we attempt to interpret the results from this experiment alone by making a simplifying assumption. It appears from Table 7.9 that factors **B** and **D**, the two N_2 flow variables, may have little effect. If these two factors are ignored, all of the aliasing disappears. Fitting a two-factor interaction model in the remaining 4 factors, we obtain a model with 10 terms that explains 97.8% of the variation in $\ln(\text{SD})$. From this fitted model we obtain predicted $\ln(\text{SD})$ values and display these in a cube plot (Figure 7.4). The noteworthy feature is that for both Target etch levels, the minimum $\ln(\text{SD})$ is achieved at **C** = 1, **E** = -1, **F** = 1. Thus, our tentative conclusion is that rotating wafers with

a high rpm and setting the water vapor flow low in the pre-etch period but high during etching, the best uniformity is achieved. Even though this combination coincides with our best two observed runs, some confirmation of these tentative conclusion is needed. We will revisit this conclusion in Chapter 9 when discussing follow-up experimentation. (In fact, follow-up runs revealed that one cannot simply ignore factors **B** and **D**.)

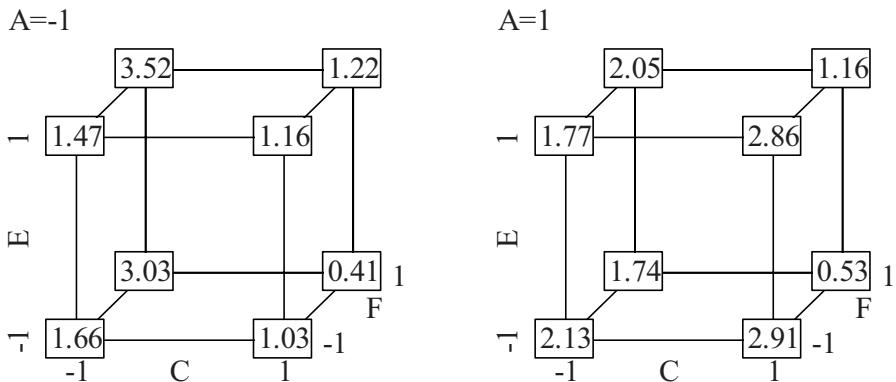


Fig. 7.4. Cube plot for predicted $\ln(\text{SD})$ for two-factor interaction model in four factors

7.1.4 Example 7.4: A larger resolution IV design with more active interactions

Choueiki, Mount-Campbell, and Ahalt (1997) used a resolution IV 2^{10-4} design to investigate the architecture and training of a neural network to predict short-term load requirements for an Ohio electric utility. The 10 factors investigated are listed in Table 7.10. The last two factors are characteristics of the load being forecast rather than factors of the neural network; that is, each combination of factors $\{\mathbf{J}, \mathbf{K}\}$ corresponds to a different electric utility's data. A lack of interactions between the neural network factors and these time series factors would indicate that the same network design is preferred, regardless of the situation to which it is applied.

The response is the root mean squared error (RMSE) for hourly predictions, expressed as a percentage of the average usage for a full year (1993). Usually we apply the log transformation to standard deviations and variances. However, here the ratio max/min is less than 2, so a transformation would have little effect except to complicate the analysis. The response data for the design appear in Table 7.11. This is a 1/16th fraction of a 2^{10} , based on the generators $\mathbf{G} = \mathbf{BCDF}$, $\mathbf{H} = \mathbf{ABDE}$, $\mathbf{J} = \mathbf{ACDF}$, and $\mathbf{K} = \mathbf{ABCE}$.

These generators produce a minimum aberration fraction, with length-4 words **ABGJ** and **CDHK**. This particular fraction was chosen because the few two-factor interactions it aliases were believed to be negligible.

Table 7.10. Factors and levels for Choueiki et al. experiment

Factors	Levels	
	-1	1
A Hidden layers	1	2
B Transfer function in output layer	Linear	Sigmoid
C Transfer function in hidden layer	Sigmoid	Sinusoid
D Backpropagation learning algorithm	Standard	Cumulative
E Gaussian noise added	No	Yes
F Stopping rule	RMSE	CD
G Network	Feedforward	Recurrent
H Years of training data	2	4
J Time of peak	Winter	Summer
K Industrial load %	Low	High

Figure 7.5 plots the 64 RMSE values, revealing how much larger one value is than the rest. We will fit models both with and without this value, to see if it alters the levels of the factors deemed optimal for minimizing the RMSE.

For this 2^{10-4} design, there are 10 df for main effects, 39 df for two-factor interactions, and the remainder ($63 - 49 = 14$ df) for three-factor interactions not aliased with lower-order terms. The sum of squares and mean squares are shown in Table 7.12. Without the maximum observation, the total sum of squares is reduced by 20%, and the mean squares are 0.2478, 0.0719, and 0.0402, respectively. This initial ANOVA supports the researchers' expectation that three-factor or higher-order interactions would be unimportant.

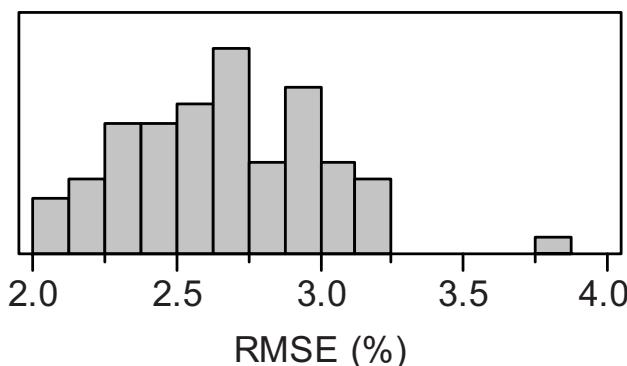


Fig. 7.5. Histogram of RMSE data from Choueiki et al. (1997)

Table 7.11. Treatment combinations for 2^{10-4} with % root mean square error for hourly predictions

A = -1										A = 1									
B	C	D	E	F	G	H	J	K	RMSSE	B	C	D	E	F	G	H	J	K	RMSSE
-1	-1	-1	-1	-1	1	1	1	1	2.1834	-1	-1	-1	-1	-1	1	-1	-1	-1	2.2931
1	-1	-1	-1	-1	-1	-1	1	-1	2.9520	1	-1	-1	-1	-1	-1	1	-1	1	2.9151
-1	1	-1	-1	-1	-1	1	-1	-1	2.4686	-1	1	-1	-1	-1	-1	-1	1	1	2.6055
1	1	-1	-1	-1	1	-1	-1	1	2.3955	1	1	-1	-1	-1	1	1	1	-1	2.0088
-1	-1	1	-1	-1	-1	-1	-1	1	2.9930	-1	-1	1	-1	-1	-1	1	1	-1	2.5495
1	-1	1	-1	-1	1	1	-1	-1	3.0491	1	-1	1	-1	-1	1	-1	1	1	3.0309
-1	1	1	-1	-1	1	-1	1	-1	2.6417	-1	1	1	-1	-1	1	1	-1	1	2.4679
1	1	1	-1	-1	-1	1	1	1	2.6906	1	1	1	-1	-1	-1	-1	-1	-1	3.0283
-1	-1	-1	1	-1	1	-1	1	-1	2.9295	-1	-1	-1	1	-1	1	1	-1	1	2.4878
1	-1	-1	1	-1	-1	1	1	1	3.8682	1	-1	-1	1	-1	-1	-1	-1	-1	3.1891
-1	1	-1	1	-1	-1	-1	-1	1	2.7197	-1	1	-1	1	-1	-1	1	1	-1	2.6658
1	1	-1	1	-1	1	1	-1	-1	2.2678	1	1	-1	1	-1	1	-1	1	1	3.2493
-1	-1	1	1	-1	-1	1	-1	-1	2.9744	-1	-1	1	1	-1	-1	-1	1	1	3.0965
1	-1	1	1	-1	1	-1	-1	1	2.6687	1	-1	1	1	-1	1	1	1	-1	2.6411
-1	1	1	1	-1	1	1	1	1	2.3445	-1	1	1	1	-1	1	-1	-1	-1	2.8845
1	1	1	1	-1	-1	1	1	-1	2.6009	1	1	1	1	-1	-1	1	-1	1	2.4098
-1	-1	-1	-1	1	-1	1	-1	1	2.6580	-1	-1	-1	-1	1	-1	-1	1	-1	2.8995
1	-1	-1	-1	1	1	-1	-1	-1	2.5533	1	-1	-1	-1	1	1	1	1	1	2.3149
-1	1	-1	-1	1	1	1	1	-1	2.5566	-1	1	-1	-1	1	1	-1	-1	1	2.6574
1	1	-1	-1	1	-1	-1	1	1	2.6208	1	1	-1	-1	1	-1	1	-1	-1	2.4210
-1	-1	1	-1	1	1	-1	1	1	3.2336	-1	-1	1	-1	1	1	1	-1	-1	2.8831
1	-1	1	-1	1	-1	1	1	-1	2.5910	1	-1	1	-1	1	-1	-1	-1	1	2.4787
-1	1	1	-1	1	-1	-1	-1	-1	3.1318	-1	1	1	-1	1	-1	1	1	1	2.7916
1	1	1	-1	1	1	1	-1	1	2.2120	1	1	1	-1	1	1	-1	1	-1	2.3436
-1	-1	-1	1	1	-1	-1	-1	-1	2.8580	-1	-1	-1	1	1	-1	1	1	1	2.7270
1	-1	-1	1	1	1	1	-1	1	2.1099	1	-1	-1	1	1	1	-1	1	-1	2.3727
-1	1	-1	1	1	1	-1	1	1	2.8065	-1	1	-1	1	1	1	1	-1	-1	2.6814
1	1	-1	1	1	-1	1	1	-1	2.2139	1	1	-1	1	1	-1	-1	1	1	2.3256
-1	-1	1	1	1	1	1	1	-1	2.5365	-1	-1	1	1	1	1	-1	-1	1	2.6597
1	-1	1	1	1	-1	-1	1	1	2.8636	1	-1	1	1	1	-1	1	-1	-1	2.7835
-1	1	1	1	1	-1	1	-1	1	2.9233	-1	1	1	1	1	-1	-1	1	-1	3.0618
1	1	1	1	1	1	-1	-1	-1	2.2124	1	1	1	1	1	1	1	1	1	2.1073

Table 7.12. ANOVA for Example 7.4

Source	df	SS	MS
Main effects	10	2.5574	0.2557
Two-factor interactions	39	4.0485	0.1038
Remainder	14	0.6568	0.0469
Total (corrected)	63	7.2626	

One could begin by fitting either the two-factor interaction model or a saturated model. For the two-factor interaction model, estimates with t statistics of 2 or more are shown in Table 7.13. Here, the standard error for each estimate is $0.027 = (0.0469/64)^{1/2}$, where the MSE comes from Table 7.12. Choueiki et al. (1997) began by fitting a simpler model, one that omitted two-factor interactions between network and time series factors; that is, they assumed the network effects would not depend on either factor **J** or **K**. The MSE for these two models are virtually identical; only the degrees of freedom for error are different. Alternatively one may fit a saturated model, obtaining Lenth's PSE = 0.029, which is also in agreement with the standard errors obtained by omitting 14 (or 24) interactions. When the saturated model is fit, 1 of the 14 three-factor interactions appears among the set of significant estimates; the estimate for **BJK** = **DEF** = **AGK** is 0.070, making it the seventh largest factorial effect. We will interpret this effect below, as we discuss the other estimates.

We discuss the estimates before examining a residual plot, because our understanding of the estimates will affect the reduced model we select. The estimates in Table 7.13 indicate that several factors prominently affect the RMSE of prediction. The large negative effects for **G**, **H**, and **C** indicate that the high level is preferred in each case—that is, a recurrent network, 4 years of training data, and a sinusoid function in the hidden layer. The largest effect is the **BF** interaction; given that b_{BF} is negative, we prefer **B** = **F**, and given that the coefficients for **BC**, **B**, and **F** are all negative, the optimal combination is **B** = **F** = 1 (i.e., a sigmoid transfer function for the output layer) and the CD stopping rule. [This stopping rule minimizes the RMSE on an independent data set; see Choueiki et al. (1997) for details.] Regarding factors **D** and **E**, any combination besides $(\mathbf{D}, \mathbf{E}) = (1, -1)$ appears to be acceptable.

Table 7.13. Largest estimates from two-factor interaction model for Example 7.4

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	2.670	0.027	98.606	.000
BF	-0.144	0.027	-5.310	.000
G	-0.114	0.027	-4.208	.001
H	-0.091	0.027	-3.378	.005
C	-0.091	0.027	-3.364	.005
JK	0.087	0.027	3.213	.006
BC	-0.074	0.027	-2.724	.016
DE	-0.067	0.027	-2.491	.026
B	-0.061	0.027	-2.240	.042
EF	-0.060	0.027	-2.209	.044
F	-0.058	0.027	-2.124	.052

The four combinations of Peak season (**J**) and Industrial load (**K**) correspond to hourly load time series for different public utilities. It was surprising to the authors that the **JK** interaction would be significant, even though neither main effect is. By re-examining the data, it was discovered that the two utilities for which **JK** = −1 reported “native load data” and the time series for the two utilities with **JK** = 1 also included “interruptible customer load.” Thus, we define the (unintended) factor **L** = **JK**. When **L** = 1, the RMSE is larger, because it is a more difficult task to forecast the occasionally clipped interruptible load. By including **L**, the design becomes a resolution III 2^{11-5} fraction, with one length-3 word (**JKL**) and four length-4 words (**ABGJ**, **CDHK**, **CEGL**, **AFHL**).

If we fit the two-factor interaction model in all 11 factors, 7 of the 11 main effects plus 8 two-factor interactions are statistically significant at $\alpha = .05$. In addition to **L** = **JK**, no estimate with aliased two-factor interactions is significant. Thus, the aliasing of effects does not appear to cause ambiguity, even for this unintended resolution III fraction.

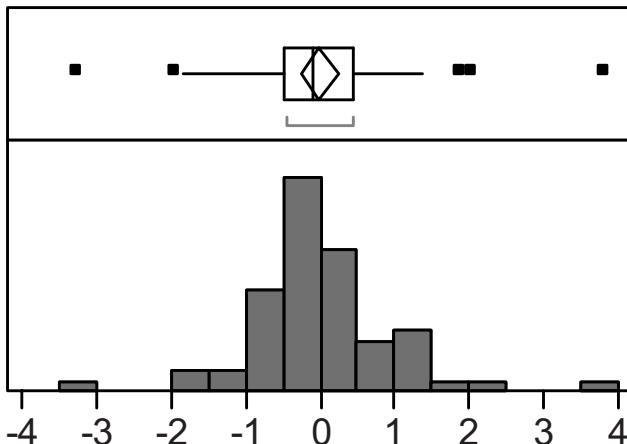
Including the main effects for **E** and **J** to make the model hierarchical, we obtain the reduced model in Table 7.14 and examine its Studentized residuals (see Figure 7.6). For this model, the combination of high levels for factors **B**–**G** minimizes the RMSE, with a predicted value of 2.0%, averaging across the levels of **J** and **L**. The only significant interactions between network and time series factors are **BL** (= **JKL**) and **HJ**. These are interpreted as follows. The preference for **B** = 1 (sigmoid transfer function in output layer) is greater when forecasting native load data. The preference for **H** = 1 (4 years of training data) is strongest for summer peak load data. The authors reasoned that, generally, 2 years of training data should be enough. However, both 1989 and 1993 were unusually hot years. In the researchers’ views, this similarity exaggerated the benefits of using 4 years of training data, especially for the summer peak load series.

Table 7.14. Reduced model for Example 7.4

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	2.670	0.021	127.230	0.000
B	-0.061	0.021	-2.891	0.006
C	-0.091	0.021	-4.340	0.000
D	0.045	0.021	2.166	0.036
E	0.025	0.021	1.207	0.234
F	-0.058	0.021	-2.741	0.009
G	-0.114	0.021	-5.430	0.000
H	-0.091	0.021	-4.359	0.000
J	0.021	0.021	0.996	0.324
L	0.087	0.021	4.145	0.000
BC	-0.074	0.021	-3.515	0.001
BD	-0.048	0.021	-2.265	0.028
BF	-0.144	0.021	-6.852	0.000
BL	0.070	0.021	3.317	0.002
CF	0.046	0.021	2.172	0.035
DE	-0.067	0.021	-3.214	0.002
EF	-0.060	0.021	-2.851	0.007
HJ	-0.050	0.021	-2.369	0.022

The residual plot for this reduced model is acceptable, except for two extreme residuals, one positive and one negative, with Studentized values 3.81 and -3.28. With these two values omitted, the reduced model's R^2 increases from 82% to 88%, the preferred combination of factors **B–H** is unchanged, and the predicted optimal RMSE is still 2.0%. The same model for RMSE is selected if we refit the model using stepwise regression (including all main effects and two-factor interactions as eligible terms), with the exception that the **CF** interaction is no longer statistically significant.

The only observation at the recommended treatment combination $\mathbf{B} = \dots = \mathbf{H} = 1$ is for the interruptible load series, $\mathbf{J} = \mathbf{K} = \mathbf{L} = 1$; its observed RMSE was 2.11, which is the second best observation out of 64. The best observed RMSE of 2.01 was for predicting one of the native load series ($\mathbf{L} = -1$), where our model estimates that the optimum treatment combination would achieve $\text{RMSE} = 1.80\%$. This prediction could be verified by using the optimal network settings to predict load for the $\{\mathbf{J} = 1, \mathbf{K} = -1\}$ series. One should also validate this performance by applying the recommended settings for the neural network to other time series. By varying \mathbf{H} and predicting load for various years, one might also resolve the question about how many years of training data are beneficial.



Studentized residuals for RMSE

Fig. 7.6. Box-plot and histogram for Studentized residuals from reduced model for RMSE

This 64-run example differs from Examples 7.1–7.3 in several respects. First, the larger number of observations facilitates estimation of standard errors for the effects (although the error here is not experimental error, but lack-of-fit of neural network models to four sampled time series). Second, the larger design makes it possible to estimate all of the two-factor interactions of interest. The design provided this benefit, since by the careful choice of generators, only interactions thought to be negligible were aliased. Finally, given the large design, one could even lose a couple of observations without seriously impacting model selection and estimation.

7.1.5 Example 7.5: Analysis of strength-3 orthogonal arrays obtained by foldover

Yi et al. (2005) investigated improving computer architecture by varying 41 parameters in a simulator model for a superscalar processor's performance. A folded-over Plackett–Burman design with $2 \times 44 = 88$ runs was used to select the treatment combinations. The particular design used is listed in Tables 7.15 and 7.16. Each table contains an $OA(44, 2^{43}, 2)$; combined, they form a strength-3 design. Here, only the first 41 columns are assigned factors.

Table 7.15. Half of the folded-over Plackett–Burman design used by Yi et al. (2005), with cycles per instruction (CPI) for first benchmark computing task

No.	Treatment combination	CPI
1	++++-+---++-++-+-----+-----+-----+-----+	1.948
2	-+---+---+---+---+-----+-----+-----+-----+	0.703
3	+---+---+---+---+-----+-----+-----+-----+	0.837
4	+---+---+---+---+-----+-----+-----+-----+	1.586
5	-+---+---+---+-----+-----+-----+-----+	1.295
6	+---+---+---+---+-----+-----+-----+-----+	1.083
7	-+---+---+---+-----+-----+-----+-----+	1.933
8	+---+---+---+-----+-----+-----+-----+	1.054
9	+---+---+---+-----+-----+-----+-----+	1.068
10	-+---+---+---+-----+-----+-----+-----+	1.458
11	-+---+---+---+-----+-----+-----+-----+	1.206
12	----+---+---+---+-----+-----+-----+-----+	0.887
13	+---+---+---+---+-----+-----+-----+-----+	0.664
14	-+---+---+---+-----+-----+-----+-----+	1.026
15	-+---+---+---+-----+-----+-----+-----+	0.944
16	----+---+---+---+-----+-----+-----+-----+	1.335
17	----+---+---+-----+-----+-----+-----+	1.453
18	----+---+---+-----+-----+-----+-----+	1.034
19	+---+---+---+-----+-----+-----+-----+	1.135
20	+---+---+---+-----+-----+-----+-----+	1.176
21	+---+---+---+-----+-----+-----+-----+	1.872
22	-+---+---+---+-----+-----+-----+-----+	1.972
23	+---+---+---+-----+-----+-----+-----+	1.604
24	-+---+---+---+-----+-----+-----+-----+	1.027
25	-+---+---+---+-----+-----+-----+-----+	1.407
26	-+---+---+---+-----+-----+-----+-----+	1.431
27	+---+---+---+-----+-----+-----+-----+	1.524
28	+---+---+---+-----+-----+-----+-----+	1.513
29	+---+---+---+-----+-----+-----+-----+	1.295
30	+---+---+---+-----+-----+-----+-----+	1.624
31	+---+---+---+-----+-----+-----+-----+	1.150
32	-+---+---+---+-----+-----+-----+-----+	0.551
33	+---+---+---+-----+-----+-----+-----+	1.223
34	+---+---+---+-----+-----+-----+-----+	1.012
35	+---+---+---+-----+-----+-----+-----+	1.649
36	-+---+---+---+-----+-----+-----+-----+	1.086
37	-+---+---+---+-----+-----+-----+-----+	1.984
38	+---+---+---+-----+-----+-----+-----+	1.085
39	-+---+---+---+-----+-----+-----+-----+	0.746
40	+---+---+---+-----+-----+-----+-----+	0.666
41	-+---+---+---+-----+-----+-----+-----+	0.761
42	-+---+---+---+-----+-----+-----+-----+	1.363
43	+---+---+---+-----+-----+-----+-----+	1.142
44	-+---+---+---+-----+-----+-----+-----+	3.157

Table 7.16. Other half of the folded-over Plackett–Burman design used by Yi et al. (2005), with CPI

No.	Treatment combination	CPI
45	-----+-----+-----+-----+-----+-----+-----+	1.261
46	+---+---+---+---+---+---+---+---+-----+	1.506
47	-+---+---+---+---+---+---+---+---+-----+	1.494
48	--+---+---+---+---+---+---+---+---+-----+	1.034
49	+-+---+---+---+---+---+---+---+---+-----+	0.891
50	+-+---+---+---+---+---+---+---+---+-----+	1.250
51	+---+---+---+---+---+---+---+---+-----+	1.019
52	-+---+---+---+---+---+---+---+---+-----+	1.783
53	--+---+---+---+---+---+---+---+---+-----+	1.548
54	+-+---+---+---+---+---+---+---+---+-----+	0.786
55	+---+---+---+---+---+---+---+---+-----+	1.135
56	++---+---+---+---+---+---+---+-----+	1.371
57	-+---+---+---+---+---+---+---+-----+	1.325
58	--+---+---+---+---+---+---+---+-----+	2.211
59	+-+---+---+---+---+---+---+-----+	1.588
60	++---+---+---+---+---+---+-----+	1.244
61	++---+---+---+---+---+---+-----+	1.280
62	++---+---+---+---+---+---+-----+	1.256
63	-+---+---+---+---+---+---+-----+	1.260
64	-+---+---+---+---+---+---+-----+	2.134
65	--+---+---+---+---+---+---+-----+	1.210
66	---+---+---+---+---+---+---+-----+	0.570
67	-+---+---+---+---+---+---+-----+	0.633
68	+---+---+---+---+---+---+-----+	1.464
69	--+---+---+---+---+---+---+-----+	1.452
70	--+---+---+---+---+---+---+-----+	1.349
71	--+---+---+---+---+---+---+-----+	0.543
72	-+---+---+---+---+---+---+-----+	0.815
73	--+---+---+---+---+---+---+-----+	0.961
74	--+---+---+---+---+---+---+-----+	1.172
75	--+---+---+---+---+---+---+-----+	1.708
76	--+---+---+---+---+---+---+-----+	1.989
77	-+---+---+---+---+---+---+-----+	1.643
78	--+---+---+---+---+---+---+-----+	1.729
79	--+---+---+---+---+---+---+-----+	1.632
80	+---+---+---+---+---+---+---+-----+	1.219
81	++---+---+---+---+---+---+---+-----+	0.515
82	-+---+---+---+---+---+---+-----+	0.933
83	+---+---+---+---+---+---+---+-----+	1.640
84	-+---+---+---+---+---+---+-----+	1.706
85	+---+---+---+---+---+---+---+-----+	1.278
86	++---+---+---+---+---+---+---+-----+	1.666
87	-+---+---+---+---+---+---+-----+	1.190
88	+++++++=+++++++=+++++++=+++++++=+++++++=	0.291

Strength-3 orthogonal arrays, like regular resolution IV designs, have main effects orthogonal to all two-factor interaction contrasts. However, rather than having two-factor interactions completely aliased with (or orthogonal to) one another, the two-factor interactions may be partially aliased, as was the case for the strength-2 orthogonal arrays discussed in Section 6.3. Miller and Sitter (2001) recommended the following sensible two-step approach to the analysis of strength-3 orthogonal arrays. Step 1 is to identify the significant main effects. Step 2 is to identify significant interactions, based on a *weak heredity* assumption; that is, two-factor interactions involving one or two active factors are considered possible, while interactions involving two inactive factors are assumed to be negligible. We now illustrate such an analysis for the data in Tables 7.15 and 7.16.

For each of the 41 factors, the levels were chosen to be just outside the range found in commercial processors. For the sake of brevity, we list only the factors and their levels that were subsequently found to be active (see Table 7.17). Yi et al. (2005) tested each of these 88 computer configurations against 48 benchmark computing tasks. Here, we analyze the cycles per instruction (CPI) data for just the first of these benchmarks. One may analyze CPI as the response, or its reciprocal, instructions per cycle. A plot of the data shows two extreme values: one high CPI when every factor is at its low level and one high speed when every factor is at its high level. Since the factor levels are assigned such that the high level should be preferred, we would expect predominantly negative main effect coefficients for the response CPI and positive estimates for its reciprocal.

Table 7.17. Active factors for CPI simulation with first benchmark

Column	Factor	Low Level	High Level
6	Branch predictor	Two-level	Perfect
7	Branch misprediction penalty	10 cycles	2 cycles
12	Re-order buffer entries	8	64
13	Integer ALUs	1	4
14	Integer ALU latency	2 cycles	1 cycle
24	Load-store queue entries	0.25ROB	ROB
29	L1 D-cache latency	4 cycles	1 cycle
33	L2 cache latency	20 cycles	5 cycles

With saturated designs of the form (7.7) (discussed at the beginning of Section 7.3) there are $N/2$ df for main effects and $N/2 - 1$ df for two-factor interactions. If the number of factors $k < N/2$, then $N/2 - k$ df correspond to combinations of three-factor and higher order interactions, which are commonly assumed to be negligible and so treated as error. Using this partitioning for the CPI data, the initial ANOVA is

Source	df	SS	MS	%SS
Main effects	41	14.2570	0.3477	81.85
Two-factor interactions	43	3.0955	0.0720	17.77
Error	3	0.0660	0.0220	0.38
Total (corrected)	87	17.4185		100.00

Since these data come from a processor simulator, there is no random error. Thus, rather than using p -values for Lenth t statistics, we may simply use parsimony as a criterion; that is, what simple model accounts for most of the variation? First, we select main effects to include in the model. Using forward selection regression, the first 15 effects to enter are shown in Table 7.18. Each of these estimates is negative, which corresponds with our expectation for CPI main effects. The column 43 contrast would be the 16th largest main effect, if it were included; b_{43} is nonzero, not because of random error (since there is none), but due to three-factor and/or higher-order interactions among the factors. Such higher-order interactions bias all of the main effects to some extent. No more main effects need be considered. In fact, for a very parsimonious model, one might select just 7 factors, since they account for nearly 73% of the variation; the remaining 24 main effects account for only an additional 9%. Thus, for our first stage of model selection we select factors **6, 12, 13, 14, 24, 29, and 33**; see Table 7.17 for a description of each.

Table 7.18. Stepwise regression for main effects in Example 7.5

Step	Factor	Estimate	SS	R ²
1	x_{12}	-0.230	4.649	26.69
2	x_6	-0.182	2.900	43.34
3	x_{13}	-0.180	2.857	59.74
4	x_{29}	-0.092	0.748	64.04
5	x_{24}	-0.085	0.634	67.68
6	x_{14}	-0.074	0.477	70.42
7	x_{33}	-0.068	0.405	72.74
8	x_7	-0.053	0.250	74.18
9	x_{40}	-0.050	0.216	75.42
10	x_{27}	-0.049	0.213	76.64
11	x_{30}	-0.047	0.193	77.74
12	x_{26}	-0.038	0.124	78.46
13	x_{34}	-0.032	0.088	78.96
14	x_2	-0.030	0.079	79.42
15	x_{31}	-0.028	0.067	79.80

To consider interactions, we fit a model with the 7 largest main effects and consider as eligible the 28 two-factor interactions involving 2 of these factors and the remaining $3 \times 34 = 102$ interactions involving one of the largest 3

factors and any of the nonsignificant ones. From this large number of possible interactions, two stand out as large using stepwise regression: $x_6 * x_7$ and $x_{12} * x_{24}$. The estimated coefficients are in Table 7.19. Note that the estimates for main effects are unchanged by the addition of two-factor interactions, because this is an orthogonal array of strength 3. These two interaction columns are weakly correlated, so their standard errors are slightly larger than for main effects.

Table 7.19. Reduced model for Example 7.5

Term	Estimate	Std Error	t-Ratio
Intercept	1.288	0.0208	61.89
x_6	-0.182	0.0208	-8.72
x_7	-0.053	0.0208	-2.56
x_{12}	-0.230	0.0208	-11.04
x_{13}	-0.180	0.0208	-8.66
x_{14}	-0.074	0.0208	-3.54
x_{24}	-0.085	0.0208	-4.08
x_{29}	-0.092	0.0208	-4.43
x_{33}	-0.068	0.0208	-3.26
$x_6 * x_7$	0.090	0.0209	4.32
$x_{12} * x_{24}$	0.090	0.0209	4.31

This model explains 83% of the variation in CPI. The unexplained variation is due almost exclusively to omitted main effects and two-factor interactions. The ANOVA is:

Source	df	SS	MS
Model	10	14.4835	1.448
(8 main effects			1.615)
(2 two-factor interactions			0.782)
Error	77	2.9349	0.038
(33 omitted main effects			0.040)
(41 omitted 2-factor interactions			0.037)
(3 omitted 3-factor interactions			0.022)
Total (corrected)	87	17.4185	

In this particular application, due to the likely monotonicity of effects, we do not expect an interaction to be larger than either of the corresponding main effects. For the $x_{12} * x_{24}$ interaction we have

$$\widehat{CPI} = \dots - 0.23x_{12} - 0.085x_{24} + 0.090x_{12} * x_{24} + \dots$$

The consequence of a positive interaction is that for $x_{12} = 1$, the effect for factor x_{24} disappears; that is, for this benchmark, if the reorder buffer is large,

a load/store queue only one-fourth its size is adequate. The $x_6 * x_7$ interaction was expected; if the branch predictor never makes a mistake ($x_6 = 1$), then it does not matter how severe the penalty for mistakes (x_7) is. Other smaller interactions are suggested using stepwise regression. However, the reduced model in Table 7.19 appears to contain all of the most prominent effects. It is possible to estimate more interactions. However, the partial aliasing begins to cause confusion about which effects to include if many more interactions are considered.

Here we have analyzed CPI data from only the first of 48 benchmark computing tasks. Yi et al. (2005) ranked the factors by amount of variation explained for each task. From these they computed an average rank to provide an overall assessment of the importance of each factor on processor efficiency. They also performed a cluster analysis of the rankings to find a small set of the benchmark tasks that are distinct from one another. For those interested in the details of experimenting with computer architecture, see also Joshi et al. (2006).

7.2 Regular Resolution IV Designs

7.2.1 Criteria for ranking regular resolution IV designs

As with resolution III designs, minimum aberration is one popular means for comparing resolution IV designs. However, the case of 2^{9-4} designs illustrates how the word length pattern fails to capture important differences among resolution IV designs. There are five distinct regular resolution IV 2^{9-4} designs, with $A_4 = 6, 7, 9, 10$, and 14, respectively (see Chen, Sun and Wu 1993, p. 138). The minimum aberration design listed in Appendix G uses columns 7 (**F=ABC**), 11 (**G=ABD**), 19 (**H=ABE**) and 29 (**J=ACDE**) as generators, which produces six length-4 words in the defining relation

$$\mathbf{I} = \mathbf{ABCF} = \mathbf{ABDG} = \mathbf{ABEH} = \mathbf{CDFG} = \mathbf{CEFH} = \mathbf{DEGH} = \dots \quad (7.3)$$

and the following aliasing among 28 of the 36 two-factor interactions:

$$\begin{array}{ll} \mathbf{AB} = \mathbf{CF} = \mathbf{DG} = \mathbf{EH} \\ \mathbf{AC} = \mathbf{BF} & \mathbf{AD} = \mathbf{BG} \\ \mathbf{AE} = \mathbf{BH} & \mathbf{AF} = \mathbf{BC} \\ \mathbf{AG} = \mathbf{BD} & \mathbf{AH} = \mathbf{BE} \\ \mathbf{CD} = \mathbf{FG} & \mathbf{CE} = \mathbf{FH} \\ \mathbf{CG} = \mathbf{DF} & \mathbf{CH} = \mathbf{EF} \\ \mathbf{DE} = \mathbf{GH} & \mathbf{DH} = \mathbf{EG} \end{array}$$

Since factor **J** does not appear in the length-4 words in (7.3), its two-factor interactions are aliased only with higher-order interactions.

Consider now the second-best design in terms of aberration, design 9-4.2 with $A_4 = 7$. By using column 14 (**BCD**) rather than column 19 as the generator for **H**, the aliasing becomes

$$\begin{array}{ll}
\mathbf{AB} = \mathbf{CF} = \mathbf{DG} & \mathbf{AC} = \mathbf{BF} = \mathbf{GH} \\
\mathbf{AD} = \mathbf{BG} = \mathbf{FH} & \mathbf{AF} = \mathbf{BC} = \mathbf{DH} \\
\mathbf{AG} = \mathbf{BD} = \mathbf{CH} & \mathbf{AH} = \mathbf{CG} = \mathbf{DF} \\
\mathbf{BH} = \mathbf{CD} = \mathbf{FG}
\end{array}$$

Neither **J** nor **E** appears in the 7 length-4 words, so the 15 two-factor interactions involving one or both of these factors are clear. The minimum aberration design is inferior to this second design in three ways:

- The minimum aberration design has 8 clear two-factor interactions, whereas design 9-4.2 has 15 clear two-factor interactions.
- The minimum aberration design has $\text{rank}(\mathbf{X}_2' \mathbf{X}_2) = 21$, whereas design 9-4.2 has $\text{rank}(\mathbf{X}_2' \mathbf{X}_2) = 22$, which provides one more degree of freedom for two-factor interactions.
- The minimum aberration design aliases a set of four two-factor interactions together. Design 9-4.2 never aliases more than three two-factor interactions together.

For resolution IV designs, we need criteria that highlight the differences in aliasing among two-factor interactions. Block and Mee (2003) introduced the alias length pattern (alp) as the frequencies of the lengths of alias sets for two-factor interactions:

$$\text{alp} = (a_1, a_2, \dots, a_L), \quad (7.4)$$

where L denotes the size of the largest alias set and a_j denotes the number of alias sets of size j ($j = 1, 2, \dots, L$). For the minimum aberration 2^{9-4} design, $\text{alp} = (8, 12, 0, 1)$, whereas for design 9-4.2, $\text{alp} = (15, 0, 7)$. The following criteria for ranking resolution IV designs are functions of alp:

- Maximize a_1 , the number of clear two-factor interactions.
- Maximize $M = a_1 + a_2 + \dots + a_L = \text{rank}(\mathbf{X}_2)$, the largest number of two-factor interactions that can be estimated for any model.
- Minimize L , the length of the longest chain of aliased two-factor interactions.
- Maximize Cheng, Steinberg, and Sun's (1999) estimation capacity sequence (E_1, E_2, \dots, E_M) , where E_j is the proportion of models containing the k main effects and j two-factor interactions that can be estimated from the design. These proportions may be determined from the alp; for instance,

$$E_2 = 1 - 4 \sum_{i=1}^M i(i-1)a_i/[k(k-1)(k^2-k-2)].$$

- Minimize the number of length-4 words, $A_4 = \sum_{j=2}^L j(j-1)a_j/6$.

All five criteria are useful for characterizing regular resolution IV designs.

As the case of 9 factors in 32 runs illustrates, a single design may not rank first on all five criteria. Which design is preferred for a given application depends on the circumstances. If the ability to estimate all two-factor interactions for 2 factors is considered critical, design 9-4.2 is ideal. If, instead, one can rule out a couple of two-factor interactions (involving 4 factors) a priori, such as **AB** and **CF**, then the minimum aberration design becomes very attractive, since the only remaining aliasing involves 13 pairs of two-factor interactions. Even if no two-factor interactions can be assumed away, the minimum aberration design can estimate more models with a small number of two-factor interactions. Because different resolution IV designs will have different aliasing structure, careful choice of a design, combined with thoughtful assignment of factors to design columns, can result in substantially more informative experiments.

7.2.2 Even resolution IV 2^{k-f} fractional factorial designs

Resolution IV designs with $k = N/2$ are optimal in at least two respects. First, $N/2$ is the maximum number of factors permitting a matrix of the form (7.2). Second, such designs have $\text{rank}[\mathbf{X}_1, \mathbf{X}_2] = N$; that is, all of the degrees of freedom are useful for estimating main effects and two-factor interactions. Designs with $\text{rank}[\mathbf{X}_1, \mathbf{X}_2] = N$ are *second-order saturated* (SOS)—terminology that mirrors the label “saturated main effect design” used in Chapter 6 when $\text{rank}[\mathbf{X}_1] = N$.

Designs for which $k = N/2$ are constructed by foldover. The regular resolution IV designs 2^{4-1} , 2^{8-4} , 2^{16-11} , 2^{32-26} , etc. each have the form

$$\mathbf{D} = \begin{bmatrix} \mathbf{S}_{N/2} \\ -\mathbf{S}_{N/2} \end{bmatrix}, \quad (7.5)$$

where $\mathbf{S}_{N/2}$ is the Sylvester-type Hadamard matrix defined in (6.1). (Non-regular designs of strength 3, with $k = N/2$, are obtained in similar fashion folding over other Hadamard matrices; see Section 7.3.)

The defining relations for fractions constructed by (7.5) have only even-length words. Thus, an alternative construction of an N -run resolution IV with $k = N/2$ factors is to construct a full factorial in $k - f$ basic factors and then to use every interaction with an odd number of factors as a generator. We illustrate this second construction for the $k = 8$, $N = 16$ fraction. We need four basic factors to create the initial 2^4 factorial. To these we append all of the interactions involving an odd number of factors:

A	B	C	D	ABC	ABD	ACD	BCD
-1	-1	-1	-1	-1	-1	-1	-1
1	-1	-1	-1	1	1	1	-1
-1	1	-1	-1	1	1	-1	1
1	1	-1	-1	-1	-1	1	1
-1	-1	1	-1	1	-1	1	1
1	-1	1	-1	-1	1	-1	1
-1	1	1	-1	-1	1	1	-1
1	1	1	-1	1	-1	-1	-1
-1	-1	-1	1	-1	1	1	1
1	-1	-1	1	1	-1	-1	1
-1	1	-1	1	1	-1	1	-1
1	1	-1	1	-1	1	-1	-1
-1	-1	1	1	1	1	-1	-1
1	-1	1	1	-1	-1	1	-1
-1	1	1	1	-1	-1	-1	1
1	1	1	1	1	1	1	1

If we assign the factors **E**, **F**, **G**, and **H**, respectively, to these interactions, the defining relation for this fraction is

$$\begin{aligned}
 I &= ABCE = ABDF = ACDG = BCDH \\
 &= CDEF = BDEG = ADEH = BCFG = ACFH = ABGH \\
 &= AEFG = BEFH = CEGH = DFGH = ABCDEFGH.
 \end{aligned}$$

This example illustrates how if only odd-length generators are used, then the defining relation contains only even-length words. This design's word length pattern is (14, 0, 0, 0, 1).

For even fractional factorial designs such as the 2^{8-4} the following hold:

- Main effects will be aliased only with interactions of odd length; for example, $A = BCE = BDF = CDG = ABCDH = ACDEF = \dots = BCDEFGH$
- Even-length interactions will be aliased together; for example, $AB = CE = DF = BCDG = ACDH = ABCDEF = \dots = GH = CDEFGH$
- There will be $N/2$ df for odd-length aliases, and $N/2 - 1$ df for two-factor interactions (and other interactions involving an even number of factors).

General results for even resolution IV designs

All regular even 2^{k-f} designs are a projection of the regular resolution IV design with $k = N/2$. This is apparent, since every regular even design uses a subset of the odd-length interactions as generators, and no even-length interactions. In addition, the following hold:

- For $k/N > 5/16$, all regular resolution IV 2^{k-f} designs are even. Thus, all resolution IV designs for 8 and 16 runs are even. For $N = 32$, all resolution IV designs with 11–16 factors are even.
- For $k/N > 5/16$, no resolution IV design has clear two-factor interactions (Chen and Hedayat 1998).
- Every regular even design with $k/N \geq 5/16$ has k df for main effects, $N/2 - 1$ df for two-factor interactions, and $N/2 - k$ df for three-factor interactions not aliased with main effects.

An even resolution IV design is appropriate if the primary attention is for estimating main effects, and the risk of two-factor interactions being present precludes the use of a resolution III design. If two-factor interactions are identified, as they were for Example 7.3, the aliasing will make it difficult to ascertain which particular interactions are active, and so more data will be required to clarify which interactions are important.

Enumeration of even resolution IV designs

Chen, Sun, and Wu (1993) enumerated all regular nonisomorphic designs of resolution IV for $N = 32$ and 64. Block and Mee (2005) documented a search for the best even designs at $N = 128$, which was subsequently confirmed by Xu (2009). The minimum aberration designs for $N = 32$, 64, and 128 and $k/N > 5/16$ are summarized in Appendix G. The generators are arranged so that the designs for different k but identical N are embedded in one another in a convenient manner. When $k/N > 5/16$, all resolution IV designs have $a_1 = 0$ and $M = N/2 - 1$; hence, the minimum aberration criterion is adequate for ranking even designs.

For larger N , Butler's (2003a) complementary design theory for even resolution IV fractions is very useful. To obtain the minimum aberration design with k factors ($5N/16 < k < N/2$), delete $d = N/2 - k$ columns from the resolution IV design with $N/2$ factors, such that the d deleted columns form a minimum aberration even design. For instance, for $N = 256$ one may use any k columns from the resolution IV $2^{128-120}$ design for $k = 125-127$. For the minimum aberration designs with $k = 120-124$, one must delete $d = 128 - k$ columns from the $2^{128-120}$ design such that the deleted columns form a full 2^d . For $k=119$, delete nine columns forming a resolution VIII design; for $k = 116-118$ (81-115), one must delete a minimum aberration fraction with resolution VI (IV).

7.2.3 Even/odd resolution IV 2^{k-f} fractional factorial designs

For $k/N \leq 5/16$, regular 2^{k-f} fractional factorial designs of resolution IV exist where half of the words in the defining relation are of even length and half are of odd length. Consider two $N = 32$ examples:

1. Minimum aberration 2^{10-5} , with word length pattern $(10, 16, 0, 0, 5)$ and $\text{alp} = (0, 20, 0, 0, 1)$; this is a SOS design, having $M = 21$ df for two-factor interactions. All other 2^{10-5} resolution IV designs are even designs having 15 or more length-4 words and only $M = 15$ df for two-factor interactions.
2. Resolution IV 2^{9-4} fraction with the most clear two-factor interactions. Recall design 9-4.2 discussed in Section 7.2.1. This fraction has word length pattern $(7, 7, 0, 0, 0, 1)$ and $\text{alp} (15, 0, 7)$. It is also SOS, since $M = 22 = N - k - 1$.

The SOS designs maximize the number of degrees of freedom for two-factor interactions. Furthermore, all even/odd resolution IV designs are the projection of one or more even/odd SOS design. The two designs just mentioned are the only even/odd resolution IV SOS designs for $N = 32$. At $N = 64$, there are eight even/odd SOS designs; they occur at $k = 20, 18, 17$ (five different designs), and 13.

The SOS designs may be used to construct larger SOS by doubling as follows. Let \mathbf{D} denote a resolution IV SOS design with N runs and k factors. Then

$$\begin{bmatrix} \mathbf{D} & \mathbf{D} \\ \mathbf{D} & -\mathbf{D} \end{bmatrix} \quad (7.6)$$

is a resolution IV SOS design with $2N$ runs and $2k$ factors. The 2^{20-14} and 2^{18-12} SOS designs may be obtained by doubling the 2^{10-5} and 2^{9-4} SOS designs, respectively. The 8 even/odd SOS designs at $N = 64$ may be doubled to produce 128-run SOS designs for 40, 36, 34, and 26 factors. There exist 79 other even/odd SOS designs for $N = 128$, for k ranging from 21 to 33 (Block 2003). Chen and Cheng (2006) showed that for $k > 1 + N/4$, all regular resolution IV SOS designs are obtained by doubling.

General results for even/odd resolution IV designs

- All minimum aberration resolution IV designs with $k \leq 5N/16$ are even/odd.
- The SOS designs with $k = 5N/16$ are obtained by repeated doubling [as shown in (7.6)] the resolution V 2^{5-1} fraction. These designs are all minimum aberration, as are many of their projections. For $N = 64$, all minimum aberration designs for $k = 14-19$ are projections of the SOS 2^{20-14} fraction; for $N = 128$, all minimum aberration designs for $k = 30-39$ are projections of the SOS 2^{40-33} fraction. Xu and Cheng (2008) used a complementary design approach to identify the best columns to omit to obtain these minimum aberration projections and made conjectures regarding the number of these that are minimum aberration designs. One implication of their result is that the other SOS designs with $k \geq 1 + N/4$ have more length-4 words than the best projections of the $k = 5N/16$ design series.

- If $k \leq N/4 + 1$, there exists a regular even/odd resolution IV fraction with clear two-factor interactions. However, this does not imply that the minimum aberration design for $k \leq N/4 + 1$ has clear two factor interactions.

Constructing good even/odd resolution IV designs for $N \geq 256$

Complete enumeration of even/odd resolution IV designs for $N = 256$ has been achieved only up to $k = 17$ factors (Xu 2009, Table 10), but all minimum aberration and weak minimum aberration 256-run designs have been identified for up to 28 factors (Xu 2009, Table 12). In addition, for 29–80 factors, Xu (2009, Table 13) lists good resolution IV 256-run designs. Unlike the case for $(5/16)N < k < (1/2)N$, where all resolution IV designs are projections of a single SOS design with $k = N/2$, there are many even/odd SOS designs, and so the search for best designs is more difficult. For instance, at $N = 256$, an incomplete search by Block (2003) turned up more than 34,000 even/odd SOS designs.

Xu's best 256-run designs for 59–79 factors are all projections of the SOS design at $k = 80$, which is a $k = 5N/16$ series design (see Butler 2005, Sect. 3). Of these, the designs for $k = 69$ –80 are guaranteed to have minimum aberration (Xu and Cheng 2008). Just as the 2^{20-14} design has $a_{10} = 3$ and the 2^{40-33} design has $a_{20} = 7$, the 2^{80-72} design has $a_{40} = 15$; that is, every factor appears in 15 alias chains. By sequentially deleting factors that shorten these 15 chains the most, one generally minimizes the number of length-4 words. Xu and Cheng (2008) explained precisely how to obtain the minimum aberration projections.

For good even/odd resolution IV designs of size 512 and $k = 24, \dots, 160$, see Xu (2009, Tables 14 and 15). We now turn attention to criteria other than minimum aberration.

7.2.4 Resolution IV designs that maximize the number of clear two-factor interactions

Chen and Hedayat (1998) showed that resolution IV designs with $k > 1 + N/4$ have no clear two-factor interactions. (Chen and Hedayat also showed that some resolution III designs with k as large as $N/2$ may have clear two-factor interactions; such designs might be of interest if one was assured that certain two-factor interactions were zero and that no two-factor interaction but these would be aliased with main effects.) Chen, Sun, and Wu (1993) listed the number of clear two-factor interactions for each design in their tables, and Block (2003) provided the complete alp for many 128-run resolution IV designs. Several other papers have appeared that provide a means of constructing designs with a large number of clear two-factor interactions (see Tang, Ma, Ingram, and Wang 2002, Yang and Liu 2006) or that explore the relation between the minimum aberration and maximum clear criteria (Wu and Wu 2002).

Table 7.20 shows the number of clear two-factor interactions for the minimum aberration design, followed by the maximum number of clear two-factor interactions among all resolution IV designs, for $N = 32$, 64, and 128. For designs where the number of factors is only slightly more than is possible for a resolution V design, the minimum aberration designs also maximize the number of clear two-factor interactions; this is the case for two 32-run designs, four 64-run designs, and three 128-run designs. However, as k increases, the number of clear two-factor interactions for minimum aberration designs diminishes rapidly. This is because minimum aberration designs tend to have more uniform size alias sets (Cheng, Steinberg, and Sun 1999). No minimum aberration design of size 64 (128) has any clear two-factor interactions for $k > 14$ (23). As k approaches $1 + N/4$, the maximum number of clear two-factor interactions equals $2k - 3$, which corresponds to all interactions involving two factors. Design 9-4.2 discussed earlier is one such example.

Table 7.20. Number of clear two-factor interactions (2fi's) for minimum aberration design versus design with maximum number clear

k	No. 2fi's	$N = 32$	$N = 64$	$N = 128$
6	15	All		
7	21	15:15		
8	28	13:13	All	
9	36	8:15	30:30	
10	45	None	33:33	
11	55		34:34	All
12	66		36:36	60:60
13	78		20:36	66:66
14	91		8:25	73:73
15	105		0:27	63:77
16	120		0:29	60:69
17	136		0:31	46:75
18	153		None	33:81
19	171			36:78
20	190			24:84
21	210			26:84
22	231			25:48
23	253			12:45
24	276			0:45
:	:			:
33	528			0:63
34	561			None

In practice, one may have a particular set of interactions that are deemed most likely. There are two approaches to constructing resolution IV designs that accommodate a specific subset of two-factor interactions.

- The more stringent requirement is to have the main effects and specific subset of two-factor interactions be clear of aliasing with other two-factor interactions. This ensures that the specified effects may be estimated without bias, provided there are no three-factor or higher-order interactions. Ke, Tang, and Wu (2005) provide 32- and 64-run designs with clear two-factor interactions for several standard subsets of interactions. Wu, Mee, and Tang (2008) enumerated all admissible graphs corresponding to resolution IV designs with clear two-factor interactions for $N = 32, 64$, and 128. An example is provided below.
- A less stringent condition is to require only that the specific subset of two-factor interactions be “eligible” (i.e., estimable, ignoring all other two-factor interactions). Wu and Chen (1991, 1992) provided graphs to aid the assignment of factors to columns, most of which are for 16-run designs.

Consider now an 11-factor example from Wu and Chen (1992). The first six factors **A–F** belong to the epoxy dispensing step for a circuit pack assembly process, and the remaining five factors **M–Q** pertain to the subsequent component placement step. Two-factor interactions among the six epoxy factors are of interest and must be estimable. Wu and Chen’s Figure 1 shows a graph from Taguchi (1987, p. 1134) corresponding to a 32-run design, for which the 11 main effects and 15 epoxy factor interactions of interest are estimable. However, this design has resolution III. A much better solution is provided by the minimum aberration (resolution IV) 2^{11-5} design by assigning factors **A–F** to columns that form a six-letter word in the defining relation. Appendix G lists columns 7, 11, 13, 21, 25, and 31 as generators for the minimum aberration 2^{11-6} design. If we assign **A–E** to the basic columns 1, 2, 4, 8, 16, use column 31 to define **F**, and the remaining generators for factors **M–Q**, then **ABCDEF** forms a word in the defining relation and the aliasing among two-factor interactions is

$$\begin{aligned}
 \mathbf{AB} &= \mathbf{CM} = \mathbf{DN} \\
 \mathbf{AC} &= \mathbf{BM} = \mathbf{DO} = \mathbf{EP} \\
 \mathbf{AD} &= \mathbf{BN} = \mathbf{CO} = \mathbf{EQ} \\
 \mathbf{AE} &= \mathbf{CP} = \mathbf{DQ} \\
 \mathbf{AF} &= \mathbf{MQ} = \mathbf{NP} \\
 \mathbf{BC} &= \mathbf{AM} = \mathbf{FQ} = \mathbf{NO} \\
 \mathbf{BD} &= \mathbf{AN} = \mathbf{FP} = \mathbf{MO} \\
 \mathbf{BE} &= \mathbf{FO} = \mathbf{MP} = \mathbf{NQ} \\
 \mathbf{BF} &= \mathbf{CQ} = \mathbf{DP} = \mathbf{EO} \\
 \mathbf{CD} &= \mathbf{AO} = \mathbf{MN} = \mathbf{PQ}
 \end{aligned}$$

$$\begin{aligned}
 \mathbf{CE} &= \mathbf{AP} = \mathbf{FN} = \mathbf{OQ} \\
 \mathbf{CF} &= \mathbf{BQ} = \mathbf{EN} \\
 \mathbf{DE} &= \mathbf{AQ} = \mathbf{FM} = \mathbf{OP} \\
 \mathbf{DF} &= \mathbf{BP} = \mathbf{EM} \\
 \mathbf{EF} &= \mathbf{BO} = \mathbf{CN} = \mathbf{DM}
 \end{aligned}$$

The first interaction in each alias set in the interaction of interest. Although this 32-run design permits estimation of the main effects and 15 two-factor interactions of interest, each interaction is aliased with 2 to 3 other two-factor interactions.

What about estimation with the 15 interactions of interest being clear? Ke, Tang, and Wu (2005) is useful for finding such designs. Ke et al. listed four classes of resolution IV designs with clear two-factor interactions. In their notation, the groups G_1 and G_2 denote a partitioning of the k factors. The four classes are as follows:

1. $G_1 \times G_1$ (i.e., all two-factor interactions involving two factors from G_1 are clear)
2. $G_1 \times G_1$ and $G_2 \times G_2$
3. $G_1 \times G_1$ and $G_1 \times G_2$
4. $G_1 \times G_2$

Our example is of class 1, where G_1 consists of the six factors **A–F**. The closest to a solution in Ke et al. (2005) are (i) a 2^{9-3} design with six factors in G_1 ; (ii) a 2^{10-4} design with five factors in G_1 ; (iii) a 2^{11-5} design with four factors in G_1 . There are 64-run designs with as many as 34 clear two-factor interactions, but the sets of clear two-factor interactions never contain all $\binom{6}{2}$ interactions for six factors. The closest that can be achieved for the problem of interest is to use Chen, Sun, and Wu's (CSW's) (1993) design 12-5.3. Assigning factors **M–Q** to columns (1, 2, 4, 8, 49) and factors **A–F** to columns (16, 32, 7, 11, 29, 46), the aliasing of two factor interactions is

$$\begin{aligned}
 \mathbf{AB} &= \mathbf{MQ} \\
 \mathbf{CD} &= \mathbf{OP} \\
 \mathbf{EF} &= \mathbf{NQ} \\
 \mathbf{MB} &= \mathbf{AQ} \\
 \mathbf{MA} &= \mathbf{BQ} \\
 \mathbf{MC} &= \mathbf{NO} \\
 \mathbf{MD} &= \mathbf{NP} \\
 \mathbf{MO} &= \mathbf{NC} \\
 \mathbf{MP} &= \mathbf{ND} \\
 \mathbf{OD} &= \mathbf{PC} \\
 \mathbf{NF} &= \mathbf{EQ} \\
 \mathbf{NE} &= \mathbf{FQ} \\
 \mathbf{MN} &= \mathbf{OC} = \mathbf{PD}
 \end{aligned}$$

Twelve of the 15 interactions of interest are clear; the other three (**AB**, **CD**, and **EF**) are each aliased with just one two-factor interaction.

Class 3 and class 4 designs are most suitable for robust parameter design applications where G_1 represents the control factors and G_2 represents noise factors. The objective of robust parameter design experiments is to identify control factor levels that make a process or product robust (i.e., insensitive) to variation in the noise factors. Section 10.3.2 contains analysis for a 32-run robust parameter design example with three noise factors and seven control factors. For a readable introduction to robust parameter design, see Abraham and MacKay (1993).

7.3 Strength-3 Orthogonal Arrays

7.3.1 Strength-3 orthogonal arrays that are even designs

Let k be a multiple of 4 and let H_k be any Hadamard matrix of order k . Then the $N = 2k$ run design obtained by foldover

$$\mathbf{D} = \begin{bmatrix} \mathbf{H}_k \\ -\mathbf{H}_k \end{bmatrix} \quad (7.7)$$

has strength 3; that is, the design projects into an equally replicated 2^3 factorial in every subset of three columns. The most commonly used strength-3 array that is not also a regular resolution IV design is the OA(24, 2^{12} , 3). Miller and Sitter (2001) recommended using this design, not only to estimate main effects clear of aliasing from two-factor interactions but also to attempt to identify a small number of important two-factor interactions.

Miller and Sitter (2001) presented a nine-factor, 24-run example where the effects of two factors and their interaction are so strong that the correct model would be detectable from a 12-run, strength-2 orthogonal array (as we did in Example 6.6). In Section 7.1.4 we analyzed data from a folded-over Hadamard design with 41 factors and 88 runs, where identifying 2 or 3 interactions is straightforward but identifying more becomes challenging.

In Section 6.3, our primary concern in constructing OA(N , 2^k , 2) was the magnitude of the correlations between main effects and two-factor interactions. For instance, the 20-run design with generalized resolution 3.4 was troubling, due to the presence of three-factor interaction columns that summed to ± 12 . After foldover, it is the magnitude of correlations among pairs of two-factor interaction columns that determines the generalized resolution. Table 7.21 lists the generalized resolution and the initial portion of the confounding frequency vector for the most commonly used OA(N , $2^{N/2}$, 3). B_4 is the generalized number of length-4 words; for instance, $B_4 = (1/3)^2 495 = 55$ for the OA(24, 2^{12} , 3).

Table 7.21. Generalized resolution for common OA($N, 2^{N/2}, 3$)

N	k	Res.	Gen. H_k	First portion of cfv	B_4
24	12	4.67	H_{12}	$A_4(0.33) = 495$	55
32	16	4.00	Hall IV	$A_4(1, 0.5) = (28, 448)$	140
40	20	4.40	Any H_{20}	$A_4(0.6, 0.2) = (285, 4560)$	285
48	24	4.67	Paley	$A_4(0.33) = 4554$	506
56	28	4.29	Williamson	$A_4(0.71, 0.43, 0.14) = (7, 2436, 18032)$	819
64	32	4.75	Paley	$A_4(0.25) = 19840$	1240
72	36	4.67	Cyclic PB	$A_4(0.33, 0.11) = (10710, 48195)$	1785
80	40	4.00	Doubled H_{20}	$A_4(1, 0.6, 0.2) = (190, 2280, 36480)$	2470
88	44	4.73	Paley	$A_4(0.27, 0.09) = (33110, 102641)$	3311
96	48	4.67	Paley	$A_4(0.33, 0.17) = (12972, 103776)$	4324

The choice of H_k can dramatically affect the generalized resolution. For instance, the maximum correlation for Example 7.5 (Tables 7.15 and 7.16) is $24/88 = 3/11$. However, folding over Williamson's H_{44} , instead of Paley's H_{44} , produces a maximum correlation of $9/11$! Note that the maximum correlation is 1 in two cases for Table 7.21; for $N=32$, $A_4(1) > 0$ is inevitable, since every H_{16} has $A_4(1) > 0$. The other case arises for $N = 80$; this is a foldover of a doubled design, which here produces complete aliasing among $3k(k - 2)/8$ pairs of two-factor interactions. The smallest correlations arise from folding over Paley's H_{32} and H_{44} matrices. No literature has yet appeared regarding the best projections from foldovers of Hadamard designs.

Cheng (1998) and Bulutoglu and Cheng (2003) guaranteed that the two-factor interaction model is estimable in every subset of five factors for any strength-3 orthogonal array where the run size is not a multiple of 16 or for arrays obtained as a foldover of Paley Hadamard designs. Only for the cases $N = 32$ and 80 in Table 7.21, where $A_4(1) > 0$, does this not apply.

7.3.2 Strength-3 orthogonal arrays that are not even designs

Any design consisting of mirror-image pairs of runs necessarily has at most $N/2 - 1$ df for estimating two-factor interactions [i.e., construction by (7.7) implies $\text{rank}(\mathbf{X}_2) \leq N/2 - 1$]. To consider designs with $N/2$ or more df for two-factor interaction in a strength-3 array, one must use a different construction. Cheng, Mee, and Yee (2008) presented two means for constructing strength-3 orthogonal arrays for which $\text{rank}(\mathbf{X}_2) > N/2$. One construction is to fold over an OA($N/2, 2^{N/4}, 3$) a second time, adding one more column and reversing just a subset of the columns for the foldover. This produces strength-3 orthogonal arrays with $k = 1 + N/4$. These OA($N, 2^k, 3$) have clear two-factor interactions and are often SOS. For instance, one may construct an OA(48, $2^{13}, 3$) with $\text{rank}(\mathbf{X}_2) = 34$, 12 clear two-factor interactions, and $A_4(1/3) = 234$. Tang (2006) discussed the case of reversing just one column. Reversing

just one column in the foldover of the OA($N/2, 2^{N/4}, 3$) maximizes the number of clear two-factor interactions, but it also has more aberration. For instance, Tang's Example 1 has 23 clear two-factor interactions but with $A_4(1/3) = 330$. Both of these OA(48, 2^{13} , 3) are SOS, with $\text{rank}(\mathbf{X}_2) = 34$; they differ in that one minimizes the G-aberration and the other maximizes the number of clear interactions. Tang (2006) also proved that, as with regular resolution IV designs, $k = N/4 + 1$ is the maximum number of factors for a strength-3 orthogonal array to have any clear two-factor interactions.

A second construction of OA($N, 2^k, 3$) without mirror-image pairs of runs is to take the Kronecker product of a Hadamard matrix H_m and the resolution V 2^{5-1} . This produces an OA($16m, 2^{5m}, 3$), which is SOS; that is, $\text{rank}(\mathbf{X}_2) = 11m - 1$. For $m = 4$ and 8, this construction produces the regular resolution IV designs with $k = 5N/16$ for $N = 64$ and 128, respectively. For $m = 12$, it produces an OA(192, 2^{60} , 3). This fraction has 131 df for two-factor interactions, much more than $N/2$; also, the correlations among two-factor interaction columns are mostly small, with $A_4(0.25, 0.125, 0.06, 0.03) = (1056, 3216, 53955, 913464)$. Analogous to the regular $5N/16$ series designs, this design is believed to have many projections with minimum G_2 aberration.

Finally, Xu (2005) constructed several minimum G_2 -aberration, strength-3 orthogonal arrays for 7–9 factors in 32 runs and for 9–14 factors in 64 runs that are not foldover designs. Xu's 64-run designs for 13 and 14 factors have much lower aberration than is possible from any regular 2^{13-7} or 2^{14-8} fraction, or as projections of any OA(64, 2^{17} , 3) constructed in Cheng, Mee and Yee (2008). Xu's OA(64, 2^{14} , 3) is obtained from the OA(128, 2^{15} , 4) design constructed in Section 8.2.1 by choosing any 1 of the 15 columns of the 128-run array and discarding the runs for which this column is -1 . The resulting 64-run design has resolution 4.5 and permits estimation of all two-factor interactions in many projections involving up to 10 factors. Taking half of the OA(64, 2^{14} , 3), Xu (2005) obtained an OA(32, 2^{13} , 2) that, although it only has resolution 3.5, does permit estimation of the two-factor interaction model in any set of five factors.

7.4 Nonorthogonal Resolution IV Designs

One can obtain main effect estimates without bias from two-factor interactions by making all main effect contrasts orthogonal to two-factor interaction contrasts. This is guaranteed by any foldover design, even if the main effect contrasts are not orthogonal. Let \mathbf{X}_1 be the first-order model matrix for any saturated main effect, two-level design (see Section 6.4). Then

$$\mathbf{D} = \begin{bmatrix} \mathbf{X}_1 \\ -\mathbf{X}_1 \end{bmatrix} \quad (7.8)$$

is a design for $k = N/2$ factors permitting estimation of the first-order coefficients without bias from two-factor interactions, whether they are included or excluded from the model. Webb (1968) showed that $N = 2k$ is the minimum run size for a design with this property, and Margolin (1969) proved that any equal occurrence design \mathbf{D} with $N = 2k$ and $\mathbf{D}'\mathbf{X}_2 = 0$ is necessarily a “mirror image pair design” given by (7.8). Margolin (1969) presented the efficiencies of 16 such nonorthogonal designs, based on saturated main effect designs available at the time.

Miller and Sitter (2005) promoted the use of these designs “when the primary goal is identification of important main effects with a secondary goal of entertaining a small number of potentially important (two-factor) interactions.” Non-orthogonal foldover designs are available for any even N , whereas orthogonal resolution IV designs are only available for N a multiple of 8. Table 6.35 lists the D-efficiency and A-efficiency for optimal saturated main effects designs for $N \leq 19$. These same efficiencies apply to the main effect estimates for the foldover design (7.8). Note that the resolution IV designs obtained by foldover become equal-occurrence designs, even when the saturated main effect design used is not.

We consider examples for $k = 5$ and 6. A D-optimal design for four factors in five runs is easily found (e.g., using JMP’s Custom Design option). One such design (from the many isomorphic ones), with the intercept column appended, is

$$\mathbf{X}_1 = \begin{bmatrix} 1 & -1 & -1 & -1 & 1 \\ 1 & -1 & 1 & 1 & 1 \\ 1 & 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & -1 \end{bmatrix}. \quad (7.9)$$

This design has A-efficiency = 0.9, since the variance for each coefficient is $\sigma^2/4.5$. Each two-factor interaction column has a correlation of $\pm 2/3$ with two main effects and $\pm 1/3$ with the other two effects.

If we fold over the first-order model matrix (7.9) by appending the five runs

$$-\mathbf{X}_1 = \begin{bmatrix} -1 & 1 & 1 & 1 & -1 \\ -1 & 1 & -1 & -1 & -1 \\ -1 & -1 & 1 & -1 & -1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & 1 & 1 & -1 & 1 \end{bmatrix}, \quad (7.10)$$

we create a $N = 10$ -run, five-factor \mathbf{D} with the following properties:

- The variances for the five main effect coefficients now equal $\sigma^2/9$. The D-efficiency and A-efficiency for the first-order model listed in Table 6.35 become the corresponding efficiency for the k main effects estimated using the foldover design.

- The correlations between the columns of \mathbf{D} are $\pm 2/10$. However, the correlations between the estimated coefficients obtained from $(\mathbf{D}'\mathbf{D})^{-1}$ are only $\pm 1/8$, causing little loss of efficiency.
- The 10 two-factor interaction columns are orthogonal to the 5 main effects, and so cannot bias the estimates for main effects.
- There are 4 df for estimating a small number of the 10 two-factor interactions. However, one-third of the correlations among the interaction columns are $\pm 2/3$. Although there is no complete aliasing among these interactions, the many large correlations imply that if more than one interaction is present, it will become difficult to correctly identify them.

As a second example of a nonorthogonal resolution IV design, we consider the 12-run, six-factor design

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & -1 & 1 & 1 \\ -1 & 1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 1 & 1 & -1 & 1 \\ -1 & 1 & 1 & 1 & 1 & -1 \\ -1 & -1 & -1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & 1 & 1 \end{bmatrix}. \quad (7.11)$$

The first six rows are the D-optimal design following Table 6.35 with an intercept column appended; the last six rows are a foldover of the first six. This design has up to 5 df for two-factor interactions. Correlations among the two-factor interaction column range from 0 to 0.707.

Miller and Sitter (2002) compared the performance of these nonorthogonal resolution IV designs with orthogonal resolution III designs. For instance, for $N = 12$, an alternative to the nonorthogonal resolution IV design is to choose six columns from the 12-run Plackett–Burman design or to use a “model robust design” from Li and Nachtsheim (2000). Miller and Sitter’s Table 4 shows the matrix

$$\begin{bmatrix} \mathbf{D}'\mathbf{D} & \mathbf{D}'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{D} & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}$$

for a design isomorphic to (7.11), and their Table 6 shows the same matrix for an OA(12, 2^6 , 2). The clear benefit of the resolution IV design is the ability to estimate main effects without confusion from two-factor interactions.

Miller and Sitter (2005) proposed using all-subsets regression in two phases, with the first phase to determine the active main effects. Once main effects are identified, the practitioner is asked to identify interactions considered possible, and all-subset regression is used to augment the main effects

model. They illustrated this analysis with a numerical example based on taking 12 runs from a complete 2^5 factorial. The full factorial leads to a reasonable hierarchical model with three main effects and two 2-factor interactions. For the 12 treatment combinations Miller and Sitter used for their resolution IV design, the two active interactions are fortuitously orthogonal. Thus, their example is a best-case scenario. Since some pairs of interaction columns have correlations of $2/3$ or more, identifying a second interaction correctly with this design can be problematic.

7.5 Summary Regarding Choice of a Design

As we have proceeded through the sections of this chapter, the run size has become increasingly flexible. Although all designs in this chapter have $N \geq 2k$, regular resolution IV 2^{k-f} designs have N a power of 2, strength-3 orthogonal arrays have run sizes that are multiples of 8 and nonorthogonal resolution IV designs exist for N a multiple of 2. (If the main effects are orthogonal but the run size is not a multiple of 8, then $\mathbf{X}_1' \mathbf{X}_2 \neq \mathbf{0}$ and estimates for main effects will depend on which interaction terms are included in the model.) This flexibility in run size is important when there are a large number of factors. For instance, with $k = 10$, regular resolution IV designs are of size 32 or 64, and strength-3 arrays exist for $N = 24, 32, 40, \dots$ and a nonorthogonal design exists for N as small as 20. Since all of the designs in this chapter permit unbiased estimation of main effects, assuming three-factor and higher-order interactions are negligible, the choice of a design depends largely on how much information is desired regarding two-factor interactions.

If expert opinion regarding the presence of particular two-factor interactions is available, this can be utilized in the design choice. Recall the example at the end of Section 7.2.4. Additional literature that takes advantage of information about likely effects includes a customization of the minimum aberration criterion to fit the set of likely effects (Ke and Tang 2003) or the construction of an orthogonal design specific to the set of likely effects (Liao, Iyer, and Vecchia 1996).

Resolution V Fractional Factorial Designs

In many applications, one would like to estimate all main effects and two-factor interactions. This chapter presents designs for estimating such models, including regular resolution V 2^{k-f} fractional factorial designs, strength-4 orthogonal arrays, and nonorthogonal designs. The sections are as follows:

Section 8.1. Regular Resolution V 2^{k-f} Fractional Factorial Designs

Section 8.2. Strength-4 Orthogonal Arrays

Section 8.3. Three-Quarter Fractions of Regular Resolution V Designs

Section 8.4. Smaller Nonorthogonal Resolution V Designs

Section 8.5. Recommendations Regarding Design Choice

Section 8.6. Analysis of Resolution V Experiments

In the strictest sense, resolution V refers only to regular 2^{k-f} fractional factorial designs with the shortest word in the defining relation having length 5. Here, however, we follow Webb (1968, p. 291) by using resolution V in a broader sense as referring to any design—orthogonal or not—that permits estimation of all main effects and two-factor interactions. When regular resolution V fractional factorial designs require too many treatment combinations, one should consider the smaller alternative designs from Sections 8.2–8.4. Section 8.5 offers recommendations. The final section illustrates how to analyze experiments based on these designs.

8.1 Regular Resolution V 2^{k-f} Fractional Factorial Designs

Table 8.1 lists the run size N of the smallest resolution V regular fractions and the number of regression coefficients $r = 1 + 0.5k(k+1)$ to be estimated for the

two-factor interaction model (1.3), for $k = 5 - 24$ factors. For larger k , Draper and Lin (1990) summarized what was known until recently regarding the relationship between minimum N for a given k (or maximum k for a given N) for resolution V 2^{k-f} designs. Xu (2009) has improved two cases, identifying 2^{33-23} and 2^{47-36} designs of resolution V. Although there is no exact formula, note that for regular resolution V designs, the maximum number of factors k is approximately $N^{1/2}$.

Table 8.1. Size of smallest regular resolution V designs for $k = 5-24$ factors

No. Factors k	Run Size N	No. Parameters r	Remaining df $N - r$
5	16	16	0
6	32	22	10
7	64	29	35
8	64	37	27
9	128	46	82
10	128	56	72
11	128	67	61
12	256	79	177
13	256	92	164
14	256	106	150
15	256	121	135
16	256	137	119
17	256	154	102
18	512	172	340
19	512	191	321
20	512	211	301
21	512	232	280
22	512	254	258
23	512	277	235
24	1024	301	723

Generators for minimum aberration resolution V designs for up to 23 factors are given in Table 8.2. The designs of size 256 and larger are taken from Franklin (1984) and Xu (2009). Note that most regular fractions are substantially larger than one needs to estimate the two-factor-interaction model. For $k \geq 7$, no more than 60% of the degrees of freedom correspond to main effects and two-factor interactions. The remaining degrees of freedom, $N - r$, are useful for blocking, for checking the adequacy of the model that assumes no interactions involving three or more factors, and for estimating the error variance. However, in many cases, economizing on the run size is appropriate.

Sections 8.2–8.4 presents smaller alternative designs than the regular 2^{k-f} fractional factorials.

Table 8.2. Generators for smallest minimum aberration designs of resolution V (or more) for 5–23 factors

No. Factors	No. Runs	Design	Generator Columns	A_5	A_6
5	16	5-1.1	15	1	
6	32	6-1.1	31	0	1
7	64	7-1.1	63	0	0
8	64	8-2.1	45, 51	2	1
9	128	9-2.1	31, 121	0	3
10	128	10-3.1	15, 51, 121	3	3
11	128	11-4.1	15, 51, 85, 120	6	6
12	256	12-4.1	31, 107, 205, 241	0	12
13	256	13-5.1	103, 121, 157, 179, 207	3	12
14	256	14-6.1	31, 39, 107, 169, 243, 254	9	18
15	256	15-7.1	78, 109, 135, 171, 181, 211, 246	15	30
16	256	16-8.1	23, 46, 92, 113, 139, 184, 197, 226	24	44
17	256	17-9.1	23, 46, 92, 113, 139, 184, 197, 226, 255	34	68
18	512	18-9.1	47, 93, 185, 227, 279, 369, 395, 453, 511	0	102
19	512	19-10.1	105, 127, 143, 181, 211, 285, 307, 327, 427, 473	12	84
20	512	20-11.1	Design 19-10.1, plus 485	16	120
21	512	21-12.1	Design 20-11.1, plus 510	21	168
22	512	22-13.1	105, 127, 155, 188, 206, 275, 298, 301, 350, 358, 369, 391, 507	63	189
23	512	23-14.1	23, 90, 99, 127, 155, 188, 206, 301, 340, 358, 391, 430, 435, 450	84	252

8.2 Strength-4 Orthogonal Arrays

Table 8.1 indicates that for 128-run regular resolution V fractional factorials, the maximum number of factors is 11. However, there exists a 128-run nonregular orthogonal design for estimating all main effects and two-factor interactions for as many as 15 factors. Furthermore, this nonregular design can be run in blocks of size 16, with each block a regular resolution III fraction. Such designs are the topic of this section.

8.2.1 Fifteen-factor orthogonal design in 128 runs

We now present a strength-4 orthogonal array of size $N = 128$ for 15 factors. Recall that strength 4 means that this array projects into an equally replicated

2^4 factorial in every subset of four factors. The coding theory derivation of this and other strength-4 and strength-5 orthogonal arrays is documented by Hedayat, Sloane, and Stufken (1999). The most helpful construction for this 15-factor, 128-run orthogonal design is as the combination of eight 2^{15-11} fractions from the same family. Begin by constructing the first of eight 2^{15-11} blocks using the generators

$$\begin{aligned} \mathbf{E} &= \mathbf{AB} & \mathbf{F} &= \mathbf{AC} & \mathbf{G} &= \mathbf{BC} & \mathbf{H} &= \mathbf{ABC} & \mathbf{J} &= \mathbf{AD} & \mathbf{K} &= \mathbf{BD} \\ \mathbf{L} &= \mathbf{ABD} & \mathbf{M} &= \mathbf{CD} & \mathbf{N} &= \mathbf{ACD} & \mathbf{O} &= \mathbf{BCD} & \mathbf{P} &= \mathbf{ABCD}. \end{aligned}$$

These 16 runs constitute the saturated regular resolution III fraction (design 15-11.1) discussed earlier in Chapters 5 and 6.

For 7 additional sets of 16 runs, reverse the signs for 6 of the generators, as shown in the following array:

Block	E	F	G	H	J	K	L	M	N	O	P
1	+	+	+	+	+	+	+	+	+	+	+
2	+	-	+	-	+	-	-	+	-	-	+
3	-	+	-	+	+	-	+	-	-	-	+
4	-	+	+	-	-	-	-	-	+	+	+
5	+	+	-	-	-	-	+	+	-	+	-
6	-	-	+	+	+	+	-	-	+	-	-
7	+	-	-	+	-	+	-	-	-	+	+
8	-	-	-	-	-	+	+	+	+	-	+

For example, the second block is formed using

$$\begin{aligned} \mathbf{E} &= \mathbf{AB} & \mathbf{F} &= -\mathbf{AC} & \mathbf{G} &= \mathbf{BC} & \mathbf{H} &= -\mathbf{ABC} & \mathbf{J} &= \mathbf{AD} & \mathbf{K} &= -\mathbf{BD} \\ \mathbf{L} &= -\mathbf{ABD} & \mathbf{M} &= \mathbf{CD} & \mathbf{N} &= -\mathbf{ACD} & \mathbf{O} &= -\mathbf{BCD} & \mathbf{P} &= \mathbf{ABCD}. \end{aligned}$$

This 128-run design in 8 blocks of size 16 is remarkable, permitting estimation of all 15 main effects and 105 two-factor interactions, plus block main effects, orthogonally. Although composed of eight regular fractions, this design is not a regular 2^{15-8} fraction. Note that the sign of the last generator, $\mathbf{P} = \mathbf{ABCD}$, is reversed in only two of the eight blocks; thus, \mathbf{ABCDP} sums to $N/2 = 64$, not to zero or N , as is required for all regular fractions. Whereas a regular resolution V design completely aliases two-factor interactions with some three-factor interactions, this strength-4 orthogonal array does not, since no five-factor interaction is “+1” for all 128 treatment combinations. Its generalized resolution is 5.5.

8.2.2 Nineteen-factor orthogonal design in 256 runs

As we saw in Table 8.1, regular resolution V fractions of size $N = 256$ permit no more than 17 factors. Here, we construct an orthogonal design for 19 factors utilizing the OA(128, 2^{15} , 4) array from Section 8.2.1.

Begin by dividing a 2^4 factorial in four additional factors, \mathbf{Q} , \mathbf{R} , \mathbf{S} , and \mathbf{T} , into eight blocks, blocking on two-factor interactions:

Block	Q	R	S	T
1	-1	-1	-1	-1
	1	1	1	1
2	1	-1	-1	-1
	-1	1	1	1
3	-1	1	-1	-1
	1	-1	1	1
4	1	1	-1	-1
	-1	-1	1	1
5	-1	-1	1	-1
	1	1	-1	1
6	1	-1	1	-1
	-1	1	-1	1
7	-1	1	1	-1
	1	-1	-1	1
8	1	1	1	-1
	-1	-1	-1	1

Note that each block consists of a mirror-image pair of runs. To create an orthogonal array with 256 runs and 19 factors, take the product of the first block of the OA(128, 2^{15} , 4) from Section 8.2.1 and the first block of size 2 from this 2^4 ; that is, take the block 1, 2^{15-11} fraction for factors $\mathbf{A} - \mathbf{P}$; to each of these 16 treatment combinations append $\mathbf{Q} = \mathbf{R} = \mathbf{S} = \mathbf{T} = -1$ and $\mathbf{Q} = \mathbf{R} = \mathbf{S} = \mathbf{T} = +1$. This produces $16 \times 2 = 32$ runs for 19 factors. Create 32 more runs by combining block 2 from the OA(128, 2^{15} , 4) with block 2 from the 2^4 . Do the same construction for blocks 3–8 and combine the 8 sets of 32 runs into a 256-run, 19-factor design of strength 4. The resulting $8 \times 32 = 256$ runs of this orthogonal array should be run in random order, not in blocks, since these blocks are only resolution II. The design has generalized resolution of 5.

8.2.3 Other strength-4 orthogonal designs

Two more strength-4 orthogonal arrays described by Hedayat, Sloane, and Stufken (1999) are as follows:

- OA(2^{63} , 4). Having up to 63 factors is a huge improvement over regular designs, where resolution V with 2048 runs is currently possible only for $k \leq 47$ factors (Xu 2009, Table 17). Analogous to the 128-run design above, this orthogonal array is composed of 32 orthogonal blocks of size 64. For details regarding generators, see Mee (2004, p. 411).
- OA(2^{69} , 4). Although this orthogonal array allows for four more factors (69) than is currently possible with a regular 2^{k-f} design of resolution V, it still utilizes fewer than 60% of its degrees of freedom for main effects and two-factor interactions. For construction details, see Mee (2004, p. 411).

Given these nonregular orthogonal designs, Table 8.3 in an updated version of Table 8.1 that includes both regular and nonregular orthogonal designs. Note the reduction in run size for $k = 12, 13, 14, 15, 18$, and 19.

Table 8.3. Size of smallest regular resolution V or strength four orthogonal designs for $k = 5\text{--}20$ factors

No. Factors k	Design from Section	Run Size N	No. Parameters r	Remaining df $N - r$
5	8.1	16	16	0
6	8.1	32	22	10
7	8.1	64	29	35
8	8.1	64	37	27
9	8.1	128	46	82
10	8.1	128	56	72
11	8.1	128	67	61
12	8.2	128	79	49
13	8.2	128	92	36
14	8.2	128	106	22
15	8.2	128	121	7
16	8.1	256	137	119
17	8.1	256	154	102
18	8.2	256	172	84
19	8.2	256	191	65
20	8.1	512	211	301

8.3 Three-Quarter Fraction of Regular Resolution V Designs

Three-quarter fraction designs for up to 11 factors were proposed by Addelman (1961) and John (1961, 1962, 1969). These designs are 25% smaller than the smallest orthogonal resolution V designs. Each is constructed by dividing an orthogonal resolution V (or higher) design into four blocks and then discarding one block. The fundamental ideas are now detailed for the simple four-factor case, followed by a list of generators for designs with $k = 7, \dots, 11$ factors.

8.3.1 Four factors with run size $N = 3(2^{4-2}) = 12$

Partition the 2^4 factorial into four sets of runs, blocking on **ABC** and **ABD**. One block is

A	B	C	D
-1	-1	1	1
1	-1	-1	-1
-1	1	-1	-1
1	1	1	1

The other three blocks are obtained by reversing the signs in one or both of the last two columns. This four-run fraction has defining relation

$$\mathbf{I} = \mathbf{ABC} = \mathbf{ABD} = \mathbf{CD}$$

and complete alias chains

$$\begin{aligned}\mathbf{A} &= \mathbf{BC} = \mathbf{BD} = \mathbf{ACD} \\ \mathbf{B} &= \mathbf{AC} = \mathbf{AD} = \mathbf{BCD} \\ \mathbf{C} &= \mathbf{AB} = \mathbf{ABCD} = \mathbf{D}.\end{aligned}$$

Ignoring three-factor and higher-order interactions, the defining relation and alias chains reduce to

$$\begin{aligned}\mathbf{I} &= \mathbf{CD} \\ \mathbf{A} &= \mathbf{BC} = \mathbf{BD} \\ \mathbf{B} &= \mathbf{AC} = \mathbf{AD} \\ \mathbf{C} &= \mathbf{AB} = \mathbf{D}.\end{aligned}$$

These 11 terms correspond to the 11 columns of a model matrix that includes up to two-factor interactions. If one omits from the 2^4 the four runs at the top of this page, the resulting 12-run design is

A	B	C	D
-1	-1	-1	-1
-1	-1	-1	1
-1	-1	1	-1
-1	1	-1	1
-1	1	1	-1
-1	1	1	1
1	-1	-1	1
1	-1	1	-1
1	-1	1	1
1	1	-1	-1
1	1	-1	1
1	1	1	-1

From this three-fourths of a 2^4 , one can estimate all main effects and two-factor interactions, provided the five higher-order interactions do not exist. Arranging the columns of the 12×11 model matrix \mathbf{X} as

$$\{\mathbf{I}, \mathbf{CD}, \mathbf{A}, \mathbf{BC}, \mathbf{BD}, \mathbf{B}, \mathbf{AC}, \mathbf{AD}, \mathbf{C}, \mathbf{AB}, \mathbf{D}\},$$

the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is block diagonal:

$$\left[\begin{array}{cccccccccc} \frac{9}{8} & \frac{3}{8} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{3}{8} & \frac{9}{8} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{3}{2} & \frac{3}{4} & \frac{3}{4} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{3}{2} & \frac{3}{4} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{3}{4} & \frac{3}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{3}{2} & \frac{3}{4} & \frac{3}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{3}{4} & \frac{3}{2} & \frac{3}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{3}{4} & \frac{3}{4} & \frac{3}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{3}{2} & \frac{3}{4} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{3}{4} & \frac{3}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{3}{4} & \frac{3}{4} \end{array} \right] /12.$$

Thus, the first two estimators, b_0 and b_{CD} , have variance $\sigma^2(1.125)/12$, and the other nine estimators have variance $\sigma^2(1.5)/12$. Note how effects that are aliased in a single $(1/4)2^4$ are now correlated in the $(3/4)2^4$. This correlation is $(3/8)/(9/8) = 1/3$ for b_0 and b_{CD} , the effects aliased in a (reduced) chain of length 2, and $(3/4)/(3/2) = 1/2$ for effects aliased in (reduced) chains of length 3. These correlations increase the variance from σ^2/N (if the columns were orthogonal) to either $1.125\sigma^2/N$ or $1.5\sigma^2/N$, respectively. Thus, the VIFs are 1.125 and 1.5.

Note that the $(3/4)2^4$ design above provides higher precision for b_{CD} than for the other 10 factorial effects. The alternative $3/4$ fraction obtained by excluding the runs with $\mathbf{I} = \mathbf{A} = \mathbf{BCD} = \mathbf{ABCD}$ results in the main effect b_A being estimated with better precision than the other factorial effects.

One cannot construct a $(3/4)2^4$ design using a defining relation that aliases four effects of interest. For instance, the four observations with $\mathbf{I} = \mathbf{A} = \mathbf{BC} = \mathbf{ABC}$ aliases $\mathbf{B} = \mathbf{AB} = \mathbf{C} = \mathbf{AC}$. Excluding this fourth of the 2^4 creates a singular $\mathbf{X}'\mathbf{X}$ matrix, making it impossible to estimate all four of these coefficients from this $3/4$ fraction.

Finally, the 12-run designs above cannot be run as a randomized block design in three blocks of size 4. Doing so would confound an effect of interest with blocks, because each 2^{4-2} is only resolution I or II. So the 12 treatment combinations should be conducted as a completely randomized design. Our use of the term “block” here is just a tool for defining the treatment combinations in the design.

We now present three-quarter fraction designs of run size $N = 48$ for 7 and 8 factors and designs of run size $N = 96$ for 9–11 factors. If resolution V or higher designs are partitioned into four blocks, and one block of runs is omitted, the remaining runs constitute a nonorthogonal design, with correlations among effects that are aliased in a single block. Provided no alias set

consists of more than three effects to be estimated, the model of interest is estimable from just three blocks, although with less precision than if the N runs formed an orthogonal design.

8.3.2 Seven factors with run size $N = 3(2^{7-3}) = 48$

Two 48-run designs are presented, each based on combining three resolution III 2^{7-3} fractions together.

1. *Three-quarters of the resolution VI 2^{7-1} design.* Generate the resolution VI fraction using $\mathbf{G} = \mathbf{ABCDF}$. Partition this into four sets of runs using **ABE** and **ACF**; each set of runs corresponds to Chen, Sun, and Wu's (CSW's) (1993) design 7-3.3. This 48-run design gives the highest precision to the four two-factor interactions **CD**, **CG**, **DF**, and **FG**, since these effects are clear of aliasing in the 2^{7-3} fraction.
2. *Three-quarters of the resolution V 2^{7-1} design.* Generate the resolution V fraction using $\mathbf{G} = \mathbf{ABCD}$. Partition this into four sets of runs using **ABE** and **ACF**; each set of runs corresponds to CSW's design 7-3.2. This design gives the highest precision to the main effects **D** and **G** and the two-factor interactions **AD** and **AG**. This is the design proposed by Addelman (1961, p. 494)

Any of these 48-run designs may be run in three blocks. Designs 7-3.2 and 7-3.3, the two 16-run fractions from which these three-quarter fractions are built, both have $A_3 + A_4 = 5$. Since words of length 3 and 4 each produce three alias pairs, these designs have the same variance efficiency. If there is a possibility that one might complete the full 2^{7-1} by running the fourth block, the resolution VI design using blocks corresponding to design 7-3.3 would be preferred.

8.3.3 Eight factors with run size $N = 3(2^{8-4}) = 48$

This design was presented first by Addelman (1961) and John (1962). See also Mee (2004). Construct the design as follows:

- Generate the resolution V 2^{8-2} design with $\mathbf{G} = \mathbf{ABCD}$ and $\mathbf{H} = \mathbf{ACEF}$.
- Divide the resolution V design into four sets of runs using the contrasts **ABE** and **CDF**. Omit one set.
- These 48 runs can be performed as a randomized block design in 3 blocks of size 16, since no main effects or two-factor interactions are aliased with **ABE**, **CDF**, or **ABCDEF**.

The omitted set of 16 runs is a resolution III fraction corresponding to CSW's design 8-4.3. No other fraction of any resolution will produce a 3/4 fraction with fewer correlations among the estimators of the two-factor interaction model.

The two designs for $k = 7$ are projections of this eight-factor design. Drop factor **A**, **C**, **G**, or **H** and get $3/4$ of the resolution VI fraction. Drop factor **B**, **D**, **E**, or **F** and get $3/4$ of the resolution V fraction.

8.3.4 Nine factors with run size $N = 3(2^{9-4}) = 96$

The optimum $3/4$ of a 2^{9-2} design is presented in Mee (2004). This $N = 96$ run design may be constructed as follows:

- Generate the resolution VI 2^{9-2} design with **H** = **BDEFG** and **J** = **ACEFG**.
- Divide the resolution VI design into four sets of runs using the contrasts **ABF** and **CDG**. Omit one set.
- The remaining 96 runs can be performed as a randomized block design in three blocks of size 32, since no main effects or two-factor interactions are aliased with **ABF**, **CDG**, or **ABCDEFG**.

Each set of 32 treatment combinations in this design corresponds to CSW's design 9-4.8, with $wlp = (2, 3, 6, 4, 0, 0, 0)$. The 5 shortest words are **ABF**, **CDG**, **ACEH**, **BDEJ**, and **FGHJ**; these produce the 15 pairs of aliases in a single block of 32 runs, and the correlation for each of 15 pairs of effects in the 96-run design. Addelman (1961) proposed a $(3/4)2^{9-2}$ design based on three blocks of CSW's design 9-4.5, with $wlp = (1, 5, 6, 2, 1, 0, 0)$. While our individual blocks have higher aberration, they produce a more efficient $3/4$ fraction design, with fewer correlations. (This design will be discussed again in Section 10.4 in the context of sequential assembly of fractions.)

8.3.5 Ten factors with run size $N = 3(2^{10-5}) = 96$

John (1969) proposed a $3/4$ fraction design for 10 factors equivalent to the following:

- Generate the minimum aberration resolution V 2^{10-3} design with **H** = **BDEFG**, **J** = **ACEFG**, and **K** = **CDEF**.
- Divide the resolution V design into 4 sets of 32 runs using the contrasts **ABF** and **CDG**. Omit one set.
- The remaining 96 runs can be performed as a randomized block design in 3 blocks of size 32, since no main effect or two-factor interaction is aliased with **ABF**, **CDG**, or **ABCDEFG**.

Each block of 32 runs is a resolution III fraction corresponding to CSW's design 10-5.7 with $wlp = (2, 7, \dots)$; all other 2^{10-5} designs have $A_3 + A_4 \geq 10$.

8.3.6 Eleven factors with run size $N = 3(2^{11-6}) = 96$

Two 96-run designs are described, each based on taking a 3/4 fraction of the resolution V 2^{11-4} fraction. John (1969) proposed elimination of a resolution II block of size 32; the second design presented here is constructed using resolution III blocks.

1. *John's (1969) 3/4 of the resolution V 2^{11-4} design.* Define the resolution V fraction using $\mathbf{H} = \mathbf{ABCF}$, $\mathbf{J} = \mathbf{ADEG}$, $\mathbf{K} = \mathbf{BEFG}$, and $\mathbf{L} = \mathbf{CDEF}$. Partition this into 4 sets of 32 runs using \mathbf{BD} and \mathbf{ACG} . Eliminating 1 set of runs produces a 96-run design with just 44 correlations among the effects in a model containing main effects and two-factor interactions ($k = 11$ from the length-2 word and 33 from the 11 words of length 3 and 4). Because the sets of runs are resolution II, this design cannot be performed as a randomized block design without confounding a two-factor interaction with blocks.
2. *Three-quarters of the resolution V 2^{11-4} design in three blocks.* Define the resolution V fraction using the same generators for \mathbf{H} , \mathbf{J} , \mathbf{K} , and \mathbf{L} as above. Partition this 128-run design into 4 sets of 32 runs using \mathbf{ACDF} and \mathbf{BCEG} ; each set of runs is a resolution III 2^{11-6} fraction with $wlp = (5, 10, \dots)$. Chen et al. (1993) did not list any 2^{11-6} design with five or more length-3 words; however, all of the designs they listed have $A_3 + A_4 > 15$, and so are inferior for the purpose of constructing a 3/4 fraction. Although this design has 45 correlations among estimates for main effects and two-factor interactions, 1 more than for the design proposed by John (1969), it can be run in 3 blocks without confounding any two-factor interactions with blocks.

We do not consider three-quarter fraction designs for $k = 12\text{--}15$ or 18 and 19 factors, since the orthogonal designs from Section 8.2 are half the size of the smallest regular resolution V fraction. Identifying a three-quarter fraction of size $N = 3(64) = 192$ for $k = 16$ or 17 factors is an open research problem.

8.4 Smaller Nonorthogonal Resolution V Designs

Smaller nonorthogonal designs than those proposed in Section 8.3 have been constructed by several methods:

1. Partition a design of resolution (\geq)V into eight (or more) blocks. Then run the minimum number of blocks that permits estimating all main effects and two-factor interactions.
2. Combine two fractions not from the same family of fractions
3. Rechtschaffner (1967) proposed a series of two-level designs for estimating all main effects and two-factor interactions with the minimal number of runs; that is, $N = 1 + k + .5k(k - 1)$, so that there are no degrees of freedom for error.

4. D-Optimal and A-optimal designs of various sizes have been constructed via search with numerical algorithms.

Resolution V designs constructed by each of these methods will now be mentioned, including only those designs considered the most useful.

8.4.1 Additional irregular fractions

Section 8.3 discussed taking three-fourths of resolution V fractions. Other irregular fractions have been proposed to construct resolution V designs. For example, Addelman (1969) proposed two different means of augmenting the resolution IV 2^{7-2} fraction. In both his design 7.1A and 7.4A, the first 32 runs constitute the regular 2^{7-2} design with defining relation $\mathbf{I} = \mathbf{ABCDE} = \mathbf{ABCDEF} = \mathbf{DEFG}$. If one were to add 32 runs defined by $\mathbf{I} = \mathbf{ABCDE} = -\mathbf{ABCDEF} = -\mathbf{DEFG}$, one would have an orthogonal resolution V 2^{7-1} design. If, instead, one adds only 8 of the 32 runs from the second fraction (e.g., the subset with $\mathbf{A} = \mathbf{B}$ and $\mathbf{F} = \mathbf{AD}$), one still may estimate the two-factor-interaction model. The subset of eight runs can be designated as a second block, but this $(5/8)2^{7-1}$ fraction cannot be performed in five blocks of eight since, then, the \mathbf{AB} interaction would be confounded with blocks.

More sequential designs of this type for $k = 3, \dots, 11$ factors appear in Addelman (1969). These will be considered more fully in Section 10.4.

8.4.2 Combining two lower-resolution regular fractions

The following 64-run design constructed from two resolution I fractions of size 32 permits estimation of all 10 main effects and 45 two-factor interactions with good precision:

- $\mathbf{F} = \mathbf{ACD}, \mathbf{G} = \mathbf{ADE}, \mathbf{H} = \mathbf{ABCE}, \mathbf{J} = \mathbf{CDE}, \mathbf{K} = +1$
- $\mathbf{F} = \mathbf{ABCD}, \mathbf{G} = \mathbf{BCE}, \mathbf{H} = \mathbf{BDE}, \mathbf{J} = -\mathbf{CDE}, \mathbf{K} = -1$

Before adding factor \mathbf{K} , each half of the design is a resolution IV 2^{9-4} fraction with $A_4 = 7$. When we combine the two fractions, some elements of the defining relations cancel, namely \mathbf{K} and \mathbf{CDEJ} , since both are $+1$ for the first 32 runs and -1 for the last 32.

After these two words cancel, we are left with a fraction that is strength 3 (meaning we have an equally replicated 2^3 in every subset of three columns). The design sums to ± 32 in 12 sets of four-factor interactions. This design supports estimation of the 10 main effects orthogonally, plus all 45 two-factor interactions. Thus, this is a nonorthogonal resolution V design that is half the size of the smallest orthogonal design of resolution V or VI.

By eliminating factor \mathbf{G} , only four subsets of four factors do not sum to zero. Thus, the recommended design for nine factors is

- $\mathbf{F} = \mathbf{ACD}, \mathbf{H} = \mathbf{ABCE}, \mathbf{J} = \mathbf{CDE}, \mathbf{K} = +1$
- $\mathbf{F} = \mathbf{ABCD}, \mathbf{H} = \mathbf{BDE}, \mathbf{J} = -\mathbf{CDE}, \mathbf{K} = -1$

This nine-factor design's D-efficiency = 92.3% relative to an orthogonal design (see Mee 2004, p. 404).

8.4.3 Saturated resolution V designs

Rechtschaffner (1967) proposed several series of saturated resolution V designs that are simple to construct. His best series of designs consists of the following treatment combinations:

- The treatment combination with all factors low.
- $k(k - 1)/2$ treatment combinations with two factors high and all other factors low.
- k treatment combinations with one factor low and all other factors high.

For instance, for $k = 6$, Rechtschaffner's design is given in Table 8.4. These designs are easily constructed from a full 2^k by defining a variable that is the sum $x_1 + \dots + x_k$ and only retaining the treatment combinations for which the sum is $-k$, $4 - k$, and $k - 2$. For $k = 6$, these sums are -6 , -2 , and 4 .

Table 8.4. Rechtschaffner's 22-run resolution V design for $k = 6$ factors

A	B	C	D	E	F	Sum
-1	-1	-1	-1	-1	-1	-6
-1	-1	-1	-1	1	1	-2
-1	-1	-1	1	-1	1	-2
-1	-1	-1	1	1	-1	-2
-1	-1	1	-1	-1	1	-2
-1	-1	1	-1	1	-1	-2
-1	-1	1	1	-1	-1	-2
-1	1	-1	-1	-1	1	-2
-1	1	-1	-1	1	-1	-2
-1	1	-1	1	-1	-1	-2
-1	1	1	-1	-1	-1	-2
1	-1	-1	-1	-1	1	-2
1	-1	-1	-1	1	-1	-2
1	-1	1	-1	-1	-1	-2
1	1	-1	-1	-1	-1	-2
-1	1	1	1	1	1	4
1	-1	1	1	1	1	4
1	1	-1	1	1	1	4
1	1	1	-1	1	1	4
1	1	1	1	-1	1	4
1	1	1	1	1	-1	4

Qu (2007) derived a closed-form expression for the least squares estimators and their covariance matrix for Rechtschaffner's designs. Table 8.5 shows the variances for the intercept and other regression coefficients and the equivalent sample size if the design were orthogonal, for $k = 4, \dots, 10, 12, 15$, and 20.

Table 8.5. Variances of estimators for Rechtschaffner's resolution V designs

k	N	$\text{Var}(b_0)/\sigma^2$	$\text{Var}(b_i)/\sigma^2$	$\sigma^2/\text{Var}(b_i)$
4	11	0.0972	0.1389	7.20
5	16	0.0625	0.0625	16.00
6	22	0.0550	0.0522	19.15
7	29	0.0764	0.0503	19.86
8	37	0.1276	0.0504	19.84
9	46	0.2090	0.0510	19.61
10	56	0.3210	0.0517	19.34
12	79	0.6374	0.0531	18.84
15	121	1.3444	0.0547	18.28
20	211	3.1451	0.0565	17.70

For $k = 4, 5$, and 6, Rechtschaffner's designs are D-optimal. For $k = 5$, Rechtschaffner's design is the regular resolution V 2^{5-1} fraction. Although Rechtschaffner's designs are simple to construct for any number of factors k , the variance of the estimators is very poor for eight or more factors, considering the run size. Note how the variances in Table 8.5 do not decrease after $k = 7$, even as the size of the design increases.

Tobias (1996) proposed a series of saturated designs that corresponds to Rechtschaffner's designs for $k = 4, 5, 6$, but is better for $k \geq 7$. For $k = 7$, see Table 8.6. Note that the difference between the Rechtschaffner design and the Tobias design for $k = 7$ is that instead of containing all $\binom{k}{2} = 21$ treatment combinations with sum $4 - k$, Table 8.6 contains 11 points with sum $4 - k$ —excluding 10 runs where both **A** and **B** are -1 and replacing these with 10 points where both **A** and **B** are -1 and the sum is $6 - k$. The design in Table 8.6 is both D-optimal and A-optimal. For $k \geq 8$, Tobias's designs are not optimal, but they are much more efficient than Rechtschaffner's designs.

8.4.4 D-Optimal and A-optimal resolution V designs

Optimal designs may be constructed for any $N \geq r$ (i.e., for any number of runs at least as large as the number of parameters to be estimated). One simply specifies the design size N , the number of factors k , and the model—here one that includes an intercept, all main effects, and two-factor interactions. Let \mathbf{X} denote the $N \times r$ model matrix. The following criteria are popular for assuring that the covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is as small as possible:

- D-Optimal designs minimize the determinant of $(\mathbf{X}'\mathbf{X})^{-1}$.
- A-Optimal designs minimize the trace of $(\mathbf{X}'\mathbf{X})^{-1}$.

Table 8.6. Tobias's 29-run resolution V design for $k = 7$ factors

A	B	C	D	E	F	G	Sum
-1	-1	-1	-1	-1	-1	-1	-7
-1	1	-1	-1	-1	-1	1	-3
-1	1	-1	-1	-1	1	-1	-3
-1	1	-1	-1	1	-1	-1	-3
-1	1	-1	1	-1	-1	-1	-3
-1	1	1	-1	-1	-1	-1	-3
1	-1	-1	-1	-1	-1	1	-3
1	-1	-1	-1	-1	1	-1	-3
1	-1	-1	-1	1	-1	-1	-3
1	-1	-1	1	-1	-1	-1	-3
1	-1	1	-1	-1	-1	-1	-3
1	1	-1	-1	-1	-1	-1	-3
-1	-1	-1	-1	1	1	1	-1
-1	-1	-1	1	-1	1	1	-1
-1	-1	-1	1	1	-1	1	-1
-1	-1	-1	1	1	1	-1	-1
-1	-1	1	-1	-1	1	1	-1
-1	-1	1	-1	1	-1	1	-1
-1	-1	1	1	-1	-1	1	-1
-1	-1	1	1	-1	1	-1	-1
-1	1	1	1	1	1	1	5
1	-1	1	1	1	1	1	5
1	1	-1	1	1	1	1	5
1	1	1	-1	1	1	1	5
1	1	1	1	-1	1	1	5
1	1	1	1	1	-1	1	5
1	1	1	1	1	1	-1	5

For an orthogonal resolution V design, $\mathbf{X}'\mathbf{X} = N\mathbf{I}_r$, where \mathbf{I}_r is an $r \times r$ identity matrix, and so $|(\mathbf{X}'\mathbf{X})^{-1}| = 1/|\mathbf{X}'\mathbf{X}| = N^{-r}$ and $\text{trace}(\mathbf{X}'\mathbf{X})^{-1} = r/N$. For most design sizes N , orthogonal designs do not exist, and the determinant and trace of $(\mathbf{X}'\mathbf{X})^{-1}$ are larger as a result.

Many commercial statistical software packages are capable of constructing D-optimal designs or A-optimal designs. Some require an initial candidate set (here generally a full 2^k). All require that a model be specified as well as a design size N at least as large as the number of parameters to be estimated.

Sometimes an optimal design has recognizable structure. For instance, the 10-factor, 64-run design in Section 8.4.2 and Tobias's 7-factor design in Table 8.6 were first obtained with a D-optimal design search, and then characterized as having a certain structure.

Nguyen and Dey (1989) produced an efficient, stand-alone computer algorithm for constructing D-optimal resolution V designs and published the maximum $|\mathbf{X}'\mathbf{X}|$ achieved for $k = 4, 5, 6$, for various run sizes. Nguyen and Miller (1997) produced tables of minimum $\text{trace}(\mathbf{X}'\mathbf{X})^{-1}$ for $k = 7\text{--}10$ and many run sizes. The Nguyen and Miller tables show the rapid gain in efficiency that results from increasing the sample size. For instance, the A-optimal saturated design for $k = 7$ (the design in Table 8.6) has $\text{trace}(\mathbf{X}'\mathbf{X})^{-1} = 1.334$, and A-efficiency of 75% relative to an orthogonal design, since

$$(r/N)/\text{trace}(\mathbf{X}'\mathbf{X})^{-1} = (29/29)/1.3304 = 0.75.$$

If one increases the design size from 29 to 34 (an increase of 17%), $\text{trace}(\mathbf{X}'\mathbf{X})^{-1}$ is reduced by nearly 25% to 1.0077, and the A-efficiency increases to 84.6%. In addition to this marked decrease in the variance, the five extra runs provide 5 df for estimating the error variance (or checking for lack of fit). Unless one is certain that a saturated design will provide sufficient precision, using a few extra runs seems prudent. Optimal design algorithms provide the most convenient means for constructing resolution V designs of size N slightly larger than r . If one is certain of the model to be estimated, these designs are reasonable choices.

8.5 Recommendations Regarding Design Choice

The following designs are highly efficient for estimating a model with all main effects and two-factor interactions, both in terms of economy of run size and in terms of being orthogonal or having high variance efficiency.

- $k = 5$: The 2^{5-1} fractional factorial design with $\mathbf{E} = \mathbf{ABCD}$.
- $k = 6$: The 2^{6-1} fractional factorial design with $\mathbf{F} = \mathbf{ABCDE}$, which can be run in two blocks. If this design is too expensive, Rechtschaffner's saturated design (in Table 8.4) or a slightly larger optimal design is recommended.
- $k = 7$: One of the 48-run $(3/4)2^{7-1}$ designs from Section 8.3.2, which can be run in three blocks. If this design is too expensive, consider the 40-run $(5/8)2^{7-1}$ design from Addelman (1969) mentioned in Section 8.4.1.
- $k = 8$: The 48-run $(3/4)2^{8-2}$ design in Section 8.3.3.
- $k = 9$: The 64-run design for nine factors in Section 8.4.2.
- $k = 10$: The $(3/4)2^{10-3}$ design in Section 8.3.5.
- $k = 11$: The $(3/4)2^{11-4}$ design in Section 8.3.6.
- $k = 12\text{--}15$: The 128-run orthogonal array presented in Section 8.2.1.

- $k = 16\text{--}17$: Partition the $n = 256$ run resolution V fraction from Table 8.2 into blocks of size 32 (or smaller) and run a subset of the blocks until one gains adequate precision. (This is analogous to smaller sequences discussed in Section 10.4.)
- $k = 18\text{--}19$: The 256-run orthogonal array presented in Section 8.2.2.

These recommendations provide a starting point for your design choice. The final choice will depend on (i) the reasonableness of the assumption that higher-order interactions can be ignored, (ii) the magnitude of the error variance, (iii) the ease of conducting follow-up runs, and (iv) any budgetary constraints. That is, uncertainty about higher-order interactions, more error variance, and difficulty in conducting follow-up runs tend to justify running larger experiments.

If running the design in blocks is necessary, this may be the deciding factor. For example, only one of the $(3/4)2^{11-4}$ designs in Section 8.3.6 can be run with each 2^{11-6} as a block without confounding effects of interest with blocks. More will be said in Chapter 10 about running fractional factorial designs as randomized block experiments.

8.6 Analysis of Resolution V Experiments

In general, one should analyze larger resolution V experiments by fitting at least three models. First, fit the two-factor interaction model (1.3). Then consider a model with additional terms to investigate the assumption that no higher-order interactions are needed. Finally, fit a parsimonious reduced model, eliminating interaction terms that appear unimportant.

A wide variety of designs have been discussed in this chapter, and some aspects of the analysis are design dependent. Therefore, it will be instructive to provide analysis details here for several cases. We begin by analyzing an orthogonal 128-run subset from Wang et al.'s (1993) 2^9 data. Our analysis in Section 4.2 of the full factorial revealed many higher-order interactions. Thus, these data will provide a challenging application for identifying effects. We then analyze a $(3/4)2^{9-2}$ fraction and, finally, a 64-run irregular fraction to show how the decrease in design size lowers the precision and limits one's ability to explore lack-of-fit. Our final analysis is for a saturated resolution V design; the data are from Le Thanh, Voilley, and Luu (1993) and involve a seven-factor experiment concerning the volatility of three food additives.

8.6.1 Example 8.1: Analysis of a regular fraction of resolution V (or more)

When Wang et al.'s (1993) 2^9 data were presented and analyzed in Section 4.2, we used $x_1\text{--}x_9$ to identify the nine factors. Here we will use uppercase letters to identify the factors, as is more common for fractional factorial designs. We

use the letters **A–E** to denote the N-terminal residues $x_1–x_5$ and **J–M** to denote the C-terminal residues $x_6–x_9$. Now, suppose only the quarter fraction with **L = BDEJK** and **M = ACEJK** is available. For this resolution VI fraction, we fit three models:

- A model with all main effects and two-factor interactions.
- A saturated model, to investigate the possible importance of higher-order terms.
- A reduced model including only terms that are statistically significant or that are required to make the model hierarchical. A few higher-order terms with p -values below .05 may be ignored to avoid making the model overly complex.

As in Section 4.2, we model the transformed response T rather than P . Even though T 's distribution is very negatively skewed, it uses the variance-stabilizing transformation for Poisson counts; the transformation also has the effect of emphasizing the variation for large P , which are the values of primary interest. A model with all main effects and two-factor interactions utilizes only 45 of the 127 df. A summary of this fitted model for T appears in Tables 8.7 and 8.8. The four C-terminal main effects (**J, K, L, M**) and their six two-factor interactions all stand out as both statistically significant and much larger than any lower-order effect involving the other five factors. The only other estimates with p -values $< .05$ are the main effect for **D** and **D**'s interaction with **L** and **M**. Any subsequent models should certainly include these 13 terms, as they form a hierarchical model.

The initial model with all 45 lower-order terms explains $5.694/6.560 = 87\%$ of the variation in T . The t -tests in Table 8.8 assume that the fitted model (or something simpler) is correct. What if some important higher-order interactions have been omitted from our initial model? What would be the consequence of this incorrect model specification?

- Missing higher-order interactions that are aliased with terms in the model will bias estimates for the lower order effects.
- Missing higher-order interactions that are not aliased with terms in the model will inflate the mean square error, making the t statistics smaller and the corresponding p -values larger than they should be.

Table 8.7. Two-factor-interaction model ANOVA for T with resolution VI 2^{9-2} fraction

Source	df	SS	MS	F-Ratio
Model	45	5.694	0.1265	11.98
Error	82	0.866	0.0106	
Total (corrected)	127	6.560		

Table 8.8. Initial model for T with resolution VI 2^{9-2} fraction; main effect and two-factor interaction estimates sorted

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	0.8236	0.0091	90.65	<.0001
M	-0.0958	0.0091	-10.54	<.0001
L	-0.0847	0.0091	-9.32	<.0001
K	-0.0759	0.0091	-8.36	<.0001
KM	-0.0588	0.0091	-6.47	<.0001
J	-0.0560	0.0091	-6.16	<.0001
KL	-0.0550	0.0091	-6.06	<.0001
LM	-0.0524	0.0091	-5.77	<.0001
JM	-0.0500	0.0091	-5.51	<.0001
JK	-0.0460	0.0091	-5.07	<.0001
JL	-0.0421	0.0091	-4.64	<.0001
D	-0.0297	0.0091	-3.27	0.0016
DL	-0.0213	0.0091	-2.34	0.0217
DM	-0.0183	0.0091	-2.01	0.0477
BM	0.0169	0.0091	1.86	0.0663
AE	-0.0131	0.0091	-1.44	0.1544
AB	-0.0129	0.0091	-1.42	0.1591
EJ	-0.0119	0.0091	-1.31	0.1936
AM	-0.0105	0.0091	-1.16	0.2499
B	0.0103	0.0091	1.14	0.2587
EM	0.0096	0.0091	1.06	0.2926
A	0.0094	0.0091	1.04	0.3029
BL	0.0092	0.0091	1.01	0.3148
BJ	-0.0088	0.0091	-0.96	0.3379
DJ	-0.0087	0.0091	-0.96	0.3391
BC	-0.0083	0.0091	-0.91	0.3631
CD	-0.0081	0.0091	-0.89	0.3751
EL	-0.0076	0.0091	-0.83	0.4063
CK	-0.0075	0.0091	-0.83	0.4086
AK	0.0069	0.0091	0.76	0.4518
DK	-0.0065	0.0091	-0.71	0.4772
AJ	-0.0064	0.0091	-0.70	0.4857
CM	0.0058	0.0091	0.64	0.5228
BD	-0.0056	0.0091	-0.62	0.5369
AL	0.0050	0.0091	0.55	0.5817
CL	0.0043	0.0091	0.48	0.6354
DE	-0.0041	0.0091	-0.45	0.6544
EK	-0.0041	0.0091	-0.45	0.6565
AD	0.0040	0.0091	0.44	0.6586
AC	0.0029	0.0091	0.31	0.7544
CE	-0.0023	0.0091	-0.25	0.8021
C	0.0019	0.0091	0.21	0.8314
CJ	-0.0017	0.0091	-0.19	0.8504
BE	-0.0013	0.0091	-0.15	0.8839
BK	-0.0007	0.0091	-0.08	0.9390
E	-0.0004	0.0091	-0.05	0.9623

To explore the possibility of model misspecification, we need to add other terms. So we fit a full factorial model, and from its 127 effect estimates, we compute Lenth's PSE = 0.00875 as the standard error for the b_i 's. The corresponding estimate for σ^2 is $N(PSE^2) = 0.0098$, which is similar to the MSE from the two-factor interaction model. Based on Lenth t statistics, the same main effects and two-factor interactions are statistically significant at $\alpha = .05$. The largest estimate for a higher-order interaction is for **KLM**, with Lenth $t = -3.20$ (p -value = .002). The next largest higher-order estimates are for **ABJK** = **ADEL** = **BCEM** = **CDJKLM** and **JLM** = **ABCDJ** = **ACEKL** = **BDEKM**, both with Lenth $|t| = 2.33$ (p -value = .02). There are five other terms with p -values between .025 and .05 (**BLM**, **ADM**, **ADJK**, **DJKM**, **JKL** and their aliases), but the likelihood of many Type I errors here makes us inclined to ignore most of these terms.

We now choose a reduced model. In addition to the five main effects and eight two-factor interactions with p -values $< .05$ in Table 8.8, we add **KLM**, **JLM**, and **JKL**. These 16 terms form a hierarchical model. We choose to ignore five terms with Lenth t p -values between .02 and .04 from the saturated model, so as not to excessively complicate the model. If there were 100–110 inactive effects in the saturated model, we would expect 5 or 6 to have p -values $< .05$. The largest Lenth t for a term we omit (**ABJK**) is $t = 2.33$; to include **ABJK** would require that an additional 11 terms be added to keep the model hierarchical. A summary of the chosen reduced model appears in Table 8.9. This parsimonious model explains 86% of the variation in T .

After fitting a reduced model, one should always examine the residuals. One benefit of having so many degrees of freedom for error is that loss of a few observations (due to being missing or as suspicious outliers) will not seriously impact our ability to estimate the model. The residual-versus-predicted plot for our reduced model (see Figure 8.1) shows the effect of truncation in the measurements for large T . Although there are no obvious outliers in this quarter fraction of the 2^9 , there is larger variation in the vicinity of $\hat{T} = 0.7$, which our model does not explain. This means our model does not predict well when $E(P) \approx 50\%$.

Before leaving this example, we compare our results to those in Table 4.3, in which we listed a reduced model based on the full 2^9 . Note first that the standard error here is double what it was in Table 4.3. This is due to having standard errors equal to $\sigma/128^{1/2}$ rather than $\sigma/512^{1/2}$; the estimate for σ is roughly the same as before. With this quarter fraction, we have selected a reduced model that contains 15 of the largest 16 estimates in Table 4.3. The only term among these we are missing is **DJKLM** (= **BEM** in our fraction, with estimate $b_{BEM} = 0.015$). For estimates that are statistically significant for this resolution VI fraction, aliasing has not caused any confusion. Analysis of this orthogonal fraction is straightforward. We now illustrate fitting models for two nonorthogonal fractions of the 2^9 .

Table 8.9. Parsimonious model for T with resolution VI 2^{9-2} fraction

(a) Analysis of Variance				
Source	df	SS	MS	F-Ratio
Model	16	5.635	0.3522	42.3
Error	111	0.925	0.0083	
Total (corrected)	127	6.560		

(b) Parameter estimates				
Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	0.824	0.008	102.06	.0000
D	-0.030	0.008	-3.68	.0004
J	-0.056	0.008	-6.93	.0000
K	-0.076	0.008	-9.41	.0000
L	-0.085	0.008	-10.49	.0000
M	-0.096	0.008	-11.87	.0000
DL	-0.021	0.008	-2.63	.0096
DM	-0.018	0.008	-2.26	.0256
JK	-0.046	0.008	-5.70	.0000
JL	-0.042	0.008	-5.22	.0000
JM	-0.050	0.008	-6.20	.0000
KL	-0.055	0.008	-6.82	.0000
KM	-0.059	0.008	-7.28	.0000
LM	-0.052	0.008	-6.50	.0000
JKL	-0.017	0.008	-2.15	.0334
JLM	-0.020	0.008	-2.53	.0129
KLM	-0.028	0.008	-3.47	.0007

8.6.2 Example 8.2: Analysis of a resolution V three-quarter fraction

In Section 8.3.4, a 96-run ($3/16^{\text{th}}$) irregular fraction for nine factors was proposed. Excluding the 32 runs for which $\mathbf{ABJ} = \mathbf{CDK} = -1$ from the resolution VI 2^{9-2} fraction just analyzed, one obtains the $3/16$ fraction examined here. We begin by fitting a model with all main effects and two-factor interactions, just as in the previous subsection. The ANOVA and the sorted parameter estimates are displayed in Table 8.10. Excluding $1/4^{\text{th}}$ of the resolution VI fraction has the following impact on our initial analysis:

- The estimates are no longer uncorrelated with common variance σ^2/N . Thirty of the estimates are correlated in pairs (with correlations $\pm 1/3$). These estimates have variance $(9/8)\sigma^2/N$.

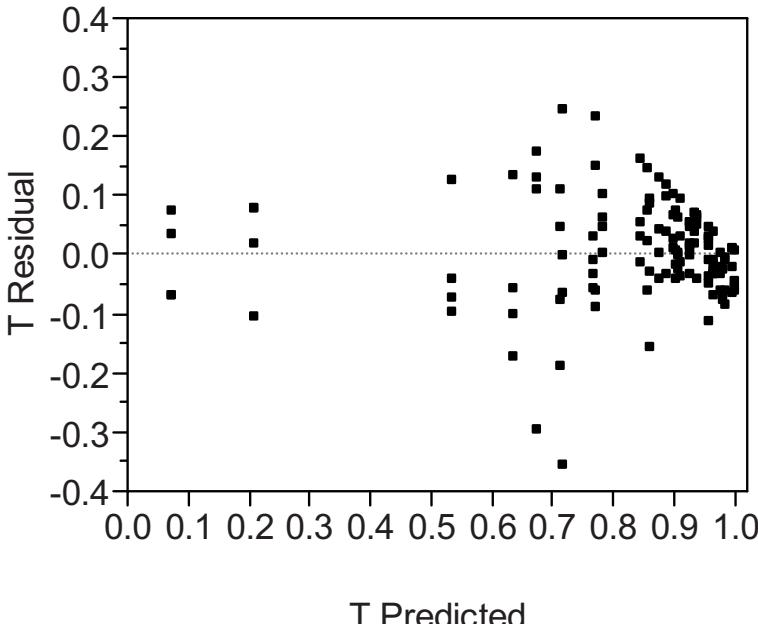


Fig. 8.1. Residual plot for reduced model in Table 8.9

- With 32 fewer observations, we have fewer degrees of freedom for error. However, 50 df is plenty for estimating σ^2 , assuming this initial model is close to being adequate.
- If we add higher-order terms, this will change our estimates for the lower-order effects, potentially decreasing their bias, while increasing their variance. The latter will be evident in the standard errors.

Our initial analysis of this 3/4 fraction of the resolution VI design is quite successful. We find 14 terms with p -values less than .05, 13 of which are the same as in Table 8.8. Only **DM** is no longer significant (with p -value = .054) while two extra effects, **AB** and **CD**, have p -value near .01 (see Table 8.10). Adding these interactions would require inclusion of three more main effects to fit a hierarchical reduced model.

With 50 df still available, it is prudent to explore additional terms. For this 3/16th fraction, specifying a saturated model is not straightforward. For the 2⁹⁻² fraction (Example 8.1) just analyzed, the aliasing was simple; there were 127 alias sets of size 4, such as

$$\begin{aligned}
 \mathbf{EJM} &= \mathbf{ACK} = \mathbf{BDKLM} = \mathbf{ABCDEJL} \\
 \mathbf{EKL} &= \mathbf{BDJ} = \mathbf{ACJLM} = \mathbf{ABCDEKM} \\
 \mathbf{CDEL} &= \mathbf{ABEM} = \mathbf{BCJK} = \mathbf{ADJKLM} \\
 \mathbf{CDEJKM} &= \mathbf{AD} = \mathbf{BCLM} = \mathbf{ABEJKL}
 \end{aligned}$$

Table 8.10. Initial model for T using resolution V $(3/4)2^{9-2}$ fraction

(a) Analysis of Variance				
Source	df	SS	MS	F-Ratio
Model	45	3.6030	0.0801	7.55
Error	50	0.5302	0.0106	
Total (corrected)	95	4.1332		

(b) Parameter estimates				
Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	0.818	0.0105	77.86	<.0001
M	-0.088	0.0105	-8.36	<.0001
L	-0.082	0.0105	-7.76	<.0001
K	-0.056	0.0111	-4.98	<.0001
LM	-0.055	0.0111	-4.90	<.0001
KL	-0.055	0.0111	-4.90	<.0001
KM	-0.054	0.0111	-4.83	<.0001
JM	-0.050	0.0111	-4.45	<.0001
JK	-0.048	0.0111	-4.33	.0001
J	-0.040	0.0111	-3.55	.0009
JL	-0.037	0.0111	-3.34	.0016
D	-0.034	0.0111	-3.01	.0041
DL	-0.028	0.0105	-2.64	.0111
AB	-0.029	0.0111	-2.63	.0114
CD	-0.029	0.0111	-2.56	.0136
DM	-0.022	0.0111	-1.97	.0540
AE	-0.019	0.0111	-1.71	.0929
A	0.017	0.0111	1.53	.1334
BJ	-0.016	0.0111	-1.47	.1489
EJ	-0.014	0.0105	-1.36	.1791
BC	-0.014	0.0105	-1.32	.1925
DJ	-0.013	0.0105	-1.25	.2153
AK	0.012	0.0105	1.19	.2415
BM	0.013	0.0111	1.16	.2529
AM	-0.011	0.0105	-1.08	.2867
BL	0.010	0.0105	0.95	.3454
AL	0.009	0.0111	0.82	.4177
DE	-0.008	0.0111	-0.73	.4710
BD	-0.008	0.0111	-0.70	.4857
EM	0.007	0.0111	0.67	.5083
EL	-0.007	0.0111	-0.65	.5209
CJ	-0.006	0.0105	-0.61	.5455
DK	-0.006	0.0111	-0.56	.5748
B	0.006	0.0111	0.56	.5809

12 smallest estimates omitted, including **C** and **E**

To fit a saturated model for the 2^{9-2} , we simply retain one effect from each alias set. However, for our $(3/4)2^{9-2}$, we can only estimate 3 of the 16 terms above. That only three can be estimated is a consequence of our omitting from the 2^{9-2} fraction a 2^{9-4} fraction that aliases these four sets together. Assuming lower-order terms are more likely, we would include, for example, **EJM**, **EKL**, and **AD** in the saturated model. These 3 coefficients can be estimated, provided the model omits the remaining 13 terms listed above. Such a careful choice could be made for the other 31 groups of 4 alias sets. This would add (up to) 50 terms to the two-factor interaction model summarized in Table 8.10. For the saturated model, the $\mathbf{X}'\mathbf{X}$ matrix is block diagonal with 32 blocks of size 3, and each estimate will have a standard error of $\sigma/64^{1/2}$ rather than $\sigma/96^{1/2}$ or $\sigma/85.3^{1/2}$ as in Table 8.10. This decrease in precision is the price one pays to investigate the possibility of higher-order terms using a nonorthogonal design.

Alternatively, some software will select a saturated model automatically. JMP's Modeling Screening platform readily computes a saturated model for this design and shows some of the aliases; see Figure 8.2. However, rather than provide the actual regression estimates, JMP orthogonalizes the model matrix and provides estimates that correspond to uncorrelated linear combinations of the effects specified in the model. To see the impact of this transformation, we fit a saturated model in the terms specified by JMP; the largest estimates are given in Table 8.11. In Figure 8.2, an asterisk following a contrast estimate indicates that this contrast column is correlated with a term previously entered into the model and so has been adjusted to make it orthogonal. To understand a specific case, consider the estimates $b_{JK} = -0.048$ and $b_{LM} = -0.055$ from Table 8.10. These two estimates are correlated for this $3/16^{\text{th}}$ fraction with a correlation of $1/3$. JMP's Modeling Screening shown in Figure 8.2 handles the correlation as follows. **M** and **L** are the largest two main effects in a main-effects-only model, so their interaction is entered first; the estimate $b_{LM} = -0.0386$ is the least squares estimate for a model with no other interactions. Further down the list in Figure 8.2, the interaction for **JK** appears. Since its contrast is correlated with **LM**, **JK** is replaced by $(3\mathbf{JK} + \mathbf{LM})/8^{1/2}$, which is uncorrelated with **LM** and has sum of squares equal to $N = 96$. The estimate for β_{LM} in Figure 8.2 is biased because it is computed omitting the **JK** interaction, which is not negligible. The difference in the estimate for β_{JK} is less objectionable, being just a matter of rescaling. These details have been provided to clarify the difference between using correlated estimates and ones that have been reparameterized to be uncorrelated. This book recommends using correlated estimates.

For the saturated model, every estimate has the standard error $\sigma/64^{1/2}$. This is estimated by the PSE = 0.01371, calculated in the usual manner; the corresponding estimate for σ^2 is $64(0.01371)^2 = .0120$. Note however, that the 95 estimates of the saturated model are correlated. Edwards and Mee (2008) investigated the effect of correlated estimates on the null distribution of Lenth t statistics. For the saturated model here, nearly 98% of the correlations among

the estimates are zero; the remaining correlations are $\pm .5$, and their effect is negligible. Thus, it is satisfactory to compare the Lenth t statistics in Table 8.11 with critical values from Appendix C.

Fitting a saturated model here allows one to check for potential lack-of-fit of the lower-order model, just as we did when we had a resolution VI fraction. The largest Lenth t of the 50 higher-order terms is 2.18 (see Table 8.11). Such an outcome is entirely consistent with all these effects being inactive. In fact, only 1 of the 50 three-factor or higher-order interactions for the saturated model has a Lenth $t > 1.7$. The lower-order model is deemed acceptable.

Table 8.11. Largest estimates from saturated model for T using resolution V $(3/4)2^{9-2}$ fraction

Term	Estimate	PSE	Lenth t
Intercept	0.80974	0.01371	59.06
L	-0.09387	0.01371	-6.85
M	-0.08574	0.01371	-6.25
LM	-0.05701	0.01371	-4.16
K	-0.05549	0.01371	-4.05
J	-0.05119	0.01371	-3.73
JK	-0.05061	0.01371	-3.69
KM	-0.04594	0.01371	-3.35
KL	-0.04546	0.01371	-3.32
JM	-0.04047	0.01371	-2.95
DL	-0.03921	0.01371	-2.86
EKM	0.02984	0.01371	2.18
JL	-0.02927	0.01371	-2.14
CD	-0.02852	0.01371	-2.08
BC	-0.02785	0.01371	-2.03
MLK	-0.02325	0.01371	-1.70
MKJ	-0.02282	0.01371	-1.66
MBE	0.02269	0.01371	1.65
D	-0.02223	0.01371	-1.62
AE	-0.02174	0.01371	-1.59

We now fit two reduced models. The simpler includes only the largest 12 terms in Table 8.10; these terms form a hierarchical model with $R^2 = 77\%$. Alternatively, one might also include **AB** and **CD** plus the main effects **A**, **B**, and **C**; note that these two interactions are statistically significant both for the saturated model and the two-factor interaction model analyses. This model has an $R^2 = 81\%$, but its residual plot is worse (see Figure 8.3); 10 of 96 predicted values exceed 1, the maximum possible value for T , and the largest negative residual is not reduced. The simpler model actually seems preferred.

<u>Term</u>	<u>Contrast</u>	<u>Lenth t</u>	<u>p-Value</u>	<u>Aliases</u>
M	-0.0878	-7.21	<.0001	
L	-0.0816	-6.70	<.0001	
K	-0.0650	-5.34	<.0001	
J	-0.0493	-4.05	0.0006	
D	-0.0348	-2.86	0.0080	
A	0.0116	0.95	0.3426	
B	0.0055	0.45	0.6633	
E	-0.0029	-0.24	0.8211	
C	-0.0004	-0.03	0.9787	
M*L	-0.0386	-3.17	0.0036	D*A*B*C
M*K	-0.0415	-3.40	0.0018	J*A*E*C
L*K	-0.0381	-3.12	0.0042	J*D*B*E
M*J	-0.0468	*	-3.84	0.0008
L*J	-0.0351	*	-2.88	0.0072
K*J	-0.0455	*	-3.74	0.0009
M*D	-0.0203	-1.67	0.1012	L*A*B*C
L*D	-0.0277	-2.28	0.0251	K*J*B*E, M*A*B*C
K*D	-0.0059	*	-0.49	0.6334
J*D	-0.0132	-1.08	0.2810	L*K*B*E
M*A	-0.0113	-0.93	0.3506	L*D*B*C, K*J*E*C
L*A	0.0085	0.70	0.4817	M*D*B*C
K*A	0.0125	1.02	0.3073	M*J*E*C
J*A	-0.0021	*	-0.17	0.8728
D*A	0.0040	0.33	0.7504	M*L*B*C
M*B	0.0156	1.28	0.2022	L*D*A*C
L*B	0.0100	0.82	0.4096	K*J*D*E, M*D*A*C
K*B	-0.0036	-0.30	0.7753	L*J*D*E
J*B	-0.0154	*	-1.27	0.2080
D*B	-0.0103	-0.85	0.3971	L*K*J*E, M*L*A*C
A*B	-0.0276	*	-2.27	0.0256
M*E	0.0070	*	0.57	0.5760
L*E	-0.0083	-0.68	0.4931	K*J*D*B
K*E	-0.0021	-0.17	0.8730	L*J*D*B, M*J*A*C
J*E	-0.0143	-1.18	0.2420	L*K*D*B, M*K*A*C
D*E	-0.0076	*	-0.63	0.5406
A*E	-0.0185	-1.52	0.1332	M*K*J*C
B*E	-0.0048	*	-0.39	0.7046
M*C	0.0037	0.30	0.7706	L*D*A*B, K*J*A*E
L*C	-0.0016	*	-0.13	0.9002
K*C	-0.0034	*	-0.28	0.7867
J*C	-0.0064	-0.53	0.6085	M*K*A*E
D*C	-0.0269	*	-2.21	0.0296
A*C	0.0030	*	0.25	0.8122
B*C	-0.0139	-1.14	0.2554	M*L*D*A
E*C	0.0017	*	0.14	0.8949
M*L*K	-0.0190	*	-1.56	0.1250
K*J*A	0.0075	*	0.61	0.5503
M*D*A	-0.0247	*	-2.03	0.0447
L*D*B	-0.0084	-0.69	0.4873	K*J*E, M*A*C
M*K*E	0.0244	*	2.00	0.0480
..... (41 Insignificant Three-factor Interaction Terms Omitted)				
M*K*D*A	-0.0012	*	-0.09	0.9264
L*K*D*A	-0.0014	*	-0.12	0.9118
M*K*A*B	0.0114	*	0.94	0.3460
L*K*A*B	-0.0111	*	-0.91	0.3589

Fig. 8.2. JMP's Modeling Screening saturated model for 3/16th fraction of 2⁹

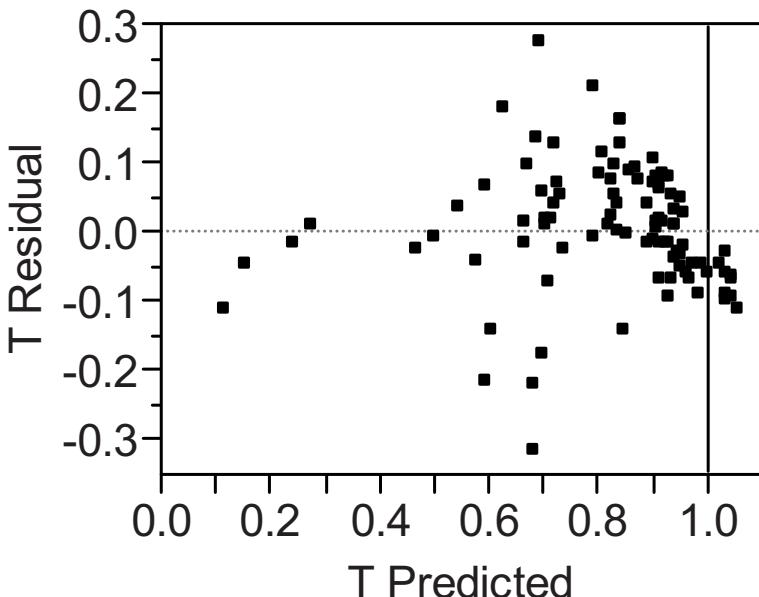


Fig. 8.3. Residual plot for 17-term model fit to 96-run irregular fraction

8.6.3 Example 8.3: Analysis of a smaller irregular fraction

The 64-run irregular resolution V design analyzed here consists of the following two regular fractions of Wang's 2^9 :

- 2^{9-4} with $\mathbf{J} = \mathbf{ACD}$, $\mathbf{L} = \mathbf{ABCE}$, $\mathbf{M} = \mathbf{CDE}$, $\mathbf{K} = +1$.
- 2^{9-4} with $\mathbf{J} = \mathbf{ABCD}$, $\mathbf{L} = \mathbf{BDE}$, $\mathbf{M} = -\mathbf{CDE}$, $\mathbf{K} = -1$.

This is identical to the nine-factor design proposed in Section 8.4.2; the only difference is that \mathbf{J} , \mathbf{L} , and \mathbf{M} are used to denote the additional factors, instead of \mathbf{F} , \mathbf{H} , and \mathbf{J} , respectively. As with the 128-run and 96-run fractions of the 2^9 , we begin our analysis of this resolution V design by fitting a model with all main effects and two-factor interactions. The resulting ANOVA and the 16 most significant parameter estimates are displayed in Table 8.12.

The following differences are noted. First, with fewer observations, there are fewer degrees of freedom for error and slightly larger standard errors. All main effects and some interactions are estimated orthogonally, but interactions with standard errors larger than $0.0127 [= \text{RMSE}/(64)^{1/2}]$ are correlated with other interactions. The largest standard errors (0.0179) are for two terms with VIFs = 2. Due to the larger standard errors, fewer estimates turn out to be statistically significant. From this model, the statistically significant estimates (using $\alpha = .05$) include the complete two-factor interaction model for factors $\mathbf{J}-\mathbf{M}$, plus \mathbf{BC} .

Table 8.12. Two-factor interaction model fit for T using resolution V irregular 64-run fraction

(a) Analysis of Variance				
Source	df	SS	MS	F-Ratio
Model	45	42500.2	944.449	4.9711
Error	18	3419.8	189.990	
Total (corrected)	63	45920.0		

(b) Largest 15 parameter estimates, sorted				
Term	Estimate	Std Error	t-Ratio	p-Value
M	-0.1055	0.0127	-8.33	.0000
L	-0.0868	0.0127	-6.86	.0000
LM	-0.0895	0.0146	-6.12	.0000
KM	-0.0762	0.0127	-6.02	.0000
K	-0.0682	0.0127	-5.39	.0000
JK	-0.0511	0.0127	-4.03	.0008
J	-0.0455	0.0127	-3.59	.0021
BC	-0.0439	0.0146	-3.01	.0076
JM	-0.0437	0.0146	-2.99	.0078
JL	-0.0287	0.0127	-2.26	.0362
KL	-0.0273	0.0127	-2.15	.0451
DL	-0.0288	0.0146	-1.97	.0642
DM	-0.0220	0.0127	-1.74	.0989
D	-0.0220	0.0127	-1.74	.0991
E	-0.0199	0.0127	-1.57	.1333
...				

One logical next step is to determine the higher-order interactions aliased with **BC**. Since the design is the combination of two regular fractions, we find the aliases of **BC** for the two halves of the design: the two alias sets are

$$\begin{aligned} \mathbf{BC} &= \mathbf{AEL} = \mathbf{JLM} = \mathbf{BCK} = \mathbf{AEKL} = \mathbf{JKLM} = \dots, \\ \mathbf{BC} &= -\mathbf{LM} = \mathbf{ADJ} = -\mathbf{BCK} = \mathbf{KLM} = -\mathbf{ADJK} = \dots. \end{aligned}$$

The partial aliasing with **LM** will not bias the estimate for β_{BC} , since the **LM** interaction is in the model. Furthermore, there is no aliasing with **BCK** due to the reverse of sign. It is the partial aliasing with three-factor interactions **JLM** and **KLM** that is the most relevant, since all lower-order terms involving these are active. Given the three-factor interaction model for **J–M**, adding **BC** interaction explains no additional variation.

This nonregular fraction has good projection properties. For example, if one ignores **A–C** (the factors with main effects not appearing in Table 8.12),

the design forms a 2^6 in the remaining factors. Using Lenth's method for this orthogonal projection, we obtain $PSE = 0.0143$; the largest estimates appear in Table 8.13. Note that the largest 13 estimates form a hierarchical model, which we would adopt as our final model.

Table 8.13. Largest estimates for full factorial model in an orthogonal projection of resolution V irregular 64-run fraction

Term	Estimate	PSE	Lenth t
M	-0.1055	0.0143	-7.35
L	-0.0868	0.0143	-6.05
KM	-0.0762	0.0143	-5.31
K	-0.0682	0.0143	-4.75
LM	-0.0675	0.0143	-4.71
JM	-0.0526	0.0143	-3.66
JK	-0.0511	0.0143	-3.56
J	-0.0455	0.0143	-3.17
KLM	-0.0346	0.0143	-2.41
JLM	-0.0311	0.0143	-2.17
JKM	-0.0297	0.0143	-2.07
JL	-0.0287	0.0143	-2.00
KL	-0.0273	0.0143	-1.90
DL	-0.0241	0.0143	-1.68
DM	-0.0220	0.0143	-1.54
D	-0.0220	0.0143	-1.54
E	-0.0199	0.0143	-1.39
JKL	-0.0198	0.0143	-1.38

8.6.4 Example 8.4: Analysis of data from Rechtschaffner's saturated design for seven factors

Le Thanh, Voilley, and Luu (1993) investigated how seven factors influence the volatility of three different aromatic food additives. The response in each case was the measured vapor–liquid equilibrium coefficient. The factors and their levels are reported in Table 8.14 and the treatment combinations are listed in Table 8.15.

Table 8.14. Factors and levels for volatility experiment

Factors	Levels	
	-1	1
A Glucose (g/kg)	20	100
B $(\text{NH}_4)_2\text{SO}_4$ (g/kg)	2	100
C KH_2PO_4 (g/kg)	1	100
D Acid (g/kg)	2	50
E Melange (g/kg)	1	10
F Temperature ($^{\circ}\text{C}$)	25	60
G pH	4	7

Table 8.15. Volatility experiment using Rechtschaffner's 29-run resolution V design for $k = 7$ factors

A	B	C	D	E	F	G	K_{VL}	$\log_{10}(K_{VL})$
-1	-1	-1	-1	-1	-1	-1	0.09	-1.046
-1	-1	-1	-1	-1	1	1	0.90	-0.046
-1	-1	-1	-1	1	-1	1	0.11	-0.959
-1	-1	-1	-1	1	1	-1	0.38	-0.420
-1	-1	-1	1	-1	-1	1	0.06	-1.222
-1	-1	-1	1	-1	1	-1	0.55	-0.260
-1	-1	-1	1	1	-1	-1	0.07	-1.155
-1	-1	1	-1	-1	-1	1	0.26	-0.585
-1	-1	1	-1	-1	1	-1	1.25	0.097
-1	-1	1	-1	1	-1	-1	0.16	-0.796
-1	-1	1	1	-1	-1	-1	0.09	-1.046
-1	1	-1	-1	-1	-1	1	0.23	-0.638
-1	1	-1	-1	-1	1	-1	2.10	0.322
-1	1	-1	-1	1	-1	-1	0.25	-0.602
-1	1	-1	1	-1	-1	-1	0.11	-0.959
-1	1	1	-1	-1	-1	-1	0.35	-0.456
1	-1	-1	-1	-1	-1	1	0.10	-1.000
1	-1	-1	-1	-1	1	-1	0.84	-0.076
1	-1	-1	-1	1	-1	-1	0.10	-1.000
1	-1	-1	1	-1	-1	-1	0.06	-1.222
1	-1	1	-1	-1	-1	-1	0.15	-0.824
1	1	-1	-1	-1	-1	-1	0.30	-0.523
-1	1	1	1	1	1	1	2.10	0.322
1	-1	1	1	1	1	1	1.06	0.025
1	1	-1	1	1	1	1	0.90	-0.046
1	1	1	-1	1	1	1	2.43	0.386
1	1	1	1	-1	1	1	1.86	0.270
1	1	1	1	1	-1	1	0.31	-0.509
1	1	1	1	1	1	-1	0.86	-0.066

The substance 2,5 dimethylpyrazine has a low vapor–liquid equilibrium coefficient of about $K_{VL} = 0.1$ at 25°C in water. Over the 29 treatment combinations in the experiment, measured coefficients ranged from 0.06 to 2.43 (see Table 8.15). Since the responses are severely skewed, we follow the authors and fit models for $\log_{10}(K_{VL})$.

We begin our analysis by fitting a saturated model for $\log_{10}(K_{VL})$. For Rechtschaffner designs such as this, the regression coefficients are correlated but estimated with equal precision. To facilitate the use of Lenth’s method for estimating the standard error of the estimates, Table 8.16 lists the estimated coefficients from largest to smallest in magnitude.

Table 8.16. Sorted estimates from a saturated model for $\log_{10}(K_{VL})$

Term	Estimate	PSE	t-Ratio
F	0.4095	0.018	22.75
B	0.1548	0.018	8.60
C	0.1210	0.018	6.72
D	-0.1007	0.018	-5.59
G	0.0722	0.018	4.01
EF	-0.0721	0.018	-4.00
CG	0.0503	0.018	2.79
BD	-0.0345	0.018	-1.92
DE	0.0309	0.018	1.72
BC	-0.0227	0.018	-1.26
AC	-0.0215	0.018	-1.19
E	-0.0207	0.018	-1.15
EG	0.0198	0.018	1.10
DF	0.0190	0.018	1.05
BG	-0.0146	0.018	-0.81
A	-0.0127	0.018	-0.70
FG	0.0123	0.018	0.69
AG	-0.0118	0.018	-0.65
AF	-0.0072	0.018	-0.40
AD	-0.0071	0.018	-0.40
CE	0.0068	0.018	0.38
CF	-0.0057	0.018	-0.32
DG	0.0049	0.018	0.27
CD	-0.0048	0.018	-0.26
BE	0.0037	0.018	0.20
AE	-0.0026	0.018	-0.14
AB	0.0022	0.018	0.12
BF	-0.0010	0.018	-0.05

Edwards and Mee (2008) examined this same example and verified that the small correlations among all the estimates have little effect on the null distribution of Lenth t statistics. Since these estimates have equal precision, the calculation of the PSE is straightforward using Table 8.16. Excluding the largest six estimates since they exceed $2.5[1.5(0.0168)] = 0.063$, Lenth's PSE = $1.5(0.012) = 0.018$, where 0.012 is the median size of the smallest 22 estimates. Table 8.15 uses this estimated standard error to construct t -ratios. Five of the seven main effects and two interactions have large t -ratios. We fit a reduced model with two interactions, including the main effect for **E**, so that the model is hierarchical.

Table 8.17 contains the results of fitting a reduced model with all seven main effects and the two significant interactions. The purpose of including the insignificant factor **A** is to make explicit that this factor (Glucose) has no apparent effect on volatility. Note that this reduced model explains 98.2% of the variability in $\log_{10}(K_{VL})$. For this reduced model, the estimates have changed slightly and their standard errors are slightly smaller and no longer equal. This is because with fewer terms, the correlations among the estimators are reduced.

Table 8.17. Reduced model fit to $\log_{10}(K_{VL})$

(a) Analysis of Variance

Source	df	SS	MS	F-Ratio
Model	9	7.2879	0.8098	118.267
Error	19	0.1301	0.0068	
Total (corrected)	28	7.4179		

(b) Parameter estimates

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	-0.3776	0.0171	-22.14	<.0001
A	-0.0032	0.0165	-0.19	.8488
B	0.1706	0.0165	10.33	<.0001
C	0.1297	0.0167	7.77	<.0001
D	-0.1100	0.0165	-6.66	<.0001
E	-0.0400	0.0167	-2.40	.0270
F	0.4040	0.0167	24.21	<.0001
G	0.0614	0.0167	3.68	.0016
EF	-0.0517	0.0170	-3.04	.0067
CG	0.0481	0.0170	2.83	.0107

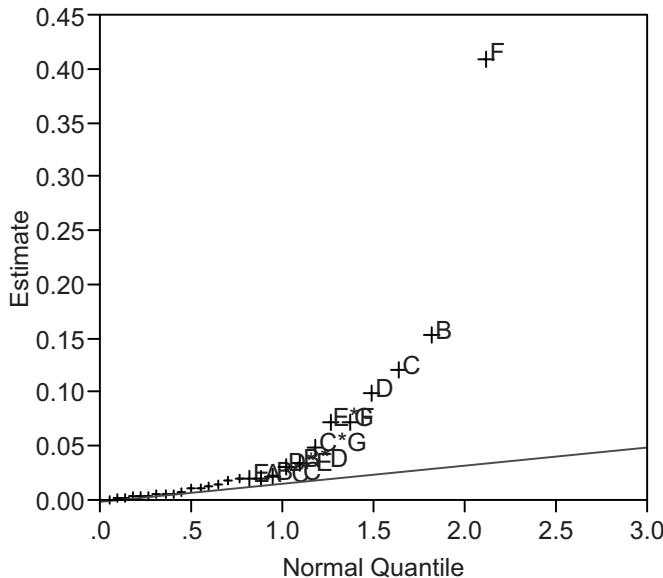


Fig. 8.4. Half-normal plot of t -ratios for original (correlated) estimates

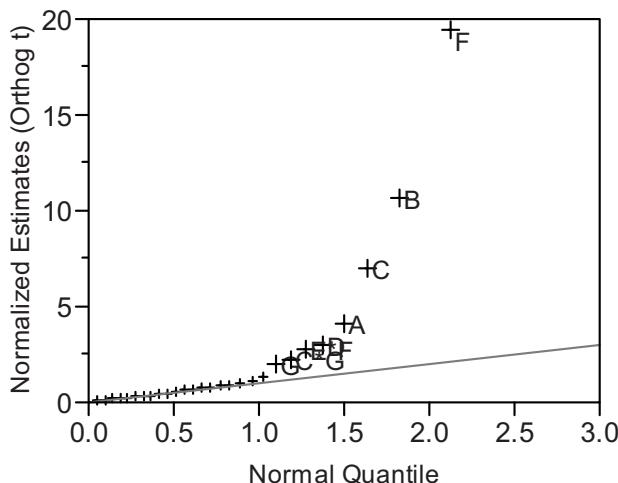


Fig. 8.5. Half-normal plot of t -ratios for orthogonal (but biased) estimates

We conclude our analysis with a reminder that when analyzing nonorthogonal designs, it is best to conduct tests using the (correlated) least squares

estimates for the models of interest, rather than uncorrelated linear combinations of the specified effects. Figures 8.4 and 8.5 illustrate the difference. Figure 8.4 is a half-normal plot of the correlated estimates from Table 8.16. Provided the two-factor interaction model is correct, these estimates are unbiased. However, software may also offer plots of uncorrelated linear combinations of these estimates; see Figure 8.5. Since the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix has no zero elements, the uncorrelated estimates here will be biased by every effect entered later in the model. Thus, with **A** listed first, its estimate plotted in Figure 8.5 is biased by every other active effect. Whereas the two-factor interaction model showed no apparent effect from **A**, Figure 8.5 assigns to **A** the fourth largest estimate. Thus, due to the potential for misinterpreting output for estimates transformed to be uncorrelated, avoid their use.

Augmenting Fractional Factorial Designs

Fractional factorial designs should often be followed by additional experimentation, either with another designed experiment or with confirmation runs. This chapter surveys the options and then discusses several particularly common situations. The sections are as follows:

- Section 9.1. Follow-up Experimentation Choices
- Section 9.2. Confirmation Runs
- Section 9.3. Steepest Ascent Search
- Section 9.4. Foldover After a Resolution III Fraction
- Section 9.5. Foldover and Semifolding After a Resolution IV Fraction
- Section 9.6. Optimal Design Augmentation
- Section 9.7. Dropping and Adding Factors

9.1 Follow-up Experimentation Choices

Follow-up experimentation is particularly important for fractional factorial designs, since estimation of a model from a fractional factorial design necessarily involves the assumption that certain factorial effects can be ignored. Follow-up experimentation may be as simple as performing confirmation runs at a single treatment combination or it may involve a new experiment larger than the initial design. Although it is impossible to describe all of the possibilities, here are some questions that follow-up experimentation may answer.

1. Does the fitted model accurately represent the response at one or more specific locations of interest within or near the initial design region?

This is perhaps the most common question following an initial experiment. Section 9.2 addresses the use and interpretation of

confirmation runs to check a fitted model or to assess the validity of outliers from the initial fraction.

2. Does the fitted model provide a useful guide for exploration into a new region of experimentation yielding better results?

Question 2 assumes that a response is not optimized in the initial region of experimentation and uses a fitted model to direct exploration outside that initial region; Section 9.3 discusses such situations.

3. How can we improve our estimates for main effects and a few interactions following an initial screening experiment?

Question 3 concerns improving estimates both in terms of precision and in terms of reducing aliasing (i.e., reducing the potential for bias). Section 9.4 will illustrate how adding a second 2^{k-f} design to the initial fraction can increase the resolution of a design, increase the number of estimable effects, and decrease the standard errors by 30%.

4. How can we estimate interactions that are aliased together in an initial experiment?

Question 4 arises following an initial experiment from Chapter 7, where two-factor interactions are aliased. Depending on the situation, one may add another experiment half the size of the original design, or even smaller (see Section 9.5).

5. What augmentation is best if we want to estimate three or more effects that are all aliased together?

Question 5 presents a situation where foldover designs are inadequate. D-Optimal design augmentation provides a practical and efficient general purpose solution. This is discussed in Section 9.6.

6. How can we repair a fractional factorial experiment that has missing observations?

D-Optimal augmentation may also be used to repair fractional factorial designs missing observations. So Section 9.6 is also pertinent here.

7. How can one add another experiment, dropping some factors from consideration while possibly adding others?

Question 7 is also common for situations with many potential factors of interest. Following one experiment where only a subset of the factors proves influential, subsequent experimentation may simply fix the levels of the apparently inactive factors at a single value; alternatively, we may actually widen the levels for such factors if we believe the initial range was too narrow. Just as some factors that were varied may be fixed in subsequent experimentation, variables that were held constant in an initial experiment can be varied as factors in the follow-up design. Section 9.7 offers advice on how to proceed.

8. When centerpoint runs provide a better response than at the observed factorial treatment combinations, how can one augment the design to fit a model accounting for such curvature?
9. When both main effects and interactions are prominent, how can one expand the experimental region in a direction of interest?

Questions 8 and 9 require that we move beyond the simplicity of two-level experiments. Chapter 12 provides an introduction to this topic.

9.2 Confirmation Runs

We suppose that a fractional factorial design has been conducted and a suitable model obtained to explain the observed data. This fitted model can be used to predict the response and a standard error calculated to determine a confidence interval for the true mean response. For instance, consider Example 6.1, where we fit an additive model for Resolution and Migration rate. Our fractional factorial represents just one-fourth of the treatment combinations in the full 2^5 . By saving the fitted models as formulas, we may obtain the predicted values for Resolution and Migration rate for any treatment combination. Figure 9.1 plots the pairs of predicted response values at all 32 treatment combinations of the full factorial. In the plot, asterisks denote treatment combinations in our fractional factorial, and square points denote treatment combinations where we have no data yet. Since larger values are preferred for both responses, the treatment combinations along the lower band of points are all inferior. These correspond to high acetonitrile ($C = 1$). Along the upper band, we prefer to compromise between high resolution/low rate and low resolution/high rate. One point of interest with predicted resolution = 1.37 and migration rate = 0.08/min is identified by an arrow in the plot; its treatment combination is $(A, B, C, D, E) = (-1, -1, -1, +1, -1)$.

If we conducted a follow-up run at this treatment combination, how far might the observed result differ from its predicted responses? To answer this question, we begin by computing the standard error for the estimated mean value as well as a standard error for the prediction error between a single observed value and this prediction. For orthogonal designs where each regression coefficient has variance σ^2/N , the variance of \hat{y} at any treatment combination in the 2^k factorial is

$$\text{Var}(\hat{y}) = \text{Var}(b_0) + r\sigma^2/N, \quad (9.1)$$

where r is the number of regression coefficients in addition to the intercept. In (9.1), for a design without blocking, $\text{Var}(b_0) = \sigma^2/(N + n_0)$, where n_0 is the number of centerpoint runs. Thus, for our additive models with five factors, $N = 2^{5-2}$ and $n_0 = 0$, $\text{Var}(\hat{y}) = 6\sigma^2/8$. The mean square error for Resolution is 0.001812, so the standard error for the predicted mean is

$$[6(0.001812)/8]^{1/2} = 0.0369$$

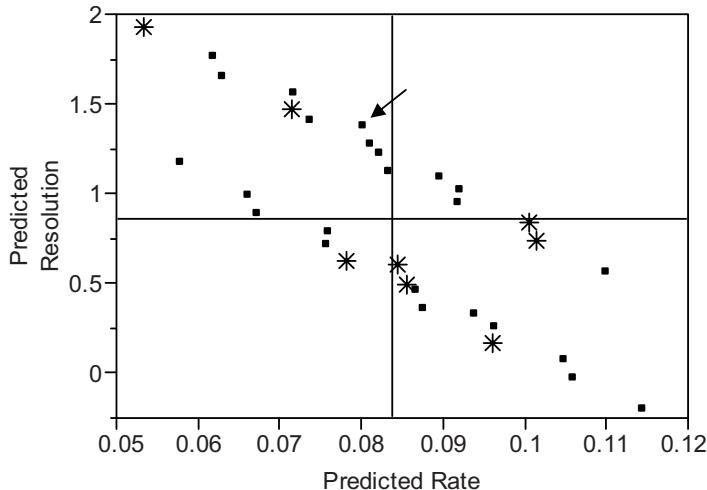


Fig. 9.1. Predicted Resolution and Migration rate for 32 treatment combinations of the 2^5 for Example 6.1

and a 95% confidence interval for the true mean resolution at the treatment combination $(-1, -1, -1, +1, -1)$ is

$$\hat{y} \pm t_{0.025, \text{df}} s_{\hat{y}} \rightarrow 1.37 \pm 4.303(0.0369). \quad (9.2)$$

Because (9.2) is a confidence interval for the mean response, the outcome of an individual confirmation run is subject to more error than for which (9.2) is intended. For an individual observation, we construct a prediction interval that is based on the $\text{Var}(y - \hat{y})$ rather than $\text{Var}(\hat{y})$. This larger variance is

$$\text{Var}(y - \hat{y}) = \sigma^2 + \text{Var}(b_0) + r\sigma^2/n. \quad (9.3)$$

For resolution at the treatment combination $(-1, -1, -1, +1, -1)$, a 95% prediction interval is

$$\hat{y} \pm t_{0.025, \text{df}} s_{y-\hat{y}} \rightarrow 1.37 \pm 4.303(0.0563), \quad (9.4)$$

since $[(1+6/8)0.001812]^{1/2} = 0.0563$. Thus, according to our model, we expect an observed resolution of 1.37 ± 0.24 at this treatment combination.

The confidence interval for the mean, (9.2), and the prediction interval for y , (9.4), ignore the possibility of any block effect—that is, of any changes between the initial experiment and the follow-up runs that would cause a shift in the response. If a block effect is likely, then confirmation runs at two or more treatment combinations will be needed. Select treatment combinations for which the comparison is of interest, such as the former operating conditions, and a new recommended treatment combination.

In addition to using confirmation runs at treatment combinations of interest to verify the adequacy of a model, confirmation runs may be used to verify outliers. If some observations in the initial experiment are not well explained by the chosen model, then collecting additional data at these treatment combinations will confirm whether they are reproducible or not.

Finally, confirmation runs following “experiments” from computer models are simpler to interpret, since there is obviously no need for a block effect. For deterministic computer models, there is no random error, so any discrepancy between the model’s prediction and the confirmation observation(s) represents lack-of-fit.

9.3 Steepest Ascent Search

When factors are continuous, one is not constrained to only consider treatment combinations with coded levels of ± 1 . We continue with Example 6.1, searching for new treatment combinations with attractive compromises for Migration rate and Resolution using the method of steepest ascent. In particular, we illustrate three variants of steepest ascent: optimizing a single response, simultaneously improving two responses, and optimizing a single response subject to a constraint. In each case, we use the regression coefficients to define a search direction where factors with large coefficients are changed more rapidly than are factors with smaller coefficients. Finally, we describe the complications that arise when extrapolating based on a model with interaction effects.

9.3.1 Steepest ascent for a single response

The observed Migration times ranged from less than 10 min to nearly 20 min (see Figure 6.1). It is desirable to have small Migration time so that the analysis time is brief. A fitted first-order model for Migration rate, the reciprocal of Migration time, is shown in Table 9.1, together with t statistics and p -values based on the t -distribution with 2 df. Using only the statistically significant estimates, we see that increased Migration rate is associated with increasing **B** and lowering **A**, **D**, and **E**.

Table 9.1. Main effects model for Migration rate

Term	Estimate	Std Error	t -Ratio	p -Value
Intercept	0.0839	0.00094	89.31	.0001
A	-0.0091	0.00094	-9.70	.0105
B	0.0048	0.00094	5.06	.0369
C	0.0022	0.00094	2.31	.1469
D	-0.0102	0.00094	-10.86	.0084
E	-0.0042	0.00094	-4.50	.0459

Steepest ascent, as proposed by Box and Wilson (1951), is the method of choosing a direction for extrapolation that maximizes the predicted response for a given “step size.” Using coded units, the distance from the center of the design to the point (\mathbf{A} , \mathbf{B} , \mathbf{C} , \mathbf{D} , \mathbf{E}) is

$$\text{Distance} = (\mathbf{A}^2 + \mathbf{B}^2 + \mathbf{C}^2 + \mathbf{D}^2 + \mathbf{E}^2)^{1/2}.$$

For any orthogonal two-level design, predicted values at treatment combinations the same distance from the center all have the same standard error for the main-effects-only model (1.2), and the standard error is an increasing function of the distance. Thus, maximizing the predicted response among all steps of a given size is equivalent to maximizing the predicted response among all points with the same standard error. The direction that maximizes the predicted response is the vector of main effect estimates. Minimizing the predicted response is accomplished by going in the opposite direction.

Here we propose to move in the direction $(-0.0091, 0.0048, 0, -0.0102, -0.0042)$, leaving \mathbf{C} unchanged because its coefficient is not statistically significant for Migration rate, but it is clearly significant and negative for the Resolution response; that is, with $b_C = 0.0022$, we are not sure that increasing \mathbf{C} will increase Migration rate, but it certainly will decrease Resolution. Dividing the recommended direction by the magnitude of the largest coefficient, we obtain the following rescaled vector with a suitable step size:

$$\Delta = \begin{pmatrix} -0.0091 \\ 0.0048 \\ 0 \\ -0.0102 \\ -0.0042 \end{pmatrix} / 0.0102 = \begin{pmatrix} -0.89 \\ 0.47 \\ 0.00 \\ -1.00 \\ -0.41 \end{pmatrix}.$$

δ -Multiples of the vector Δ produce coded treatment combinations along a ray emanating from the design center. These coded treatment combinations must be converted to natural units. For the natural units here (see Table 6.1), the treatment combination $\delta\Delta$ is

$$\begin{pmatrix} 8.5 \\ 5 \\ 45 \\ 45 \\ 30 \end{pmatrix} + \delta \begin{pmatrix} -.89(0.5) \\ .47(5) \\ .00(5) \\ -1.00(5) \\ -.41(10) \end{pmatrix} = \begin{pmatrix} 8.5 - .445\delta \\ 5 + 2.35\delta \\ 45 \\ 45 - 5\delta \\ 30 - 4.1\delta \end{pmatrix}.$$

For $\delta = 1$ and 2, the steepest ascent points in natural units are $(8.05, 7.33, 45, 40, 25.86)$ and $(7.61, 9.66, 45, 35, 21.71)$, respectively. The predicted response is $b_0 + \delta\Delta'\mathbf{b}$, which equals

$$\begin{aligned} & 0.0839 + \delta[0.89(0.0091) + 0.47(0.0048) + 0 + 1(0.0102) + 0.41(0.0042)] \\ &= 0.0839 + 0.0223\delta. \end{aligned}$$

For $\delta = 1$ (2), the predicted response is 0.106 (0.128), both higher than the best Migration rate in the 2^{5-2} .

These calculations should be done in advance of collecting data, using a spreadsheet; that is, for a series of treatment combinations along the path of steepest ascent, one should compute the coded and natural units, as well as the predicted response at each location. Calculating the natural units is necessary to ensure that each treatment combination is feasible. Calculating the predicted value is useful for determining if the fitted model is still valid, by comparison with observed values as one collects data along the path. If the standard deviation is small compared to the improvement seen as one explores out the path of steepest ascent, then a single point at each location will be sufficient to determine whether the response is still improving, as the model predicts. That is the case here, with RMSE = 0.00266, one-eighth of the expected improvement for each step of 1Δ . Once extrapolation begins to produce minimal gains, it may be necessary to replicate steepest ascent runs to determine where extrapolation ceases to be productive. Alternatively, one may collect some data beyond the perceived optimum and then use the steepest ascent data to fit a polynomial model as a function of distance; a suitably chosen smooth function is better than individual observations at discrete points for determining the optimum along a path. Finally, to aid in interpreting agreement of the new data with the previous model, it is useful to compute prediction intervals. Statistical software generally furnishes both confidence intervals for the mean and prediction intervals for a single observation, as discussed in Section 9.2. For our example, the standard error for the mean response at the point 1Δ is 0.0017, whereas the standard error relevant for an individual response is $(0.0017^2 + \text{MSE})^{1/2} = 0.0031$. The larger standard error reflects the average error between an individual observation and the predicted response, assuming the additive model is valid. For a 90% prediction interval, we take $t_{2,.05} = 2.92$ from Appendix A, since the MSE is based on just 2 df. The prediction interval is $0.106 \pm 2.92(0.0031)$, or 0.106 ± 0.009 ; thus, an observed Migration rate within 0.009 of the predicted rate (at $\delta = 1$) would be consistent with our model.

The direction of steepest ascent (descent) is very sensitive to the original design's scaling. If an active factor's effect is essentially linear over a wide range, then the wider the choice of levels for that factor in the two-level factorial design, the larger the regression coefficient will be for the coded model and the larger will be the multiplier when converting back to natural units. Having one b_i of much larger magnitude than the rest means that the search direction is not too different than just a one factor search. An experiment such as this with many statistically significant b_i 's of similar magnitude is indicative of an experiment with appropriately chosen factors and levels.

9.3.2 Steepest ascent for two responses

Suppose we have two responses that we wish to improve simultaneously. For instance, suppose both are larger-the-better responses, so that we are interested in steepest ascent. We need only consider non-negative linear combinations of the two paths of steepest ascent (as illustrated below), since these provide the optimal compromise directions (Mee and Xiao 2008b). Consider now attempts to improve both Resolution and Migration rate above what is achieved at the design center. For our first response, Resolution, we compute

$$\Delta_1 = \begin{pmatrix} 0.19 \\ -0.14 \\ -0.38 \\ 0.27 \\ 0.09 \end{pmatrix} / 0.38 = \begin{pmatrix} 0.50 \\ -0.37 \\ -1.00 \\ 0.69 \\ 0.23 \end{pmatrix},$$

using coefficients from the fitted model for Resolution. Earlier, we calculated Δ for Migration rate, zeroing out the nonsignificant coefficient for \mathbf{C} . Here we retain $b_C = 0.0022$ in the calculation of Δ_2 , to avoid biasing the required trade-off;

$$\Delta_2 = \begin{pmatrix} -0.0091 \\ 0.0048 \\ 0.0022 \\ -0.0102 \\ -0.0042 \end{pmatrix} / 0.0102 = \begin{pmatrix} -0.89 \\ 0.47 \\ 0.21 \\ -1.00 \\ -0.41 \end{pmatrix}.$$

Any direction $\Delta = \delta_1\Delta_1 + \delta_2\Delta_2$, with $\delta_i > 0$ ($i = 1, 2$), is a positive linear combination of the two paths of steepest ascent. Since $\Delta_1'\Delta_2 < 0$, following either path of steepest ascent will reduce the other predicted response. Thus, we consider compromises. If one chooses $\delta_1 = \delta_2 = 1$, the compromise direction is

$$\Delta = \Delta_1 + \Delta_2 = \begin{pmatrix} 0.50 \\ -0.37 \\ -1.00 \\ 0.69 \\ 0.23 \end{pmatrix} + \begin{pmatrix} -0.89 \\ 0.47 \\ 0.21 \\ -1.00 \\ -0.41 \end{pmatrix} = \begin{pmatrix} -0.39 \\ 0.10 \\ -0.79 \\ -0.31 \\ -0.18 \end{pmatrix}$$

At the design center, the predicted (Resolution, Migration rate) is (0.859, 0.084). For each additional step Δ from the center, predicted (Resolution, Migration rate) increase by (0.116, 0.006). The predicted responses at Δ and 2Δ are (0.975, 0.090) and (1.091, 0.096), respectively. The later point requires extrapolation only in \mathbf{C} ; its distance from the center

$$[(-0.78)^2 + (0.20)^2 + (-1.58)^2 + (-0.62)^2 + (-0.36)^2]^{1/2} = 3.66^{1/2}$$

is smaller than $5^{1/2}$, the distance of the factorial treatment combinations from the center. Data could be collected at the treatment combination 2Δ . If results are promising, collect data at additional points further in this direction.

One is not restricted to weighting Δ_1 and Δ_2 equally. If a trade-off of more improvement in Resolution is desired—with the consequence of less improvement in Rate than is indicated by the above direction—then recalculate the compromise direction $\Delta = \delta_1\Delta_1 + \delta_2\Delta_2$ using coefficients with $\delta_1 > \delta_2$. For instance, for $\delta_1 = 1.2$ and $\delta_2 = 1$, the direction is $(-0.29, 0.026, -0.99, -0.172, -0.134)$ and the predicted increase for each additional step from the center is $(0.264, 0.003)$. A step of distance $5^{1/2}$ in this direction is $(-0.62, 0.055, -2.10, -0.36, -0.28)$, where predicted Resolution = 1.417 and predicted Migration rate = 0.090, which is superior to any outcome at a factorial point. In Figure 9.2 we augment Figure 9.1 by adding predicted values for non-negative linear combinations of Δ_1 and Δ_2 , scaled to a common step size of $5^{1/2}$. These points are “Pareto optimal”; that is, there does not exist any treatment combinations with coded distance from the center $\leq 5^{1/2}$ which yields both a higher predicted Resolution and higher predicted Migration rate than any point in this set. The two arrows in Figure 9.3 correspond to the Pareto optimal points for $\delta_1/\delta_2 = 1$ and $\delta_1/\delta_2 = 1.2$ discussed above. Which weighting is preferred depends on the relative importance of improving the response variables.

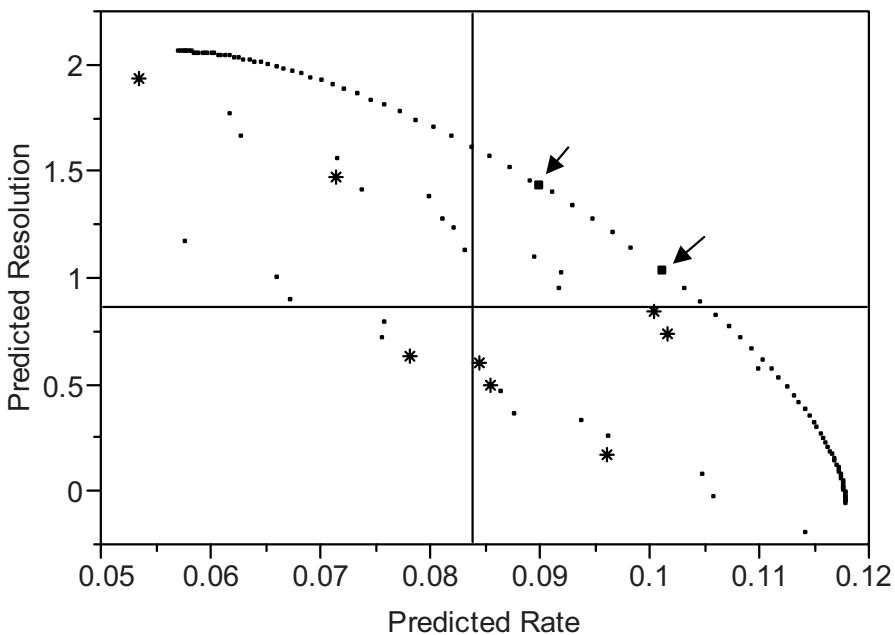


Fig. 9.2. Figure 9.1 augmented with predicted Resolution and predicted Migration rate for non-negative linear combinations of paths of steepest ascent at coded distance $5^{1/2}$ from the design center; arrows point to outcome for two different δ_1/δ_2 ratios discussed in text.

The correlation between predicted Resolution and predicted Migration rate at the design points is -0.7824 . This correlation equals the cosine of the angle between the two steepest ascent vectors. If the correlation is positive, the angle between them is less than 90° and every positive linear combination increases both predicted responses. When the inner product is negative, as it is for our example, the set of directions that improve both responses is more restricted. For our example, the angle is $\arccos(-0.7824) = 141.5^\circ$. If the correlation were nearer to -1 , the directions would be more nearly opposite and there would be little or no opportunity to improve both responses. Such a situation is mentioned below. When an experiment yields a strong negative correlation between two larger-the-better responses, it may be necessary to experiment again with additional factors, since these provide the possibility of finding another factor which, like **C**, influences one response more than the other.

9.3.3 Steepest ascent for a single response, with additional constraints

The previous subsection explained how to choose a direction for a follow-up experiment devoted to improving two responses simultaneously. When the two steepest ascent vectors are diametrically opposed, the objective may become instead to optimize one response subject to a minimally sufficient value for the second. This situation also arises when one has responses for which a target value is specified, such as maximizing yield, subject to achieving the specified target viscosity.

Petersson, Lundell, and Markides (1992) experimented with three factors influencing the retention times and the resolution for separating two components of a chiral compound. Similar to Vindevogel and Sandra (1991), trade-offs are required to achieve adequate resolution without making the time excessive. Petersson et al. conducted a series of experiments seeking to minimize time, subject to achieving adequate resolution. Their sequence of experiments is discussed and reanalyzed as a case study in Section 11.2. Here we just briefly summarize their approach. Their first experiment allowed them to identify a treatment combination where the minimally acceptable resolution of 0.274 was achieved. The second experiment was centered at this treatment combination. Using models estimated from the second experiment, they determined a starting point on the minimally acceptable contour for Resolution and then a steepest ascent direction from this point that was constrained to that contour. Collecting data along this path did achieve some decrease in time, but with a slight drop in Resolution. They persisted with additional experimentation until a suitable treatment combination was found.

Mee and Xiao (2008b) showed how to restrict the path of steepest ascent to a contour for a second response. That result will be illustrated in Chapter 11 using the case study. For more details about multiresponse optimization from a first-order model, including the use of Derringer and Suich's (1980) desirability function, see Mee and Xiao (2008b).

9.3.4 Steepest ascent for a model with curvature

In the previous example, main effects models were used to determine the exploration direction. Once a direction was determined, exploration consisted of different step sizes in the same direction. Under these simple models, standard errors for the predicted response depend only on distance from the center, not on direction, and the expected improvement is proportional to the step size. Here we consider how the presence of interactions affects the search to increase or decrease a response.

Recall the example from Section 5.1, the 2^{5-1} experiment by Hu and Bai (2001). Adding a main effect for \mathbf{C} to the reduced model adopted by Hu and Bai to make the model hierarchical, phosphorus content is estimated by

$$\widehat{\%P} = 4.86 + 0.97\mathbf{A} + 1.53\mathbf{B} - 0.23\mathbf{C} + 0.99\mathbf{AB} - 2.67\mathbf{BC}. \quad (9.5)$$

Clearly, higher phosphorus content can be achieved by increasing \mathbf{B} , provided $\mathbf{C} = -1$. Furthermore, this model indicates that when $\mathbf{B} > 0$, increasing \mathbf{A} increases $\%P$. This observation is consistent with the fact that the best two responses from the 2^{5-1} experiment both had $\mathbf{A} = \mathbf{B} = 1$ and $\mathbf{C} = -1$. Thus, extrapolating in this general direction is recommended. However, what direction corresponds to the path of steepest ascent? We choose to follow a path with the steepest gradient.

The gradient for the model (9.5) at a treatment combination $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ is

$$\begin{aligned}\nabla \widehat{\%P} &= [\partial(\widehat{\%P})/\partial\mathbf{A}, \partial(\widehat{\%P})/\partial\mathbf{B}, \partial(\widehat{\%P})/\partial\mathbf{C}] \\ &= (0.97 + 0.99\mathbf{B}, 1.53 + 0.99\mathbf{A} - 2.67\mathbf{C}, -0.23 - 2.67\mathbf{B}).\end{aligned}$$

At the center $(0, 0, 0)$, the gradient is $(0.97, 1.53, -0.23)$. Suppose we take a step in this direction of size 1:

$$\text{Step}_1 = (0.97, 1.53, -0.23)/[0.97^2 + 1.53^2 + (-0.23)^2]^{1/2} = (0.53, 0.84, -0.13).$$

From this point, the gradient is

$$(0.97 + 0.99(0.84), 1.53 + 0.99(0.53) - 2.67(-0.13), -0.23 - 2.67(0.84)),$$

which equals $(1.80, 2.4, -2.47)$. Note that the surface is steeper at Step_1 than it was at the center, since each partial derivative in the gradient has increased in magnitude. Rescaling the gradient we obtain

$$\text{Step}_2 = (1.8, 2.4, -2.47)/[(1.8)^2 + (2.4)^2 + (-2.47)^2]^{1/2} = (0.46, 0.62, -0.64).$$

Taking Step_2 after Step_1 results in the treatment combination

$$\text{Step}_1 + \text{Step}_2 = (0.53 + 0.46, 0.84 + 0.62, -0.13 - 0.64) = (0.99, 1.46, -0.77).$$

By continuing in this manner, we define a sequence of steps of equal size, at each point stepping in the direction of the greatest improvement. Note that

the direction changes at each step. For instance, for Step₁, we changed **C** the least, whereas for Step₂, we changed **C** the most.

A similar path may be obtained by using the Lagrangian method to maximize (or minimize) the predicted response, subject to the constraint of a constant distance from the center. The resulting set of points define a smooth curve. See Box and Draper (2007, Ch. 7). Points along this path optimize the response for any given step size from the center. Note, however, that for models with interactions, predicted responses at points a constant distance from the center do not have the same standard error.

Hu and Bai (2001) did something simpler than follow the gradient. Instead, they chose to fix **C** at its low level (i.e., pH = 1) and then to extrapolate beyond the initial range for **A** (Temperature) and **B** (Current density). Assuming values of pH below 1 were not permissible, this simplification is reasonable. In addition to fixing pH at its low level, factors **D** and **E** were set at the high level, and factors **A** and **B** were increased from (0, 0) to (2.3, 3.5) in seven steps of size (0.3, 0.5) in a steepest ascent search. It is important to realize that many directions will lead to improvement. There is no single correct direction for extrapolation. The important idea is simply to utilize the fitted model to identify treatment combinations with better outcomes than is possible in the previous region of experimentation.

9.4 Foldover After a Resolution III Fraction

Because regular resolution III fractional factorial designs completely alias some main effects with two-factor interactions, a common follow-up to a resolution III fraction is to add another fraction, so that together the two fractions form a resolution IV design; that is, by conducting another 2^{k-f} fraction using the same factors and the same levels, the combined experiments now form a $2/2^f = 1/2^{f-1}$ fraction of the 2^k . In the context of design augmentation, *foldover* is a term often used to describe adding another fraction that differs from the original by the reversal of levels for one or more factors.

To illustrate the main ideas of this section, consider the 2^{7-4} fraction with generators **D** = **AB**, **E** = **AC**, **F** = **BC**, and **G** = **ABC**. Its treatment combinations are

Initial 2^{7-4} fraction

t.c.	A	B	C	D	E	F	G
1	-1	-1	-1	1	1	1	-1
2	1	-1	-1	-1	-1	1	1
3	-1	1	-1	-1	1	-1	1
4	1	1	-1	1	-1	-1	-1
5	-1	-1	1	1	-1	-1	1
6	1	-1	1	-1	1	-1	-1
7	-1	1	1	-1	-1	1	-1
8	1	1	1	1	1	1	1

This fraction has seven length-3 words in its defining relation, **ABD**, **ACE**, **AFG**, **BCF**, **BEG**, **CDG**, and **DEF**. Seven length-4 words and the length-7 word **ABCDEFG** complete the defining relation. The length-3 words alias three two-factor interactions with each main effect, making this a risky design for estimating main effects.

The eight treatment combinations above are 1 of 16 fractions from this same family. The other 15 fractions have the same four generators except with some signs reversed. For resolution III fractions, one of the remaining fractions is the mirror image of the first, which may be obtained by reversing every column of the design. Thus, the eight treatment combinations of the mirror image fraction are

Mirror image 2^{7-4} fraction

t.c.	A	B	C	D	E	F	G
9	1	1	1	-1	-1	-1	1
10	-1	1	1	1	1	-1	-1
11	1	-1	1	1	-1	1	-1
12	-1	-1	1	-1	1	1	1
13	1	1	-1	-1	1	1	-1
14	-1	1	-1	1	-1	1	1
15	1	-1	-1	1	1	-1	1
16	-1	-1	-1	-1	-1	-1	-1

The generators for the mirror image fraction are **D** = **-AB**, **E** = **-AC**, **F** = **-BC**, and **G** = **ABC**. Hence, this mirror-image fraction can be obtained by reversing just the columns for **D**, **E**, and **F**. In general, the mirror-image fraction is obtained by reversing all k columns or, equivalently, by reversing just the columns with even-length interactions as generators.

Augmenting an initial resolution III design with its mirror-image fraction has the following benefits:

- All main effect estimates become clear of aliasing with two-factor interactions.

- Combinations of two-factor interactions formerly aliased with main effects are now estimable, assuming no higher-order interactions of even length.
- The precision of all estimates is improved. Assuming the error variance σ^2 is unchanged, the standard error for coefficients will decrease by a factor of $1/2^{1/2}$ (i.e., about 30%, through the addition of the second fraction).

The strategy of reversing all columns can also be applied to orthogonal arrays of strength 2 (Li, Lin, and Ye 2003) and to nonorthogonal designs (Webb 1968), with similar benefits.

One implicit assumption when a foldover design is performed is that the effects are stable from the first fraction to the second. If not, the instability of the main effect would be incorrectly attributed to the interactions aliased with that main effect in the initial fraction. For instance, if our estimates from the two fractions are

$$b_A + b_{BD} + b_{CE} + b_{FG} = 100,$$

$$b_A - b_{BD} - b_{CE} - b_{FG} = 40,$$

then the estimate for β_A is $(100 + 40)/2 = 70$ and the estimate for $\beta_{BD} + \beta_{CE} + \beta_{FG}$ is $(100 - 40)/2 = 30$. Thus, any instability in the effect of factor **A** over time between the first and second experiment is likely to be interpreted as an active two-factor interaction.

For most resolution III designs, the mirror-image fraction is the only one that will reverse the signs of all length-3 words and so increase the resolution. However, for the 2^{9-5} minimum aberration design and many other designs with k slightly larger than $N/2$, there exist better follow-up alternatives than the mirror-image fraction. Li and Mee (2002) presented several examples and offered a simple procedure for determining whether other fractions will increase the resolution. We illustrate the ideas for the 2^{9-5} case. From Appendix G, the minimum aberration design for 9 factors in 16 runs may be obtained using columns 7 and 11–14 as generators. With letters $\{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$ denoting the basic columns $\{1, 2, 4, 8\}$, the generators are

$$\begin{aligned}\mathbf{E} &= \mathbf{ABC} && \text{since } 7 = 1 \cdot 2 \cdot 4, \\ \mathbf{F} &= \mathbf{ABD} && \text{since } 11 = 1 \cdot 2 \cdot 8, \\ \mathbf{G} &= \mathbf{CD} && \text{since } 12 = 4 \cdot 8, \\ \mathbf{H} &= \mathbf{ACD} && \text{since } 13 = 1 \cdot 4 \cdot 8, \\ \mathbf{J} &= \mathbf{BCD} && \text{since } 14 = 2 \cdot 4 \cdot 8.\end{aligned}$$

This design has four length-3 words in the defining relation:

$$\mathbf{I} = \mathbf{CDG} = \mathbf{AGH} = \mathbf{BGJ} = \mathbf{EFG} = \dots$$

and produces the following aliasing among main effects and two-factor interactions:

$$\begin{aligned}
\mathbf{A} &= \mathbf{GH} \\
\mathbf{B} &= \mathbf{GJ} \\
\mathbf{C} &= \mathbf{DG} \\
\mathbf{D} &= \mathbf{CG} \\
\mathbf{E} &= \mathbf{FG} \\
\mathbf{F} &= \mathbf{EG} \\
\mathbf{G} &= \mathbf{AH} = \mathbf{BJ} = \mathbf{CD} = \mathbf{EF} \\
\mathbf{H} &= \mathbf{AG} \\
\mathbf{J} &= \mathbf{BG} \\
\mathbf{AB} &= \mathbf{CE} = \mathbf{DF} = \mathbf{HJ} \\
\mathbf{AC} &= \mathbf{BE} = \mathbf{DH} = \mathbf{FJ} \\
\mathbf{AD} &= \mathbf{BF} = \mathbf{CH} = \mathbf{EJ} \\
\mathbf{AE} &= \mathbf{BC} = \mathbf{DJ} = \mathbf{FH} \\
\mathbf{AF} &= \mathbf{BD} = \mathbf{CJ} = \mathbf{EH} \\
\mathbf{AJ} &= \mathbf{BH} = \mathbf{CF} = \mathbf{DE}
\end{aligned}$$

The mirror-image fraction may be obtained by simply reversing **G**, because **CD** is the only even-length interaction used as a generator. Since **G** appears in each of the length-3 words, adding a fraction with only **G** reversed does increase the resolution. However, one may also eliminate the four length-3 words by reversing **E**, **F**, and **G**; this has the advantage of increasing the resolution **and** eliminating 8 of the 14 length-4 words, so that the aliasing among two factor interactions is reduced to

$$\begin{array}{ll}
\mathbf{AH} = \mathbf{BJ} = \mathbf{CD} = \mathbf{EF} & \\
\mathbf{AB} = \mathbf{HJ} & \mathbf{CE} = \mathbf{DF} \\
\mathbf{AC} = \mathbf{DH} & \mathbf{BE} = \mathbf{FJ} \\
\mathbf{AD} = \mathbf{CH} & \mathbf{BF} = \mathbf{EJ} \\
\mathbf{BC} = \mathbf{DJ} & \mathbf{AE} = \mathbf{FH} \\
\mathbf{BD} = \mathbf{CJ} & \mathbf{AF} = \mathbf{EH} \\
\mathbf{AJ} = \mathbf{BH} & \mathbf{CF} = \mathbf{DE}
\end{array}$$

Li and Mee (2002) listed additional resolution III designs for which fractions other than the mirror-image fraction will increase the resolution. However, this possibility exists only when $k < (5/8)N$.

If an initial resolution III fractional factorial design does not make clear which factors are important, adding a foldover fraction is often recommended. Recall Example 6.3, a 2^{13-9} experiment, where none of the 13 main effects were statistically significant for Yield, but four estimates (**E**, **G**, **H**, **N**) were somewhat larger than the rest (see Figure 6.3). Aliasing among main effects and two-factor interactions is extensive; each main effect is aliased with five or six two-factor interactions. One of the 22 three-letter words in the defining relation involves three of the four likely factors (**I** = **-EGH**). If there are only four important main effects or interactions here, it might be **GH** that is active rather than **E**, or **EH** that is active rather than **G**, etc. Reversing all 13 main effect columns produces the mirror-image fraction. Adding

these additional 16 runs would eliminate all 22 length-3 words from the defining relation, clearing main effects from aliasing with two-factor interactions and improving the precision of estimates by doubling the number of factorial treatment combinations. If the four large estimates for Yield do correspond to active effects, we expect their estimates to be similar in the new fraction and the resulting t statistics based on the combined 2^{13-8} to increase by about 40%, since $2^{1/2} = 1.414$.

As a final example of follow-up for main effect designs, recall Example 6.6, the 12-run thermostat experiment by Bullington et al (1993). Due largely to the choice of levels, 1 of the 11 factors stood out as active, overwhelming every other effect in size. In this situation, two possible follow-up fractions should be considered. One is the mirror-image fraction already discussed. Adding this fraction to the original OA(12, 2^{11} , 2) would double the run size and increase the orthogonal array's strength to 3. Miller and Sitter (2001) discussed this specific case and proposed methods for identifying interaction effects. Li, Lin, and Ye (2003) made the point that for strength-2 orthogonal arrays with $N = 12$ and 20, the only foldover that will increase the strength is obtained by reversing all the columns.

However, the Bullington et al. data presented a situation where an alternative follow-up fraction seems advisable. Suppose one runs another OA(12, 2^{11} , 2) obtained by simply reversing the sign of the active factor \mathbf{E} . This would produce a 24-run design still of strength 2, but where at both $\mathbf{E} = 1$ and $\mathbf{E} = -1$ one has an OA(12, 2^{10} , 2) in the remaining factors. This design would permit the estimation of the 10 two-factor interactions $\mathbf{AE}, \mathbf{BE}, \dots, \mathbf{EL}$. Since \mathbf{E} 's effect is so dominant, it is quite reasonable that the effect of other active factors would depend on the level of \mathbf{E} .

For regular fractions, reversing just one column has the effect of eliminating from the defining relation every word containing this particular factor. Such a strategy is more common for follow-up to resolution IV designs, which is the topic for Section 9.5. The analysis of a mirror-image foldover experiment for a 20-run orthogonal array involving 15 factors is presented as a case study in Chapter 11.

9.5 Foldover and Semifolding After a Resolution IV Fraction

Unless three-factor interactions are a major concern, regular resolution IV fractional factorial designs typically produce at most only a few statistically significant estimates corresponding to strings of aliased effects considered of interest. For instance, in Example 7.2, the only ambiguity concerned a single pair of two-factor interactions, whereas for the more perplexing Example 7.3, four estimates pertaining to aliased interactions were significant. In neither situation would adding another 2^{6-2} fraction be justified. Unless more precision is needed, rarely does one consider adding a second fraction the same size

as the first for resolution IV designs. Instead, adding a second fraction half the size of the first generally suffices in most cases, which we now illustrate by the continuation of Example 7.3.

9.5.1 Semifolding example

Recall the experiment from Barnett, Czitrom, John, and León (1997) analyzed in Section 7.1.3. From their 2^{6-2} fraction, four of the seven estimates corresponding to (combinations of) two-factor interactions were statistically significant, whereas main effects showed little importance. In Chapter 7, we proceeded by assuming that two of the six factors could be ignored, because a two-factor interaction model in the remaining factors explained 98% of the variation. In actuality, the experimenters conducted the eight-run follow-up design that appears in Table 9.2. Since this design is half of a foldover fraction, it is referred to as a semifolding follow-up design.

Table 9.2. Semifolding follow-up design for etching uniformity by Barnett et al. (1997), sorted by predicted values obtained from the Figure 7.4 model

A	B	C	D	E	F	$\ln(\widehat{SD})$	SD	$\ln(SD)$
-1	1	1	1	-1	1	0.41	6.63	1.89
1	-1	1	1	-1	1	0.53	9.81	2.28
1	-1	1	-1	1	1	1.16	6.19	1.82
-1	1	1	-1	1	1	1.22	6.60	1.89
1	1	-1	-1	-1	1	1.74	8.59	2.15
1	1	-1	1	1	1	2.05	9.23	2.22
-1	-1	-1	-1	-1	1	3.03	12.96	2.56
-1	-1	-1	1	1	1	3.52	13.67	2.62

The eight treatment combinations in Table 9.2 represent half of a foldover fraction of the original 2^{6-2} (Table 7.8). The original design's defining relation is

$$I = \mathbf{ABDE} = \mathbf{ABCF} = \mathbf{CDEF}. \quad (9.6)$$

If we reverse column **F** (or **C**), the defining relation for the new fraction is

$$I = \mathbf{ABDE} = -\mathbf{ABCF} = -\mathbf{CDEF}. \quad (9.7)$$

The treatment combinations in Table 9.2 are those that satisfy (9.7) and also have **F** = 1. Predicted $\ln(SD)$ values for these treatment combinations are included in Table 9.2. These were taken from (the rear face of each cube in) Figure 7.4. The actual results of the follow-up fraction are quite disappointing. The first six actual $\ln(SD)$ values in Table 9.2 are larger than what our earlier

model predicted they would be, and none are close to what we expected for the optimum. Second, the correlation between observed and predicted $\ln(\text{SD})$ is only .8, far below the correlation between observed and predicted values from the initial experiment. Figure 9.3 displays the near-perfect agreement between the observed and predicted $\ln(\text{SD})$ values from the initial 18-run experiment as well as the much weaker correlation between predicted values and the follow-up experiment values. The follow-up experiment shows surprisingly small differences among the eight new standard deviations.

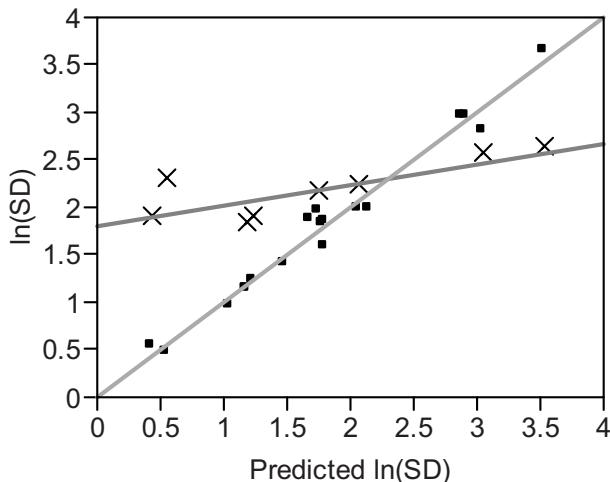


Fig. 9.3. Observed $\ln(\text{SD})$ versus predicted $\ln(\text{SD})$; initial experiment (●), follow-up experiment (✕)

Even though the results of the follow-up experiment did not match our expectation, we proceed to fit a model to the combined data. Fitting a two-factor interaction model (plus a block main effect) to the 26 runs, the only aliasing is the result of the word **ABDE**, which appears in both (9.6) and (9.7). By comparing the estimates before and after the additional eight runs, one sees that the main effect estimates are unchanged. For the five alias chains that were split in two, the initial estimate is split into two estimates based on the information in the semifolding experiment (see Table 9.3).

To present clearly how the estimates in Table 9.3 combine the information from the initial fractional factorial and the follow-up semifold fraction, we list in Table 9.4 the estimates obtained by fitting the two-factor interaction model to only the semifold data. Combining $b_A = 0.103$ and $b_{AF+BC} = -0.440$ from the original fraction with $b_{A+AF-BC} = -0.060$ from the semifolding fraction, we have three equations in three unknowns. The solution for the interaction estimates is

Table 9.3. Estimates for two-factor interaction model for Example 7.4, before and after the semifolding follow-up design

Before Semifolding		After Semifolding		
Term	Estimate	Term	Estimate	Std Error
Intercept	1.791	Intercept	2.026	0.0349
		Block	-0.235	0.0349
A	0.103	A	0.103	0.0354
B	-0.058	B	-0.058	0.0354
C	-0.380	C	-0.380	0.0354
D	-0.019	D	-0.019	0.0354
E	0.111	E	0.111	0.0354
F	-0.082	F	-0.082	0.0354
AB=CF=DE	-0.497	AB=DE	-0.334	0.0354
		CF	-0.162	0.0354
AC=BF	0.353	AC	0.218	0.0354
		BF	0.135	0.0354
AD=BE	0.079	AD=BE	0.073	0.0289
AE=BD	-0.044	AE=BD	-0.047	0.0289
AF=BC	-0.440	AF	-0.301	0.0354
		BC	-0.139	0.0354
CD=EF	0.168	CD	0.161	0.0354
		EF	0.007	0.0354
CE=DF	0.079	CE	-0.007	0.0354
		DF	0.086	0.0354

Table 9.4. Estimates for two-factor interaction model using only the semifolding data from Table 9.2

Term	Estimate
Intercept = F	2.179
A = AF = -BC	-0.060
B = -AC = BF	-0.141
C = -AB = CF = -DE	-0.208
D = -CE = DF	0.074
E = -CD = EF	-0.042
AD = BE	0.059
AE = BD	-0.055

$$b_{AF} = (-0.440 - 0.060 - 0.103)/2 = -0.301,$$

$$b_{BC} = (-0.440 + 0.060 + 0.103)/2 = -0.139.$$

Only the aliased interactions $\mathbf{AD} = \mathbf{BE}$ and $\mathbf{AE} = \mathbf{BD}$ are estimable in both the original and follow-up fraction. Thus, the estimates for these effects in Table 9.3 combine the estimates from the separate fractions, weighting by the relative sample sizes for the factorial designs:

$$b_{AD+BE} = (2/3)0.079 + (1/3)0.059 = 0.073,$$

$$b_{AE+BD} = (2/3)(-0.044) + (1/3)(-0.055) = -0.047.$$

It is reassuring that these estimates from the initial and follow-up fractions were so consistent, as that gives some support for our assumptions that higher-order terms can be ignored and that there are no Block*Factor interactions. These are the only two estimates for which the additional runs improved the precision. All the other main effect and interaction estimates have standard errors equal to $\text{RMSE}/16^{1/2}$.

The simplifying conjecture made in Section 7.1.3 that **B** and **D** have no effect is contradicted by the follow-up design. That conjecture supposed the statistically significant interaction estimates found in the first fraction were each due to a single interaction. Instead, after the follow-up design, it appears that **AF** and **BC** are both active, **AC** and **BF** are both active, **CF** and either **AB** or **DE** are active, as well as **CD**. Thus, all the factors appear in statistically significant interactions. At this point, we consider the initial choice of fraction very fortunate, since in each case, aliased interactions had like sign, making them easier to detect, rather than having them be opposite sign, making the combination negligible. If the sign of interaction effects can be guessed in advance of the design, one can choose a fraction purposely to avoid cancellation of estimates. Ambiguity remains, however, about the largest interaction estimate, for which **AB** and **DE** are still aliased.

If additional runs could be performed, it would be desirable to collect data at treatment combinations for which **ABDE** = −1. Since the best outcomes are consistently for **C** = +1, we recommend that its level be fixed. Low rotation speed never seems to be preferred, so we decide to fix **F** = 1 as well. An eight-run 2^{4-1} experiment in factors **A**, **B**, **D**, and **E** is one reasonable option. We compare this with another possibility in Section 9.6.

Semifold follow-up designs are based on two choices. The first choice is which columns to reverse. The second choice is which half of the new fraction to run. One key advantage of follow-up experiments is that these choices can be made in light of the results of the initial experiment. The next two subsections provide insight regarding these two choices.

9.5.2 Which columns to reverse?

For the resolution IV 2^{6-2} and 2^{7-3} designs, each of the remaining fractions can be obtained by reversing a single factor. However, for the 2^{8-4} , being a $1/16^{\text{th}}$ fraction, only eight of the other fractions are obtained by reversing a single factor; the remaining seven fractions are found by reversing a pair of

factors. Montgomery and Runger (1996) noted that for the 2^{8-4} design, it is often preferable to reverse two columns rather than one, since reversing any single factor eliminates half of the 14 length-4 words, whereas reversing any pair of factors eliminates 8 length-4 words. Li and Lin (2003) designated as “optimal foldover plans” choices that minimize the aberration of the combined design. For the 32-run resolution IV designs in Table G.3, the optimal foldover plans are given in Table 9.5. Reversing one factor is always suitable for designs with alias chains of length 2 or 3. However, for designs with alias chains of length 4 or more, one must reverse more columns to minimize the number of length-4 words, A_4 , in the combined design. The ideal foldover in terms of minimizing both A_4 and the length of the longest alias chain is to break each alias chain in half. For the 2^{15-10} and 2^{16-11} designs, this ideal is achieved, but it requires reversing five and six factors, respectively, to achieve this. In these cases, the optimal foldover eliminates more than twice as many four-letter words as is gained by reversing a single factor. For more details, see Li and Lin (2003), who showed the foldover that minimizes A_4 for dozens of designs of possible interest with $k \leq 11$.

Table 9.5. Optimal foldover plans for 32-run, resolution IV designs with minimum aberration

Initial Design	Initial A_4	Min. A_4 After Foldover	No. of Factors to Reverse
2^{7-2}	1	0	1
2^{8-3}	3	1	1
2^{9-4}	6	2	2
2^{10-5}	10	4	2
2^{11-6}	25	10	3
2^{12-7}	38	16	3
2^{13-8}	55	23	4
2^{14-9}	77	33	4
2^{15-10}	105	45	5
2^{16-11}	140	60	6

Mee and Xiao (2008a) presented more details regarding foldovers to 32-run even resolution IV designs. The results in Table 9.5 for 11–16 factors are published there, along with details about which specific columns achieve this best foldover. Ai, Xu, and Wu (2008) report minimum aberration foldovers of initial blocked designs of size $N = 16$ ($k \leq 12$), 32 ($k \leq 21$), and 64 ($k \leq 19$); that is, both the initial fraction and the foldover fraction are divided into 2, 4, or 8 blocks each.

Mee and Xiao (2008a) discussed the following example in detail. Jones, Marrs, Young, and Townend (1995) utilized a 32-run (2^{15-10}) fractional factorial with generators $\mathbf{F} = \mathbf{BCE}$, $\mathbf{G} = \mathbf{ACE}$, $\mathbf{H} = \mathbf{ABC}$, $\mathbf{I} = \mathbf{BDE}$, \mathbf{J}

$= \mathbf{ADE}$, $\mathbf{K} = \mathbf{ABD}$, $\mathbf{L} = \mathbf{ABCDE}$, $\mathbf{M} = \mathbf{ACD}$, $\mathbf{N} = \mathbf{BCD}$, and $\mathbf{O} = \mathbf{CDE}$ to investigate the forming process for ceramic composite furnace linings. Using Lenth's PSE, the initial fraction showed conclusively only three main effects. Six of the next seven largest estimates were two-factor interactions. Two follow-up designs seem worthy of consideration. First, since the three large main effects explain 67% of the variation, one might use steepest ascent to extrapolate in these factors, seeking further improvement in mass. Alternatively, if extrapolation is not feasible, one might consider a full foldover fraction, in order to split the 15 sets of aliased interactions into 30 smaller sets, as well as to improve the precision of the estimates. By adding another 32 runs, one would have 15 df for main effects, 30 df for two-factor interactions, 1 df for blocks, and the rest for error (assuming no higher-order interactions). For the 2^{15-10} , any foldover reversing one (two) column eliminates 28 (44) length-4 words. However, by reversing five columns, as recommended in Table 9.5, one may eliminate 60 length-4 words, and split every alias chain of length 7 into chains of length 3 and 4. Since we have several prominent main effects, such a foldover fraction is much preferred to one that makes the 14 interactions for 1 factor clear.

9.5.3 Choosing which half of the foldover

When the initial fraction provides adequate precision, running just half of a foldover fraction, as in Table 9.2, is a useful strategy for augmenting resolution IV designs. For Table 9.2, the foldover fraction was split using factor \mathbf{F} , with the choice to omit the treatment combinations with $\mathbf{F} = -1$ and to run only those with $\mathbf{F} = 1$. Mee and Peralta (2000) refer to this subset of the foldover fraction as “ss = $\mathbf{F}+$.” For even resolution IV designs, one may “subset” on any one of the factors, and the resulting semifold fraction provides the same estimable two-factor interactions as if the full foldover were performed (see Mee and Xiao 2008a). Recall from Section 7.2.2 that all resolution IV designs with $k > (5/16)N$ are even. These are the designs for which semifolding is most appropriate. Generally, the subset of choice corresponds to the preferred level of a prominent factor.

9.6 Optimal Design Augmentation

Adding a foldover fraction to an initial fractional factorial design cannot separate a chain of aliased effects into more than two pieces. Therefore, adding a foldover fraction or even a semifold fraction seems ill-suited for the following situation that appeared originally in Box, Hunter, and Hunter (1978, p. 402), with four aliased interactions that require estimates. The strategy presented here is based on the D-optimal design criterion, which chooses augmenting treatment combinations that maximize the determinant of the information matrix $\mathbf{X}'\mathbf{X}$, where \mathbf{X} is the model matrix for a user-specified model. For a

detailed account of optimal design, see Atkinson and Donev (1992) or Atkinson, Donev and Tobias (2007).

Eight factors for an injection molding process were varied according to a 2^{8-4} design to determine their effect on some shrinkage measurement. Following Meyer, Steinberg, and Box (1996), we denote the factors as **A–H**, corresponding to Mold temperature, Moisture content, Holding pressure, Cavity thickness, Booster pressure, Cycle time, Gate size, and Screw speed, respectively. The design and shrinkage measurements appear in Table 9.6.

Table 9.6. Injection molding experiment from Box, Hunter and Hunter (1978)

A	B	C	D	E	F	G	H	<i>y</i>
-1	-1	-1	1	1	1	-1	1	14.0
1	-1	-1	-1	-1	1	1	1	16.8
-1	1	-1	-1	1	-1	1	1	15.0
1	1	-1	1	-1	-1	-1	1	15.4
-1	-1	1	1	-1	-1	1	1	27.6
1	-1	1	-1	1	-1	-1	1	24.0
-1	1	1	-1	-1	1	-1	1	27.4
1	1	1	1	1	1	1	1	22.6
-1	-1	-1	-1	-1	-1	-1	-1	22.3
1	-1	-1	1	1	-1	1	-1	17.1
-1	1	-1	1	-1	1	1	-1	21.5
1	1	-1	-1	1	1	-1	-1	17.5
-1	-1	1	-1	1	1	1	-1	15.9
1	-1	1	1	-1	1	-1	-1	21.9
-1	1	1	1	1	-1	-1	-1	16.7
1	1	1	-1	-1	-1	1	-1	20.3

Fitting a saturated model and sorting the estimates, we have overwhelming evidence for two main effects and a linear combination of four interactions (see Table 9.7). We conclude that “+1” Holding pressure and “-1” Booster pressure produce high shrinkage. (Since experience suggests that raising holding pressure should decrease shrinkage and the levels for the factors were not reported, perhaps the lower holding pressure was actually assigned to **C** = 1.)

Table 9.7. Injection molding experiment estimates with Lenth t statistics

Term	Estimate	Lenth t
C	2.75	7.33
AE=BF=CH=DG	2.30	6.13
E	-1.90	-5.07
H	0.60	1.60
AC=BG=DF=EH	0.45	1.20
A	-0.35	-0.93
AB=CG=DH=EF	-0.30	-0.80
AH=BD=CE=FG	-0.30	-0.80
G	0.30	0.80
AD=BH=CF=EG	-0.20	-0.53
AF=BE=CD=GH	-0.15	-0.40
D	-0.15	-0.40
AG=BC=DE=FH	-0.10	-0.27
B	-0.05	-0.13
F	-0.05	-0.13

The highly significant estimate for **AE** and its aliases prompts the need for follow-up runs. Effect heredity assumptions would suggest that the large estimate is likely due to **CH** or **AE** rather than **BF** or **DG**. However, rather than assume the later two interactions are inactive, a four-run follow-up experiment was conducted for which the **CH**, **BF**, and **DG** contrasts are orthogonal to one another and with **AE** = +1. The four new treatment combinations and response values are shown in Table 9.8.

Table 9.8. Follow-up injection molding experiment from Box, Hunter and Hunter (1978, p. 414)

A	B	C	D	E	F	G	H	y
-1	1	1	1	-1	-1	-1	1	29.4
-1	1	-1	-1	-1	1	1	1	19.7
1	1	-1	-1	1	-1	-1	1	13.6
1	1	1	1	1	1	1	1	24.7

This follow-up design provides three additional df for interactions, plus a block effect. Fitting a model containing **AE** and its three aliases, the estimate for **CH** stands out as the only active interaction among the four. The fitted model is given in Table 9.9, where the standard errors are based on the MSE = 1.2133 with 6 df (from omitting all other two-factor interactions).

Table 9.9. Reduced model after follow-up runs

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	19.75	0.28	71.72	0.000
Block	1.50	1.40	1.07	0.326
A	-0.35	0.28	-1.27	0.251
B	-0.05	0.28	-0.18	0.862
C	2.75	0.28	9.99	0.000
D	-0.15	0.28	-0.54	0.606
E	-1.90	0.28	-6.90	0.000
F	-0.05	0.28	-0.18	0.862
G	0.30	0.28	1.09	0.318
H	0.60	0.28	2.18	0.072
AE	0.05	1.20	0.04	0.968
BF	0.10	0.67	0.15	0.887
DG	-0.45	0.67	-0.67	0.529
CH	2.60	0.67	3.85	0.008

Although this follow-up design was apparently successful at isolating one interaction, it provides very poor precision for **AE**, with standard error $\hat{\sigma}/(0.84)^{1/2} = 1.20$. A better follow-up design could have been obtained by requesting the additional four runs that would maximize the determinant of $\mathbf{X}'\mathbf{X}$ for the model in Table 9.9. The results are not unique, but one such D-optimal augmentation is in Table 9.10. The four predicted values for each new treatment combination differ by ± 2.3 based on which interaction is included in the model. If these treatment combinations would have been performed as the four-run follow-up, the variances for the four interactions would have ranged from $\sigma^2/(3.636)$ to $\sigma^2/(5.3)$, roughly half what they were based on Box, Hunter, and Hunter's follow-up runs.

Table 9.10. D-Optimal follow-up injection molding experiment

\hat{y} from Model with:											
A	B	C	D	E	F	G	H	b_{AE}	b_{BF}	b_{DG}	b_{CH}
1	1	-1	-1	1	1	1	1	18.0	18.0	13.4	13.4
-1	-1	-1	-1	-1	1	-1	1	22.0	17.4	22.0	17.4
1	1	-1	1	-1	1	1	1	16.9	21.5	21.5	16.9
1	1	-1	-1	-1	-1	1	-1	16.1	16.1	16.1	20.7

Mitchell (1974) provided an early example of D-optimal augmentation of fractional factorial designs. His Example 4 involves augmentation of the resolution IV 2^{7-3} with 14 extra runs (the minimum number) in order to estimate all two-factor interactions. Optimal design algorithms do require that

we specify a particular model of interest. Although this is generally reasonable for design augmentation, alternative approaches have been proposed that do not completely specify the model of interest. Instead, they utilize a Bayesian perspective and various effect heredity assumptions. The Bayesian approach advocated by Meyer, Steinberg, and Box (1996) selected follow-up designs intended to identify which factors are active, as distinct from identifying which terms are active. The discussion by Chipman and Hamada (1996) highlights this distinction; see also Jones and DuMouchel (1996).

The role of effect heredity assumptions must not be overlooked for these Bayesian approaches. For the injection molding example just considered, the negligible estimates for **B**, **D**, **F**, and **G** essentially exclude **BF** and **DG** from further consideration, if the Bayesian prior assumes that the only likely interactions are those involving at least one active main effect. By contrast, the D-optimal follow-up design in Table 9.10 entertained the possibility of any of these four interactions.

Section 2.10 addressed the problem of analyzing data when a few treatment combinations are missing from an intended full factorial design. Missing observations from a fractional factorial are somewhat more problematic. D-Optimal augmentation provides a means of simultaneously repairing such experiments and resolving aliasing among potential effects. Although not specific to two-level designs, Hebble and Mitchell (1972) discussed D-optimal augmentation to both repair and extend a design.

9.7 Adding and Dropping Factors

Foldover designs involve follow-up experiments using the same factors and levels as an initial experiment. Semifolding as described in Section 9.5 involved a follow-up experiment in which one factor is dropped, in that its level is fixed. Gilmour and Mead (1996) discussed fixing one or more factors at a preferred level in subsequent stages of experimentation. The same ideas apply in reverse—that is, when a variable that was fixed in the initial experiment is chosen as a factor to be varied in a subsequent experiment. When this is done, the fixed level in the initial experiment should be one of the two levels for the new factor in the follow-up experiment. Alternatively, for continuous factors, the new levels may be centered around the original fixed level.

The idea behind Gilmour and Mead's (1996) work is that the experiments are conducted separately in time (i.e., in blocks) but are to be analyzed in combination. Dropping seemingly irrelevant factors and including new factors is surely a common practice. For instance, Lai, Pan, and Tzeng (2003), discussed previously as Example 6.2, dropped Ammonium sulfate and added Glycerol in their second experiment (see Table 6.6). However, they did not think to analyze the combined data. As we learned from Table 6.9, analyzing the combined data can be quite enlightening.

Fractional Factorial Designs with Randomization Restrictions

Run order restrictions were discussed for full factorial designs in Chapter 3. Here we consider how to conduct fractional factorials as randomized block or split-unit experiments. The sections are as follows:

- Section 10.1. Randomized Block Designs for Fractional Factorials
- Section 10.2. Split-Unit Designs for Fractional Factorials
- Section 10.3. Analysis of Four Fractional Factorial Experiments with Randomization Restrictions
- Section 10.4. Sequences of Fractional Factorial Designs

10.1 Randomized Block Designs for Fractional Factorials

Regular fractional factorial design of resolution III or higher may be conducted as a randomized block design rather than as a completely randomized design. To conduct the experiment in two blocks, one simply chooses one factorial effect contrast that is not aliased with any lower-order effects and uses it to split the design into two blocks. For instance, consider a 2^{6-2} example with generators $\mathbf{E} = \mathbf{ABC}$ and $\mathbf{F} = \mathbf{BCD}$. The defining relation and 15 contrasts are

$$\begin{aligned}\mathbf{I} &= \mathbf{ABCE} = \mathbf{BCDF} = \mathbf{ADEF} \\ \mathbf{B} &= \mathbf{ACE} = \mathbf{CDF} = \mathbf{ABDEF} \\ \mathbf{D} &= \mathbf{ABCDE} = \mathbf{BCF} = \mathbf{AEF} \\ \mathbf{F} &= \mathbf{ABCDF} = \mathbf{BCD} = \mathbf{ADE} \\ \mathbf{AC} &= \mathbf{BE} = \mathbf{ABDF} = \mathbf{CDEF} \\ \mathbf{AE} &= \mathbf{BC} = \mathbf{ABCDEF} = \mathbf{DF} \\ \mathbf{BD} &= \mathbf{ACDE} = \mathbf{CF} = \mathbf{ABEF} \\ \mathbf{ABD} &= \mathbf{CDE} = \mathbf{ACF} = \mathbf{BEF}\end{aligned}$$

$$\begin{aligned}\mathbf{A} &= \mathbf{BCE} = \mathbf{ABCDF} = \mathbf{DEF} \\ \mathbf{C} &= \mathbf{ABE} = \mathbf{BDF} = \mathbf{ACDEF} \\ \mathbf{E} &= \mathbf{ABC} = \mathbf{BCDEF} = \mathbf{ADF} \\ \mathbf{AB} &= \mathbf{CE} = \mathbf{ACDF} = \mathbf{BDEF} \\ \mathbf{AD} &= \mathbf{BCDE} = \mathbf{ABCF} = \mathbf{EF} \\ \mathbf{AF} &= \mathbf{BCEF} = \mathbf{ABCD} = \mathbf{DE} \\ \mathbf{BF} &= \mathbf{ACEF} = \mathbf{CD} = \mathbf{ABDE} \\ \mathbf{ACD} &= \mathbf{BDE} = \mathbf{ABF} = \mathbf{CDE}\end{aligned}$$

Either **ABD** or **ACD** should be utilized to partition the 16 treatment combinations into 2 eight-run miniexperiments (i.e., blocks). In Section 3.3.2 we saw that a full factorial is partitioned into two blocks using the highest-order interaction. Analogous to that simpler result, for fractional factorials one chooses a contrast with no lower-order effects in its alias set. Since all four-, five-, and six-factor interactions are aliased with main effects or two-factor interactions for this 2^{6-2} , the best choice for creating the blocks is a set of aliased three-factor interactions.

Bisgaard (1994) provided a simple introduction to the task of choosing blocking for fractional designs with 8 or 16 runs. Bisgaard then described one six-factor, 16-run experiment that was part of a long series of experiments involving printed circuit boards. The particular experiment involved a hot air solder leveler (HASL). Because the HASL process was prone to disturbances from external conditions, the 16-run experiment was divided into four blocks of size four, in hopes of being able to maintain consistency of external conditions across the four runs within each block. The fractional factorial chosen was identical to the one above. The four blocks were constructed by confounding **ABD** and **ACD** with blocks. This is the best choice, since it sacrifices information for only one alias set involving two-factor interactions—the set **AE = BC = ABCDEF = DF**—since **BC** is the generalized interaction of **ABD** and **ACD**. Whereas a full factorial in four blocks confounds three factorial effects with blocks, a fractional factorial in four blocks confounds three alias sets of effects with blocks. Since this 2^{6-2} design is a one-fourth fraction, there are a total of $3(4) = 12$ factorial effects confounded with block differences.

The best choice of design generators may be different for fractional factorials in blocks than for completely randomized fractional factorials. Consider, for instance, a 16-run five-factor experiment in four blocks. Figure 10.1 considers two possible designs. Design 1 is the resolution V 2^{5-1} design blocking on the three two-factor interactions **CE**, **DE**, and **CE(DE) = CD**. Design 2 is the resolution IV 2^{5-1} design (with **E = ABC**), blocking on **ACD**, **BCD**, and **ACD(BCD) = AB**. Design 1 confounds three two-factor interactions with blocks, whereas Design 2 only confounds two two-factor interactions with blocks. However, Design 1 provides 12 df for main effects and two-factor interactions, one more than Design 2. Surprisingly, by several criteria proposed in the literature, Design 2 is preferred to Design 1. We now summarize and compare the criteria for ranking fractional factorial designs in blocks, advocating a criterion that does favor Design 1.

Suppose a regular 2^{k-f} fraction is to be conducted in 2^b blocks of size 2^{k-f-b} . Then the defining relation for this $(1/2)^f$ fraction will contain $2^f - 1$ factorial effects, and $(2^b - 1)2^f$ factorial effects will be confounded with blocks. The word length pattern (wlp) of the defining relation is used to summarize the aberration of the fractional factorial; see Section 5.2.5. Similarly, it is useful to create a word length pattern for the factorial effects confounded with blocks:

Column	Design 5-1.1 (E = ABCD) Blocking on Columns 7 and 11 (W_1 -Optimal Design)	Design 5-1.4 (E = ABC) Blocking on Columns 13 and 14 (W_2 -Optimal Design)
1	A	A
2	B	B
3	AB	AB = CE = Block
4	C	C
5	AC	AC = BE
6	BC	BC = AE
7	DE	= Block
8	D	E
9	AD	D
10	BD	AD
11	CE	BD
12	CD	= Block
13	BE	ACD = Block
14	AE	BCD = Block
15	E	DE

Fig. 10.1. Comparison of two five-factor designs in four blocks of size 4

$$\text{wlp}_b = (A_{2,1}, A_{3,1}, \dots, A_{k,1}), \quad (10.1)$$

where $A_{j,1}$ denotes the number of j -factor interactions confounded with blocks. For Design 1, $\text{wlp}_b = (3, 3, 0, 0)$, whereas for Design 2, $\text{wlp}_b = (2, 4, 0, 0)$.

The dilemma in ranking these two designs is that Design 1 has higher resolution than Design 2, but it has a worse blocking word length pattern. Neither design dominates the other. For this reason, both are listed by Sun, Wu, and Chen (SWC) (1997) as admissible designs. SWC tabulated fractional factorial designs for up to nine factors for run sizes 128 and less, where admissible designs are identified by the labeling $k - f.i/Bb.1$, where i denotes the ranking in terms of aberration for the 2^{k-f} fraction, and $Bb.1$ indicates that this design has the best wlp_b among all blocking schemes for putting this fraction in 2^b blocks. For instance, Designs 1 and 2 are labeled 5-1.1/B2.1 and 5-1.2/B2.1, respectively, by SWC. Any design listed with the suffix $Bb.2$ is inadmissible in that a better wlp_b is possible for this fraction.

Several authors have proposed means of combining the fractional factorial wlp and wlp_b into a single criterion to use in defining minimum aberration blocked fractional factorial designs. The following four sequences have been featured the most prominently, with the most recently proposed criteria listed first:

- $W_1 = [A_3, A_4, A_{2,1}, A_5, A_6, A_{3,1}, A_7, A_8, A_{4,1}, \dots];$
- $W_2 = [A_3, A_{2,1}, A_4, A_5, A_{3,1}, A_6, A_7, A_{4,1}, \dots];$
- $W_{CC} = [A_3 + A_{2,1}/3, A_4, A_5 + A_{3,1}/10, A_6, A_7 + A_{4,1}/35, \dots];$
- $W_{SCF} = [A_3, A_{2,1}, A_4, A_{3,1}, A_5, A_{4,1}, A_6, A_{5,1}, \dots].$

The W_1 and W_2 sequences were proposed by Cheng and Wu (2002); note that W_1 weights the word length pattern for the fractional factorial design the most heavily. Actually, W_2 was mentioned previously by Chen and Cheng (1999), but they favored W_{CC} over W_2 based on an estimation capacity argument. Sitter, Chen, and Feder (1997) proposed the fourth sequence. However, Chen and Cheng persuasively criticized W_{SCF} for placing $A_{4,1}$ before A_6 ; Chen and Cheng showed an example where W_{SCF} prefers a design that aliases pairs of three-factor interactions over a design for which all effects up to three-factor interactions are estimable. Thus, W_{SCF} contradicts the hierarchical principle that values lower-order effects over higher-order effects, so we consider it no further.

For five factors in four blocks of size 4, Design 1 in Figure 10.1 is preferred by W_1 , whereas Design 2 is preferred by W_2 and W_{CC} . This is typical of these criteria, in that W_1 has a stronger aversion to aliasing caused by words of length $2j$ in the defining relation, whereas W_2 and W_{CC} have a stronger aversion to confounding j -factor interactions with blocks. This is the essential difference in Designs 1 and 2. Cheng and Wu (2002) argued that W_2 is justified for situations where follow-up experiments are expected, since they will likely undo the worst aliasing among factorial effects. However, if no follow-up design is anticipated, or block effects are not more likely than two-factor interactions among the factors, then the W_1 sequence is preferred. Indeed, I would prefer Design 1 in most situations, with its higher resolution, even though by blocking one forfeits information about three two-factor interactions.

To illustrate again the difference among these criteria, consider the case of nine factors in four blocks of size 8. The minimum aberration design 9-4.1 listed in Table G.3 is optimal under criterion W_1 if one blocks on columns 3 (**AB**) and 29 (**ACDE**). As reported by Xu and Lau (2006, p. 4102), the optimal design under the W_2 criterion is obtained by blocking a higher aberration design. Figure 10.2 contrasts these two designs. W_1 favors the design on the left because it has smaller A_4 (6 vs. 8), whereas W_2 favors the design on the right because it only confounds two two-factor interactions with blocks. However, this better confounding with blocks comes at a serious price, requiring an even resolution IV design with only 15 df for two-factor interactions, whereas the minimum aberration design provides 21 df for two-factor interactions and even has eight clear two-factor interactions.

The W_1 sequence begins with A_3 and A_4 . Therefore, the W_1 criterion corresponds closer to the (unblocked) minimum aberration designs of Appendix G than do the W_2 and W_{CC} criteria. For $N = 16, 32, 64$, and 128 the minimum aberration designs are W_1 optimal except for the following cases:

- $N = 16$, in 8 blocks of size 2, for $k = 5$: use a resolution IV fraction.
- $N = 32$, in 16 blocks of size 2, for $k = 7, 8, 9, 10$: an even resolution IV design is used, not the minimum aberration designs, since otherwise the blocks would confound a main effect.

Column	Design 9-4.1 Blocking on Columns 3 and 29 (W_1 -Optimal Design)	Design 9-4.3 Blocking on Columns 6 and 26 (W_2 -Optimal Design)
1	A	A
2	B	B
3	$AB = EG$	= Block
4	C	C
5	$AC = EH$	$AC = BF = EH$
6	$BC = GH$	$BC = AF$
7	DF	F
8	D	D
9	$AD = EJ$	$AD = BG$
10	$BD = GJ$	$BD = AG = HJ$
11	CF	G
12	$CD = HJ$	$CD = FG$
13	BF	
14	AF	
15	F	$CG = DF = EJ$
16	E	E
17	$AE = BG = CH = DJ$	$AE = CH$
18	$BE = AG$	$BE = FH$
19	G	
20	$CE = AH$	$CE = AH = GJ$
21	H	H
22	FJ	
23	$CG = BH$	$BH = DJ = EF$
24	$DE = AJ$	$DE = FJ$
25	J	
26	FH	
27	$DG = BJ$	$CJ = EG$
28	FG	
29	$DH = CJ$	= Block
30		$BJ = DH$
31	EF	$AJ = GH$
		J

Fig. 10.2. Comparison of two nine-factor designs in four blocks of size 8

- $N = 32$, in 8 blocks of size 4, for $k = 21$, where a weak minimum aberration design is slightly better for blocking.
- $N = 64$, in 32 blocks of size 2, for $k = 8, \dots, 20$: an even resolution IV design is used, not the minimum aberration designs, since otherwise the blocks would confound a main effect.
- $N = 64$ in 16 or 32 blocks, for $k = 7$: the resolution VI design is used.
- $N = 64$ in 8 blocks of size 8, for $k = 15$: a weak minimum aberration design is used.
- $N = 128$ in 64 blocks of size 2, for $k = 10, \dots, 40$: an even resolution IV design is used, not the minimum aberration designs, since otherwise the blocks would confound a main effect.

- $N = 128$, for $k = 12, 13, 17, 18, 21, 23, 35$ and various block sizes: a weak minimum aberration design is used.
- $N = 128$ in blocks of size 4 or 8, for $k = 8$: the resolution VII design is used.
- $N = 128$ in blocks of size 4, for $k = 25, 28, 29$: the design with second lowest A_4 is used.

In the remaining cases, the minimum aberration design listed in Table G provides the fractional factorial that is optimal for blocking according to the W_1 criterion. In Appendix H, we list the optimal blocking according to the W_1 criterion. These results were first obtained by Cheng and Wu (2002) for $N = 16$, by Xu and Lau (2006) for designs of size 32 and 64, and by Xu (2008) for $N = 128$ up to $k = 40$. For situations where one of the other criteria is preferred, see the tables in Sun, Wu, and Chen (1997) or Xu and Lau (2006).

To illustrate the use of Appendix H, consider the situation of experimenting with seven factors, where blocks of size 4 are dictated by the circumstances (i.e., batches of a critical raw material, or the time required to perform the runs within a day impose this constraint). Then the possibilities, depending on the number of blocks to be performed are as follows:

- Four blocks of size 4: Design 7-3.1, blocking on columns 3 and 5.
- Eight blocks of size 4: Design 7-2.1, blocking on columns 5, 11, and 19.
- Sixteen blocks of size 4: Design 7-1.2 ($\mathbf{G} = \mathbf{ABCDE}$) blocking on columns 3, 12, 21, and 33 ($\mathbf{AB}, \mathbf{CD}, \mathbf{ACE}, \mathbf{AF}$). Five of the 21 two-factor interactions are confounded with blocks.
- Thirty-two blocks of size 4: 2^7 with blocking as specified in Appendix E. Once again, 5 of the 21 two-factor interactions are confounded with blocks.

For 16 runs, the minimum aberration design from Table G.2 uses columns 7, 11, and 13 as generators; blocking on columns 3 and 5, three of the seven sets of aliased two-factor interactions are confounded with blocks

$$\begin{aligned}\mathbf{AB} &= \mathbf{CE} = \mathbf{DF} \leftarrow \text{Block} \\ \mathbf{AC} &= \mathbf{BE} = \mathbf{DG} \leftarrow \text{Block} \\ \mathbf{AD} &= \mathbf{BF} = \mathbf{CG} \\ \mathbf{AE} &= \mathbf{BC} = \mathbf{FG} \leftarrow \text{Block} \\ \mathbf{AF} &= \mathbf{BD} = \mathbf{EG} \\ \mathbf{AG} &= \mathbf{CD} = \mathbf{EF} \\ \mathbf{BG} &= \mathbf{CF} = \mathbf{DE}\end{aligned}$$

This design would be primarily for estimating main effects; the aliasing among two-factor interactions and confounding of nearly half of these with blocks severely limits ability to identify interaction effects.

If eight blocks of size 4 are possible, then the 21 two-factor interactions are spread among 18 orthogonal contrasts, 4 of which are confounded with blocks:

AD	AE	AB = CF
AG	BD	AC = BF ← Block
BE	BG	AF = BC
CD	CE	DE ← Block
CG	DF	DG ← Block
EF	FG	EG ← Block

This design is much better than the 16-run design for estimating two-factor interactions. From its 31 df, the 24 df not confounded with blocks permit estimation of the 7 main effects and 14 two-factor interactions. One may analyze data from this experiment by constructing two normal probability plots—one for the 24 contrasts not confounded with blocks and one for the 7 contrasts confounded with blocks, assuming block effects may be viewed as random. The second of these normal probability plots analyzes the interblock information, to determine whether one or more of the interactions confounded with blocks might be active.

Performing a 2^{7-1} or a full 2^7 in blocks of size 4 would eliminate the aliasing among two-factor interactions, but each would still confound five two-factor interactions with blocks. Neither of these options seems worthwhile. If increased precision is required or three-factor interactions are believed important, one should use partial confounding (i.e., using different blocking schemes for two or more 2^{7-2} fractions). Butler (2006) provided some theory for such design construction and considered similar examples for up to six factors.

The blocked designs in Appendix H avoid confounding any main effects with blocks. Thus, the blocks of size 2 are always mirror-image pairs (i.e., with treatment combinations \mathbf{x} and $-\mathbf{x}$). Only even fractional factorial designs can be performed in such blocks, since even/odd fractional factorials do not contain mirror-image treatment combinations. For these blocked designs, two-factor interactions are estimable only from interblock information. Such information should not be ignored. The $N/2$ odd-effect estimates for a saturated model should be plotted in one normal plot, and the $N/2 - 1$ even-effect estimates should be plotted in a second normal plot.

In general, for a regular 2^{k-f} factorial in 2^b blocks, there are $2^{k-f} - 2^b$ contrasts orthogonal to blocks and $2^b - 1$ contrasts confounded with blocks. Unless block effects are expected to be negligible, it is recommended that Lenth's PSE be calculated using only the $2^{k-f} - 2^b$ estimates not confounded with blocks. Loepky and Sitter (2002) presented critical values for Lenth t statistics calculated using a PSE based on $2^{k-f} - 2^b$ contrasts (for designs of size 64 and smaller). When there are eight or more blocks, computing a separate PSE from the 2^b estimates confounded with blocks is recommended, in hopes of identifying large interaction effects that were confounded with blocks. This analysis assumes random block effects. Section 10.3.1 illustrates such an analysis.

10.2 Split-Unit Designs for Fractional Factorials

In Section 3.5, we considered full factorial designs conducted as split-unit designs, where the treatment combinations for the whole-unit factors were randomly assigned to groups of experimental units referred to as whole units, and then within each whole unit, the treatment combinations for the remaining factors were randomly assigned to the smaller, split units. Here we consider the same type of nested randomization, but for fractional factorial designs. For the blocked fractional factorial designs considered in Section 10.1 and listed in Appendix H, we only considered blocking schemes where the main effect contrast for each factor is orthogonal to blocks; that is, each individual block is a fractional factorial with resolution II or more. For split-unit designs, each whole unit is a block of split-unit runs with resolution I, since whole-unit factors are held fixed within each whole unit. Thus, different designs are required to provide fractional factorial designs with split-unit randomization.

Addelman (1964) was the first to provide tables of split-unit fractional factorial designs for the convenience of practitioners. However, Addelman's list is quite limited, presenting only 40 designs. Specifying a split-unit fractional factorial design entails specifying the following:

- k_w , the number of whole-unit factors
- k_s , the number of split-unit factors
- N_w , the number of whole units
- N_s , the number of split-unit runs per whole unit.

The requirements are for a particular fractional factorial design with $k = k_w + k_s$ factors, $N = N_w \times N_s$ runs, and two additional properties. First, the number of distinct treatment combinations for the whole-unit factors must not exceed N_w . Second, no split-unit factor may be aliased with a contrast that is constant within whole units.

For an initial example, consider the situation where 7 whole-unit factors are explored using 16 whole-unit treatment combinations and an additional 8 factors are varied in each group of $N_s = 4$ runs per whole unit. The required design involves 15 factors in 64 runs. However, not just any 2^{15-9} design can be employed. For instance, the minimum aberration 2^{15-9} from Appendix G, with $A_4 = 30$, is not suitable, since every seven-factor projection results in more than 16 treatment combinations. Thus, although this fraction might accommodate a split-unit design with 32 whole units of size 2, it cannot produce a design with 16 whole units of size 4. Kulahci, Ramirez, and Tobias (2006) report the minimum aberration split-unit design for this situation to be the fourth best 2^{15-9} design in terms of aberration, with $wlp = (0, 33, 54, \dots)$. It can be constructed using columns 7, 11, and 13, plus the first four basic columns as whole-unit factors, and columns 19, 21, 35, 37, 57, and 58, plus the last two basic columns as split-unit factors. Three other designs have lower aberration, but none have a seven-factor projection with just 16 distinct treatment combinations. Thus, the design obtained by Kulahci et al. (2006)

is the minimum aberration split-unit design for this situation. This design has six clear two-factor interactions. Four of the clear interactions involve two split-unit factors and so are denoted $S * S$ interactions. The other two clear interactions are $W * S$ interactions; that is, they are the interaction of one whole-unit factor and one split-unit factor. This is the first of three alternative split-unit designs listed in Table 10.1. We now consider two more alternatives.

Table 10.1. Alternative fractional factorial split-unit designs with a 2^{7-3} whole-unit design and a 2^{15-9} combined design

Criterion	A_4	M	alp
Minimum aberration	33	43	6, 19, 15, 0, 2, 1
Maximum degrees of freedom	34	45	12, 18, 5, 9, 0, 1
Maximum clear two-factor interactions	55	42	27, 0, 0, 0, 12, 3

As we saw in Chapter 7, there are various criteria besides minimum aberration for ranking resolution IV designs. One other criteria is to maximize M , the degrees of freedom for two-factor interactions. The maximum M 2^{15-9} design may be constructed using columns 7, 11, 13, 19, 21, 25, 35, 60, and 63 as generators. This design has $A_4 = 34$, one more than the minimum aberration split-unit design, but it has $M = 45$ df for two-factor interactions and twice as many clear two-factor interactions ($a_1 = 12$). We illustrate the construction of this design, utilizing the notation of Addelman (1964) and others who use the letters **A**, **B**, ... for whole-unit factors and **P**, **Q**, ... for split-unit factors. Since we must have 16 distinct whole-unit factor treatment combinations, factors **A–D** are taken as the first four basic factors; using **E** = **ABC**, **F** = **ABD**, and **G** = **ACD**, we obtain the required 2^{7-3} whole-unit design. Factors **P** and **Q** are assigned to the remaining basic columns, and the other six generators, which all involve both whole-unit and split-unit factors are used to define **R–W** as follows: **R** = **ABE**, **S** = **ACE**, **T** = **ADE**, **U** = **ABF**, **V** = **CDEF**, and **W** = **ABCDEF**. This design provides estimation of the 15 main effects, 12 clear two-factor interactions, 23 sets of aliased two-factor interactions, and 3 df for higher-order interactions (**BCD** plus two split-unit contrasts). The following combinations of two-factor interactions are estimable:

Seven whole-unit contrasts	
AB = CE = DF	= PR = QU = VW
AE = BC = FG	= RS
AF = BD = EG	= RT
AC = BE = DG	= PS
AD = BF = CG	= PT
AG = CD = EF	= ST
BG = CF = DE	

38 Split-unit contrasts

AR = BP = ES = FT	DW = FV
AP = BR = CS = DT	DV = FW
AS = CP = ER = GT	PQ = RU
AT = DP = FR = GS	PW = RV
BS = CR = PE	PV = RW
BT = DR = PF	QW = UV
CT = DS = PG	QV = UW
ET = FS = GR	GQ
AU = BQ	GU
AQ = BU	GV
AW = BV	GW
AV = BW	SQ
CU = QE	SU
DU = QF	SV
PU = QR	SW
CQ = EU	TQ
CW = EV	TU
CV = EW	TV
DQ = FU	TW

This is the second design summarized in Table 10.1.

Kulahci et al. (2006) also present the split-unit fractional factorial design for this situation that maximizes the number of clear two-factor interactions. This, the last design summarized in Table 10.1, has 27 clear two-factor interactions, but $A_4 = 55$ and $M = 42$, both of which are inferior to the other options. All three designs summarized in Table 10.1 use the same 2^{7-3} design for the whole-unit factors with seven length-4 words among **A–G**. They differ in the remaining six generators. For the maximum df design with aliasing shown above, eight $S * S$ interactions appear in the whole-unit contrasts, since they are aliased with $W * W$ interactions. The minimum aberration and maximum clear designs alias 7 and 15 $S * S$ interactions with whole-unit contrasts, respectively.

Which criterion is best in general? First, maximum clear designs are not relevant unless the particular set of clear two-factor interactions can accommodate the interactions of greatest interest. Product array designs discussed later in this section sometimes provide the maximum number of clear two-factor-interactions for a particularly relevant set of interactions; Section 10.3.2 analyzes such a design. The maximum clear 2^{15-9} design provides clear estimates for all two-factor interactions involving two split-unit factors. All other two-factor interactions are aliased in sets of size 5 or 6. Thus, for many situations, this is a very poor design, and the maximum M and minimum A_4 designs are much preferred. This author prefers designs that maximize the number of estimable two-factor interactions, especially when A_4 is close to the minimum achievable for a split-unit design. Kulahci et al. (2006) advocated custom split-

unit designs that incorporate the specific constraints and estimation priorities of the application. Such design construction typically requires the assistance of a knowledgeable statistician, who knows how to utilize lists of alternative fractional factorial designs and/or design construction software. Kulahci et al. (2006) utilized SAS[®] *Proc Factex* to find alternate designs. One may also construct these designs by writing a computer program that specifies a particular whole-unit factor design and the overall design size $N_w \times N_s$, identifies the eligible columns to use as generators for the split-unit factors, and then repeatedly chooses generators at random; each randomly generated design is evaluated and the best few are kept for further consideration.

Bingham and Sitter (2001) discuss at length many issues regarding the choice of a split-unit design. For each combination of parameters (k_w , N_w , k_s , N_s) considered, one must consider what fractional factorial designs are possible. Existing lists of split-unit designs do cover many cases, especially for $N_w \times N_s = 16$ and 32. Bingham and Sitter (1999) list minimum aberration fractional factorial split-unit designs of size 16. See also Mukerjee and Wu (2006, Table 8.1). Huang, Chen, and Voelkel (1998) provided a table of designs of sizes 16, 32, 64, and 128, but not all their designs are minimum aberration. For example, Huang et al. (1998) listed a 2^{15-9} design with $A_4 = 49$ and $M = 30$, which is clearly inferior to the minimum aberration split-unit design noted in Table 10.1. Bingham, Schoen, and Sitter (2004, Corrigendum 2005) listed minimum aberration split-unit designs with a small number of whole-unit factors (specifically, $k_w \leq 3$, $N_w = 8$ or 16, and $N_s = 2$ or 4); their original Table 2 has several errors, so the entire corrected table was published in a 2005 corrigendum. McLeod and Brewster (2004) provided many similar 32-run designs. Bingham and Sitter (2001) listed 16- and 32-run designs that seek as a secondary criterion to minimize the number of $S * S$ interactions confounded with blocks. Yang, Zhang, and Liu (2007) provide theoretical results and list a few designs for large k_w and small N_s (2 or 4).

Robust parameter design is one arena where split-unit designs often naturally arise (Box and Jones 1992, Bisgaard 2000). In these cases, a simple and sometimes efficient design is to select one fractional factorial design for the factors that can be conveniently controlled, and a second fractional factorial design for the “noise factors”—inputs to which one would like to make the process insensitive. If every treatment combination of the control factors is assigned the same combinations for the noise factors, then the design is a product array and every Control*Noise interaction is clear of aliasing with main effects and other two-factor interactions, provided the main-effects-only model is estimable. The example of Section 10.3.2 has this form.

For certain situations, split-unit designs can also be constructed from orthogonal arrays that are not regular 2^{k-f} designs. For $N_s = 2$, Tyssedal and Kulahci (2005) proposed designs of the form

$$\begin{bmatrix} \mathbf{D}_w & \mathbf{D}_s \\ \mathbf{D}_w & -\mathbf{D}_s \end{bmatrix}$$

where $[D_w, D_s]$ denotes the partitioning of any k -factor orthogonal array into k_w columns for the whole-unit factors and the remaining $k_s = k - k_w$ columns for the split-unit factors. See also Kulahci and Bisgaard (2005), who considered product array designs involving regular and nonregular designs of the form $2^{k_w-f} \times \text{OA}(N_s, 2^{k_s}, 2)$ and $\text{OA}(N_w, 2^{k_w}, 2) \times 2^{k_s-f}$. They also considered designs that are half (or quarter) of a product array, such as designs where half the whole-unit factor treatment combinations are assigned the split-unit design D_s and the other half are assigned the foldover design $-D_s$. Analysis of these designs involves the complication of partial aliasing for the whole-unit and/or split-unit contrasts.

JMP's Custom Design platform permits flexibility in the choice of k_w , k_s , N_w , N_s , and the particular model of interest. When N_w and N_s are powers of 2, optimal design algorithms will choose a regular fractional factorial design for some models; see Goos (2002). JMP's Custom design algorithm accommodates searches for split-split-unit designs as well.

Regular fractional factorial designs may be used to construct both split-split-unit designs and multiway split-unit designs, as was done in Section 3.6 for full factorial treatment structures. The split-split-split-unit Example 3.7 was actually a fractional factorial design; it was presented in Chapter 3 as a full 2^4 rather than as a 2^{6-2} by ignoring two insignificant factors. Schoen (1999) described the construction of a split-split-unit design to investigate cheese making. The fractional factorial design and its analysis are detailed in Section 10.3.3. Miller (1997) discussed how to construct and analyze two-way split-unit designs (also known as strip-plot or strip-block designs) for fractional factorials. Miller's 2^{10-5} numerical example is reanalyzed Section 10.3.4. For construction and analysis of three-way split-unit (i.e., strip-strip-block) designs and multiway split-unit designs, see Paniagua-Quiñones and Box (2008, 2009), Bates and Mee (1998), and Bingham et al. (2008).

10.3 Analysis of Fractional Factorials with Randomization Restrictions

10.3.1 Example 10.1: Foundry experiment

Young, Abraham, and Whitney (1991) presented a wonderful example of using experimental design for process improvement in the course of actual production. A multifactor experiment was conducted over a period of several months to improve the quality of crankshafts. The experiment's primary objectives were to decrease hardness variation in the crankshafts while preventing defects due to gas holes and incomplete nodularity. To assess variability, it was determined that several molds would be examined for each run of the experiment and that multiple hardness measurements would be taken on each of the six crankshafts within each mold. Two additional objectives were to see (i) whether throughput might be increased by raising the line speed without

causing any problems and (ii) whether new hardness specifications might be achieved without the use of tin and chrome in the alloy.

Each hour of production the iron casting process required 55 tons of iron and 300 tons of sand (used to pack around the molds). Six factors related to iron chemistry were varied (carbon, silicon, manganese, copper, chrome, and tin). In addition, pour temperature and line speed were varied, along with compactibility of the sand and the percentage of clay mixed with the sand. The authors noted that “Once it is established that there are many potential causes of the quality problem(s) under study and that, on the basis of available data, there is not a clear prioritization of which are the most important, a designed experiment is the most efficient way to proceed.” The article does not mention the exact levels used, other than that the current line speed and a 20% increase in speed were selected. Other choices of levels were made with the intent of seeing an effect without causing any runs with excessive scrap.

Given the huge volume of sand required, it was decided that the clay percentage factor must be held fixed each day. Given this constraint, four treatment combinations could be explored in an 8-hour shift. Recall that once each new set of factor settings is achieved, the process must operate long enough to adequately sample short-term mold-to-mold variation. In order to estimate the main effects and several two-factor interactions, it was decided that a 32-run experiment was needed. Thus, 8 different days were required. Since some potentially important sources of variation vary little within a week, it was decided to spread out the 8 blocks over a two-month period.

Consider now two different means of handling clay percentage. If clay percentage is to remain a factor in the experiment, then it must be confounded with days and we must construct a split-unit design with $k_w = 1$, $N_w = 8$, $k_s = 9$, and $N_s = 4$. The minimum aberration design with $A_4 = 10$ and $M = 21$ is suitable, provided we confound clay percentage and six two-factor interaction contrasts with blocks. Alternatively, we might conduct a 2^{9-4} design in eight blocks of size 4 as prescribed in Appendix H. If clay percentage varies from block to block (perhaps as a result of natural process variation, then its effect is confounded with other changes over time. The description by Young et al. (1991) is somewhat ambiguous regarding which of these strategies was adopted.

Young et al. (1991) do not provide the data for any of the responses considered. However, one can reconstruct data to match the contrast totals they do provide. Using their factor labels **B–J** for the optimal W_1 blocked 2^{9-4} design from Appendix H, a design that matches the description by Young et al. may be constructed as follows:

- Assign the factors Silicon (**C**), Carbon (**B**), Manganese (**D**), Tin (**G**), and Copper (**H**) to the five basic columns 1, 2, 4, 8, and 16 respectively.
- Use column 15 for Line speed, **F** = **CBDG**.
- Use column 19 for Compactibility, **E** = **CBH**.
- Use column 21 for Chrome, **I** = **CDH**.

- Use column 25 for Temperature, $\mathbf{J} = \mathbf{CGH}$.
 - Construct eight blocks by blocking on columns 3, 5, and 24 (i.e., \mathbf{CB} , \mathbf{CD} , and \mathbf{GH}).
 - If Clay percentage is a factor that is purposely varied, use column 30 to assign its levels ($\mathbf{A} = \mathbf{BDGH}$).

This design, with response values reconstructed to correspond to the contrast totals in Young et al. (1991), is provided in Table 10.2. The design is arranged in blocks. We now illustrate the analysis of randomized block and split-unit experiments using this reconstructed data.

Table 10.2. Reconstructed design and data for Young et al.'s (1991) foundry example

Since this experiment was conducted in 8 blocks, there are 7 between-block contrasts and 24 within-block contrasts. If Clay percentage is not purposely changed, but simply allowed to vary from block to block, then we have a 2^{9-4} design in the factors **B–J**, and estimates for a saturated model are as given in Table 10.3. For convenience of interpretation, the estimates are sorted from largest to smallest. Lenth's PSE is calculated separately for the 7 between-block contrasts and the 24 within-block contrasts. All between-block contrasts are insignificant, and three active main effects are clearly evident among the within-block estimates.

Table 10.3. Lenth *t* statistics for Young et al.'s (1991) foundry example

Effects	Estimate	PSE	Lenth <i>t</i>
Between Blocks			
GE = BJ	47.0	24	1.96
GI = DJ	-26.0	24	-1.08
GH = CJ	24.0	24	1.00
BD = EI	-16.0	24	-0.67
BDGH = ····	13.5	24	-0.56
BC = EH	6.0	24	0.25
CD = HI	-1.0	24	-0.04
Within Blocks			
H	-68.5	18	-3.81
D	68	18	3.78
C	-65	18	-3.61
CG = HJ	-23	18	-1.28
DE = BI	23	18	1.28
CF	-19.5	18	-1.08
FG	-17	18	-0.94
E	-17	18	-0.94
DH = CI	-17	18	-0.94
B	-16	18	-0.89
BG = EJ	15	18	0.83
DF	13	18	0.72
I	-12.5	18	-0.69
BE = CH = DI = GJ	12	18	0.67
HF	10.5	18	0.58
BF	10	18	0.56
G	10	18	0.56
J	-8	18	-0.44
EF	-7.5	18	-0.42
DG = IJ	-7	18	-0.39
F	4.5	18	0.25
FJ	4	18	0.22
FI	-2.5	18	-0.14
CE = BH	0	18	0.00

Table 10.3 illustrates the analysis as a randomized block design. The analysis is essentially the same if Clay percentage is included as a factor confounded with blocks. If Clay percentage is included as a 10th factor, it must be aliased with the **BDGH** interaction; any other choice creates a resolution III design. The minimum aberration 2^{10-5} design obtained by using the generator **A** = **BDGH** adds **AF** to the long alias set in Table 10.3 and aliases each of the other two-factor interactions involving **A** with a two-factor interaction involving **F**; the resulting alias length pattern (alp) has $a_1 = 20$ and $a_5 = 1$. The interpretation of the estimates is unchanged, since the three statistically significant estimates only involve main effects.

The minimum aberration 2^{10-5} design is also a maximum M design, since it is second-order saturated. This same design can accommodate one, two, three, or four whole-unit factors in eight blocks of four runs. The only difference in these designs is which effects are confounded with blocks; the aliasing is unchanged. Bingham, Schoen, and Sitter (2005) listed the split-unit designs for $k_w = 1$ and 2, and Huang, Chen, and Voelkel (1998) listed the designs for $k_w = 3$ and 4. In each case, the analysis is performed by computing Lenth's PSE separately for the 7 whole-unit contrasts and the 24 split-unit contrasts. This 2^{10-5} design can also accommodate 5 or 6 whole-unit factors if the number of whole units is increased to 16 (see Huang et al. 1998). It cannot accommodate 7 (or 8) whole-unit factors even in 16 whole units, because no 2^{7-3} design has aliasing that is embedded in the aliasing for the minimum aberration 2^{10-5} ; this becomes obvious by considering the alp for the 7-factor and 10-factor designs.

10.3.2 Example 10.2: Injection molding robust parameter design experiment

Engel (1992) briefly described a 10-factor, 32-run injection molding experiment where percent shrinkage was measured. The objective of the experiment was to determine which levels for seven controllable factors would make percent shrinkage insensitive to fluctuations in the three other factors that naturally vary during production. The seven controllable factors and three naturally varying factors are shown in Table 10.4. The 32 treatment combinations and the percent shrinkage for each are shown in Table 10.5 in a format that emphasizes the structure of the design: a 2^{7-4} design for the control factors crossed with a 2^{3-1} design for the noise factors. Although this design is resolution III, and so not maximum resolution for a 2^{10-5} fraction, it has the maximum number of clear two-factor interactions. (The product of two saturated main effect designs always produces a combined design with the maximum number of clear interactions.)

Table 10.4. Factors for Engel's (1992) injection molding experiment

Controllable Factors							Noise Factors			
A: Cycle time							M: Percent regrind			
B: Mold temperature							N: Moisture content			
C: Cavity thickness							O: Ambient temperature			
D: Holding pressure										
E: Injection speed										
F: Holding time										
G: Gate size										

Table 10.5. Engel's (1992) injection molding experiment treatment combinations, with observed percent shrinkage

A	B	C	D	E	F	G	(M, N, O) treatment combination			
							−,−,−	−,+,+	+,-,+	+,+,-
-1	-1	-1	-1	-1	-1	-1	2.2	2.1	2.3	2.3
-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

Engel's work does not describe any unit structure for the design; we do not know whether this was performed as a completely randomized design or as a split-unit design. Since ambient temperature could not be randomly assigned, we will assume that each noise factor combination (i.e., each column) represents a whole unit; that is, each time a desired combination of levels for **M**, **N**, and **O** was achieved, the eight treatment combinations for the controllable factors were performed—we will assume—in random order. Thus, this is a 2^{7-4} design performed four separate times, each time under a different set of conditions for Ambient temperature and the plastic's Regrind percentage and Moisture content. Such a unit structure is much more convenient than performing this as a completely randomized design. All that is sacrificed is our ability to test the significance of the whole-unit main effects **M**, **N**, and **O**. The other 28 effects of interest, namely the 7 control factor main effects and the 21 Control*Noise interactions, are all estimable as split-unit (i.e., within-block) contrasts.

We compute Lenth's PSE using only the 28 split-unit estimates. These are arranged in order in Table 10.6. The .05 critical value for Lenth's *t* computed from 28 estimates is $c_{.05}^{\text{IER}} = 2.067$, so the estimates for **CN**, **A**, and **EN** are

statistically significant. We conclude that shrinkage percentage is greater at the high level for Cycle time. More important for the purpose of the experiment, the effect of Moisture content on shrinkage depends on the levels for Cavity thickness and Injection speed. No other effects seem to matter. Since no effects involving **M** or **O** are evident, the process seems to be robust to Percent regrind and Ambient temperature. Depending on the levels of Cavity thickness and Injection speed, Moisture content may or may not have an effect. In particular, when **C** and **E** are both low (rows 1 and 8 in Table 10.5) or both high (rows 4 and 5), the b_{CN} and b_{CE} terms effectively cancel and so the Moisture content can vary without affecting shrinkage percentage.

The analysis just performed is known as response modeling in the robust parameter design literature. For product array designs such as in Table 10.5, one may also perform an analysis by computing a performance measure statistic for each control factor combination. For “nominal is best” responses such as shrinkage percent, $\log(\text{standard deviation})$ is commonly used as one performance measure. However, as Steinberg and Bursztyn (1994) pointed out, such an analysis can be misleading. Since rows 1, 4, 5, and 8 in Table 10.5 all have low Holding time, one might conclude that Holding time is critical to achieving consistent shrinkage. However, the response modeling analysis does not support this conclusion, as no effect involving **F** appears important. This apparent contradiction can arise whenever the control factor design has low resolution; Rosenbaum (1994, 1996) illuminated this point most effectively. For the injection molding example, $\mathbf{F} = -\mathbf{CE}$. Although modeling Shrinkage percentage indicates that $\mathbf{C} = \mathbf{E}$ is beneficial, this is mistakenly interpreted by response modeling as desiring $\mathbf{F} = -1$. Sometimes fitting a model for an appropriate performance measure is the simplest analysis for robust parameter design applications, but this example makes the point that such an analysis alone is not sufficient if the control array has low resolution.

One other detail needs to be examined for these data. There are hints of a problem with these data that becomes evident in two ways. First, a half-normal plot of the estimates in Table 10.6 show an unexpected concentration of $|\text{estimates}|$ near 0.15; see Figure 10.3. When a half-normal plot shows a concentration distant from zero, an outlier is suspected. Second, when we fit a hierarchical model including the three statistically significant terms **A**, **CN**, and **EN**, plus main effects **C**, **E**, and **N**, an extreme outlier is detected corresponding to the treatment combination $\mathbf{A} = \mathbf{B} = \mathbf{C} = -1$, $\mathbf{D} = \mathbf{E} = \mathbf{F} = \mathbf{G} = 1$, $\mathbf{M} = \mathbf{N} = \mathbf{O} = -1$. Figure 10.4 shows a histogram of the residuals, with the minimum residual of $0.3 - 2.64 = -2.34$, far removed from the other residuals. If this observation is omitted, the factorial effect contrasts are no longer orthogonal. However, using forward selection (or using Lenth’s method for a saturated model with 30 correlated main effects and interactions), we find some evidence for **DM** and **BM**. Thus, it appears that Holding pressure and/or Mold temperature can mitigate any Regrind percent effect. The outlier had previously concealed this insight.

Table 10.6. Analysis for Engel's (1992) injection molding experiment

Term	Estimate	PSE	Lenth <i>t</i>
Whole-unit contrasts			
O	0.150		
N	0.138		
M	-0.050		
Split-unit contrasts			
CN	0.450	0.197	2.29
A	0.425	0.197	2.16
EN	-0.419	0.197	-2.13
D	-0.281	0.197	-1.43
G	-0.231	0.197	-1.17
FO	0.169	0.197	0.86
AN	-0.163	0.197	-0.83
GN	0.156	0.197	0.79
GO	0.156	0.197	0.79
CO	-0.150	0.197	-0.76
E	0.144	0.197	0.73
EO	0.144	0.197	0.73
FN	0.144	0.197	0.73
BO	-0.138	0.197	-0.70
AO	-0.125	0.197	-0.63
CM	-0.125	0.197	-0.63
DO	0.119	0.197	0.60
BN	-0.113	0.197	-0.57
DN	0.106	0.197	0.54
EM	0.106	0.197	0.54
DM	-0.094	0.197	-0.48
B	-0.075	0.197	-0.38
C	0.062	0.197	0.32
BM	0.062	0.197	0.32
AM	-0.050	0.197	-0.25
FM	-0.044	0.197	-0.22
F	-0.019	0.197	-0.10
GM	0.019	0.197	0.10

10.3.3 Example 10.3: Split-split-unit cheese making experiment

Cheese making involves progressively smaller units of material. When the raw milk arrives from farms, it is stored in huge tanks. At some future time, the milk is transported to smaller tanks where curds are produced. The curds

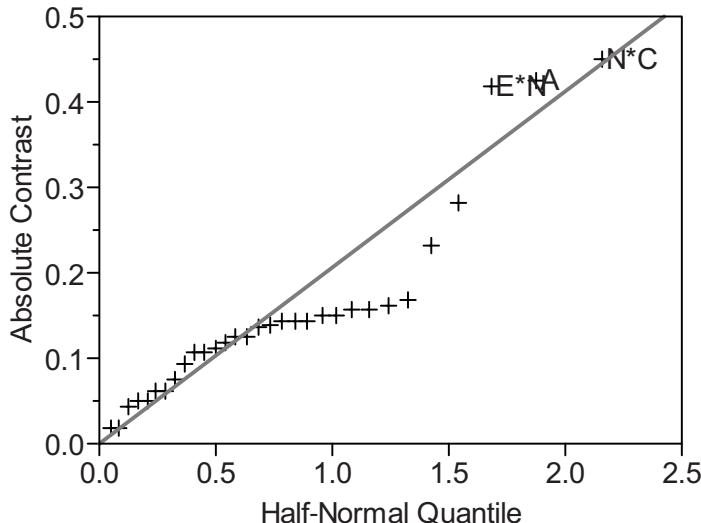


Fig. 10.3. Half normal plot of estimates for shrinkage saturated model

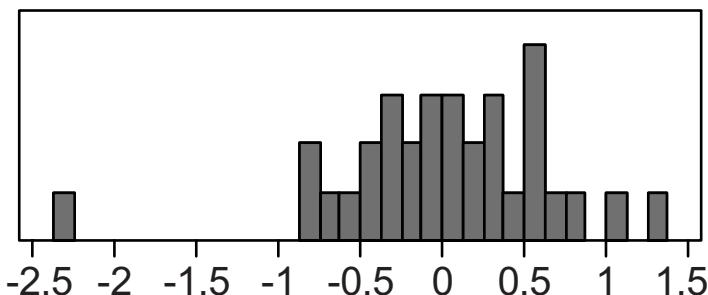


Fig. 10.4. Histogram of residuals from shrinkage reduced model

from each smaller tank are subsequently used to manufacture many individual cheeses. Schoen (1997, 1999) described an experiment involving two milk factors (**A**, **B**), five curd factors (**C**, **D**, **E**, **F**, **G**), and two cheese factors (**K**, **L**), all with two levels, plus one additional cheese factor **M** with four levels. As described by Wu (1989) and others, two-level designs are easily modified to accommodate four level factors. The factor **M**, with levels (0, 1, 2, 3), is constructed from two-level factors **H** and **J** as $\mathbf{M} = 1.5 + \mathbf{J} + 0.5\mathbf{H}$. The contrasts **J**, **H**, and **HJ** represent three orthogonal contrasts that together account for the **M** main effect. Here, **M**'s levels actually correspond to equally spaced levels for a quantitative factor. In such cases, **J** is highly correlated with **M**'s linear contrast, and **HJ** is **M**'s quadratic contrast. For more details about constructing designs with one or two four-level factors, see Section 13.4.

Schoen's design is best understood as being built in stages. The milk experiment involved eight milk storage tanks and two factors. Thus, the milk stratum is a completely randomized 2^2 design. For the curds stratum, each milk unit is viewed as a block containing four curd units. The five curd factors and two milk factors together then form a 2^{7-2} design in eight blocks of size 4. The generators for the design were $\mathbf{F} = \mathbf{ABCD}$ and $\mathbf{G} = \mathbf{ABDE}$, with blocking on the whole-unit factors \mathbf{A} and \mathbf{B} and the contrast $\mathbf{CE} = \mathbf{FG}$. This design is shown in Table 10.7.

Table 10.7. Milk and curd treatment combinations for Schoen's (1997) cheese making experiment

Milk Unit	Curd		Unit					
	A	B		C	D	E	F	G
1	-1	-1	1	-1	-1	-1	1	1
1	-1	-1	2	1	1	1	1	1
1	-1	-1	3	-1	1	-1	-1	-1
1	-1	-1	4	1	-1	1	-1	-1
2	1	1	5	1	1	1	1	1
2	1	1	6	-1	-1	-1	1	1
2	1	1	7	-1	1	-1	-1	-1
2	1	1	8	1	-1	1	-1	-1
3	-1	-1	9	-1	-1	1	1	-1
3	-1	-1	10	1	1	-1	1	-1
3	-1	-1	11	1	-1	-1	-1	1
3	-1	-1	12	-1	1	1	-1	1
4	-1	1	13	1	1	1	-1	-1
4	-1	1	14	-1	-1	-1	-1	-1
4	-1	1	15	1	-1	1	1	1
4	-1	1	16	-1	1	-1	1	1
5	1	-1	17	1	1	-1	-1	1
5	1	-1	18	-1	-1	1	-1	1
5	1	-1	19	-1	1	1	1	-1
5	1	-1	20	1	-1	-1	1	-1
6	1	1	11	1	1	-1	1	-1
6	1	1	22	1	-1	-1	-1	1
6	1	1	23	-1	-1	1	1	-1
6	1	1	24	-1	1	1	-1	1
7	-1	1	25	1	-1	-1	1	-1
7	-1	1	26	1	1	-1	-1	1
7	-1	1	27	-1	-1	1	-1	1
7	-1	1	28	-1	1	1	1	-1
8	1	-1	29	-1	1	-1	1	1
8	1	-1	30	1	-1	1	1	1
8	1	-1	31	-1	-1	-1	-1	-1
8	1	-1	32	1	1	1	-1	-1

Table 10.8. Coded responses and levels for **M** in Schoen (1997)

Curd Unit	M(H,J)			
	0(-1, -1)	1(1, -1)	2(-1, 1)	3(1, 1)
1	100.8	96.6	94.5	96.6
2	100.8	100.8	98.7	92.4
3	98.7	86.1	81.9	92.4
4	111.3	98.7	102.9	100.8
5	147.0	140.7	140.7	140.7
6	138.6	140.7	147.0	140.7
7	151.2	138.6	140.7	136.5
8	149.1	151.2	140.7	142.8
9	96.6	98.7	92.4	96.6
10	107.1	109.2	107.1	107.1
11	107.1	102.9	90.3	86.1
12	98.7	92.4	100.8	92.4
13	100.8	102.9	92.4	96.6
14	90.3	94.5	88.2	79.8
15	100.8	96.6	98.7	90.3
16	96.6	96.6	94.5	88.2
17	144.9	142.8	138.6	138.6
18	138.6	134.4	138.6	130.2
19	144.9	132.3	134.4	134.4
20	147.0	134.4	134.4	138.6
21	138.6	134.4	138.6	126.0
22	138.6	134.4	142.8	126.0
23	147.0	144.9	121.8	147.0
24	138.6	134.4	140.7	147.0
25	96.6	96.6	96.6	88.2
26	88.2	96.6	84.0	84.0
27	88.2	79.8	81.9	86.1
28	105.0	90.3	92.4	90.3
29	142.8	142.8	128.1	138.6
30	147.0	140.7	140.7	136.5
31	138.6	138.6	128.1	134.4
32	149.1	147.0	138.6	132.3

From each curd unit, cheeses were made according to each of 4 treatment combinations, for a total of 128 different combinations of the factors. The two-level pseudo-factors **H** and **J** may be taken as basic factors. Thus, from these, the level of **M** is determined using the formula $\mathbf{M} = 1.5 + \mathbf{J} + 0.5\mathbf{H}$. In addition, the other two cheese factors **K** and **L** were determined as **K** = **BCGHJ** and **L** = **ACGH**. Table 10.8 shows the coded response for each cheese and the levels for **M**. For compactness, levels for **K** and **L** are not

shown but can be determined using the generators. In terms of the 11 two-level factors **A–H** and **J–L**, the defining relation for this design is

$$\begin{aligned}\mathbf{I} &= \mathbf{ABCDF} = \mathbf{ABDEG} = \mathbf{CEFG} \\ &= \mathbf{BCGHJK} = \mathbf{ADFGHJK} = \mathbf{ACDEHJK} = \mathbf{BEFHJK} \\ &= \mathbf{ACGHL} = \mathbf{BDFGHL} = \mathbf{BCDEHL} = \mathbf{AEFHL} \\ &= \mathbf{ABJKL} = \mathbf{CDFJKL} = \mathbf{DEGJKL} = \mathbf{ABCEFGJKL}.\end{aligned}$$

Since **HJ** does not appear in any length-4 or length-5 words, all two-factor interactions involving **M** are estimable; the aliasing among two-factor interactions is from **CEFG**.

The analysis of a regular fraction split-split-unit experiment can be done by fitting a saturated model, sorting the contrasts by stratum, computing the PSE for each stratum, and then determining which estimates are statistically significant. Alternatively, one may perform three separate analyses, one for each stratum. Here we follow the first approach; see Table 10.9. A full factorial model in the seven basic factors **A–E**, **H** and **J** is easily constructed. The seven contrasts that do not change within milk units are listed first; these are **A**, **B**, and **CE**, and their generalized interactions. The PSE for these seven estimates is 1.5504. Only the **A** main effect is statistically significant.

Next, we list 24 more terms corresponding to contrasts that do not change within curd units, but do change between milk units. These are all terms involving the basic factors **A–E** (or their aliases) that did not appear in the list of milk unit contrasts. This is a very efficient design for estimating two-factor interactions because, of the 24 split-unit contrasts, only (**CDE**) is not aliased with a main effect or two-factor interaction. Even **HJK** should be viewed as a two-factor interaction, since **HJ** corresponds to part of the **M** main effect. The PSE from these 24 estimates is 0.911, and all Lenth *t* statistics are less than 2. Thus, none of the 24 split-unit estimates is statistically significant at the .05 level.

There are $127 - 31 = 96$ additional contrasts in the saturated model. The PSE for this stratum is 0.394, much smaller than the PSE for the whole-unit and split-unit strata. The .05 critical value is 1.98, so 10 estimates are statistically significant. However, with 96 Lenth *t* statistics, we would expect about five statistically significant estimates if all the true regression coefficients were zero. Using simulation as discussed in Appendix C, we determined that $P(\text{Maximum Lenth } |t| > 2.67) = .52$; that is, over half the time when none of the 96 effects are active, the largest estimate will exceed $2.67 \times \text{PSE}$. The three largest estimates correspond to two contrasts involving the four-level factor **M**, plus the **FK** interaction. The next eight estimates, with Lenth *t* statistics ranging from 2.67 to 1.83 all correspond to higher-order interactions. Thus, controlling the experimentwise error rate for the cheese stratum and desiring a parsimonious model, we adopt an overall model that includes **A**, **FK**, and the main effects for **F** and **K** in order to have a hierarchical model, plus a linear effect for **M**.

Table 10.9. Analysis for Schoen's (1997) split-split-unit cheese experiment

Term	Estimate	PSE	Lenth <i>t</i>
Milk stratum			
A	22.083	1.550	14.24
AB	1.903	1.550	1.23
BCE = ...	1.280	1.550	0.83
CE	1.050	1.550	0.68
B	-1.017	1.550	-0.66
ACE = ...	0.558	1.550	0.36
CDG =	0.197	1.550	-0.13
Curd stratum			
C	-1.739	0.911	-1.91
FA	-1.181	0.911	-1.30
CB	-1.050	0.911	-1.15
CA	-0.984	0.911	-1.08
F	-0.952	0.911	-1.05
CG = EF	-0.919	0.911	-1.01
E	-0.886	0.911	-0.97
CF = EG	-0.787	0.911	-0.86
ACG = HL =	0.755	0.911	-0.83
CDE = ...	0.722	0.911	0.79
GA	0.722	0.911	0.79
G	0.623	0.911	0.68
EB	0.591	0.911	0.65
BCG = HJK =	0.558	0.911	-0.61
FD	-0.427	0.911	-0.47
CD	-0.427	0.911	-0.47
D	-0.394	0.911	-0.43
GB	-0.328	0.911	-0.36
GD	0.230	0.911	0.25
DA	-0.230	0.911	-0.25
EA	0.197	0.911	0.22
C	-0.098	0.911	-0.11
FA	0.066	0.911	0.07
CB	0.033	0.911	0.04
Cheese stratum			
J	-2.428	0.394	-6.17
FK	1.378	0.394	3.50
H	-1.181	0.394	-3.00
BEH = ...	1.050	0.394	2.67
BFL = ...	1.017	0.394	2.58
CDH = ...	0.952	0.394	2.42
CHJ =	0.886	0.394	-2.25
BGJ = ...	0.788	0.394	2.00
CFJ = ...	0.787	0.394	2.00
BHK =	0.787	0.394	-2.00
AGJ = ...	0.722	0.394	1.83
(85 smaller estimates not displayed)			

Basing the analysis on half-normal plots, Schoen (1997) selected a model with only **A**, **C**, and **M**. The half-normal plots for the split-unit and split-split-unit contrasts, constructed using Z_Q as defined in (2.4), are shown in Figures 10.5 and 10.6, respectively. One would need more knowledge about the response y and the factors **C**, **F**, and **K** to know whether the estimates $b_C = 1.739$ and $b_{FK} = 1.378$ make sense in the context of this application. Whatever the factor **A** represents, its effect on the response is enormous. It is unusual to have a factor with such a large main effect yet not appear in significant interactions.

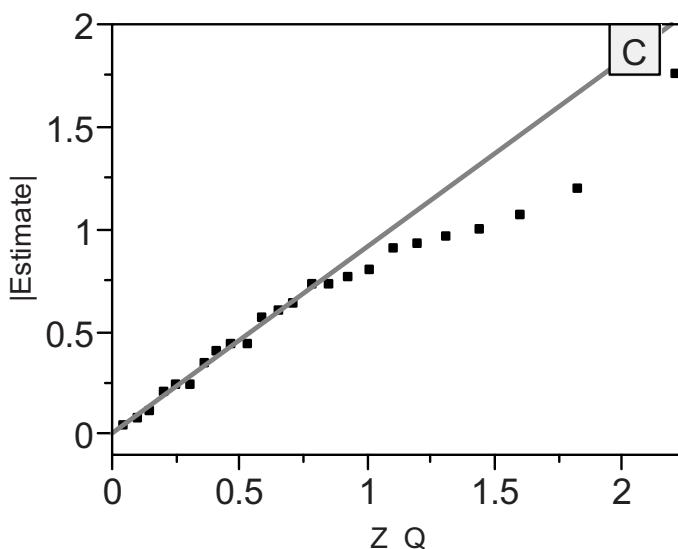


Fig. 10.5. Half-normal plot for 24 split-unit contrasts

10.3.4 Example 10.4: Two-way split-unit washer/dryer experiment

Miller (1997) presented a fractional factorial experiment involving six washing machine factors (**A–F**) and four dryer factors (**P–S**). The experiment's purpose was to minimize wrinkling of clothes after washing and drying. The six washing machine factors were explored using a 2^{6-3} design, running each of four machines twice. When the four machines finished their first run, the wet clothes were distributed to four different dryers, putting some garments from every washer in each dryer. Thus, the first run of all machines produced 16 treatment combinations ($2^{6-4} \times 2^{4-2}$); another 16 treatment combinations were obtained using the second run of the machines. The full design is one-half of a $2^{6-3} \times 2^{4-1}$. The generators used were

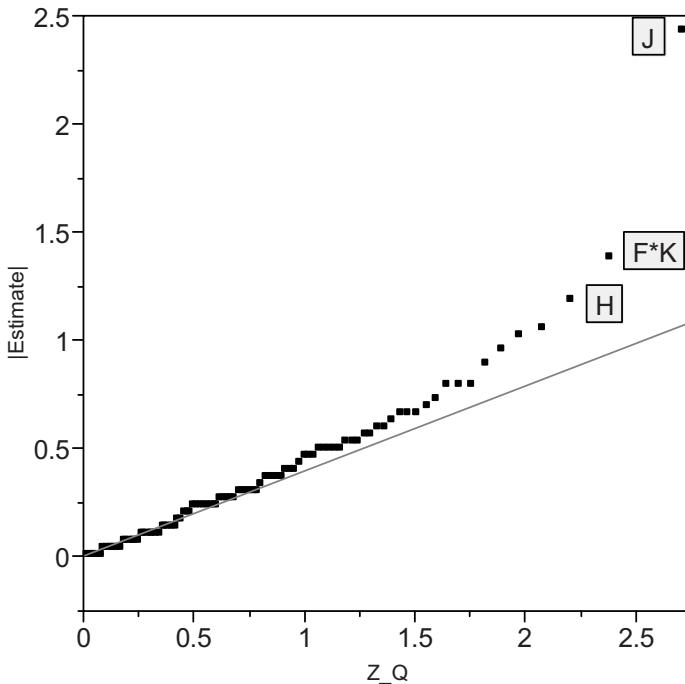


Fig. 10.6. Half-normal plot for 96 split-split-unit contrasts

- $\mathbf{C} = \mathbf{AB}$, $\mathbf{E} = \mathbf{ABD}$, $\mathbf{F} = \mathbf{BD}$ to define the washer treatment combinations;
- $\mathbf{S} = \mathbf{PQR}$ to define the dryer treatment combinations;
- $\mathbf{AD} = -\mathbf{PQ}$ to define which washer and dryer treatment combinations were assigned together.

Table 10.10 lists the 32 treatment combinations defined by these generators and shows the wrinkle response measurement for garments corresponding to each combination.

The layout of Table 10.10 emphasizes that this was an eight-run experiment for Washer factors conducted in two blocks of size 4. The first four rows were each assigned to a different washing machine; later, when Block 2 was performed, the last four row were subsequently randomly assigned to the four machines. Including an effect for blocks and the main effects $\mathbf{A-F}$ utilizes all of the degrees of freedom. Note that this analysis assumes that there are no systematic differences among the four washing machines. Otherwise, the row-to-row errors would be correlated across blocks. Similarly, the four columns of data in Block 1 and the four columns in Block 2 correspond to an eight-run experiment for Dryer factors. Here a resolution IV design is used, so there is no aliasing of two-factor interactions with the Dryer factor main effects.

Table 10.10. Treatment combinations and wrinkle measurements for Miller's (1997) laundry experiment

	A	B	C	D	E	F	(P, Q, R, S)			
Block 1							(-, -, -, -)	(+, +, -, -)	(-, -, +, +)	(+, +, +, +)
	-1	-1	1	1	1	-1	3.19	2.75	3.02	2.63
	1	-1	-1	-1	1	1	4.01	3.33	3.79	2.82
	-1	1	-1	1	-1	1	3.77	3.36	3.47	3.08
	1	1	1	-1	-1	-1	3.83	3.48	4.25	3.94
Block 2							(-, +, -, +)	(+, -, -, +)	(-, +, +, -)	(+, -, +, -)
	-1	-1	1	-1	-1	1	2.28	1.88	2.91	2.37
	1	-1	-1	1	-1	-1	2.95	3.25	3.11	2.85
	-1	1	-1	-1	1	-1	2.40	1.89	3.51	2.38
	1	1	1	1	1	1	4.05	3.68	3.24	3.31

The proper analysis of a two-way, split-unit design such as this entails identifying which factorial effects are estimated with the same precision. The preceding discussion should make the analysis intuitive. There are four categories of contrasts, each impacted differently by the sources of variation:

1. One contrast between washers and between dryers: $\mathbf{AD} = -\mathbf{PQ}$ and its aliases.
2. Six contrasts between washers and within dryers: $\mathbf{A-F}$ and their aliases.
3. Six contrasts within washers and between dryers: $\mathbf{P-S}$, \mathbf{PR} , and \mathbf{PS} , and their aliases.
4. Eighteen contrasts within washers and within dryers: all remaining contrasts, including 12 pairs of aliased Washer*Dryer interactions.

These 31 estimates are displayed by category in Table 10.11, with Lenth t statistics provided for the last 3. At least one effect is evident in each group, **A**, **B**, **P**, and **AS** = **-DR**. The $\alpha = .05$ critical value for six contrasts is 2.211 [Loeppky and Sitter (2002); alternatively, use the code in Appendix C]. If small values of the response correspond to fewer wrinkles, then we prefer **A** = **B** = -1, **P** = +1 and, to take advantage of the interaction term, either **S** = 1 or **D** = **R**.

Because the PSE is similar for the Dryer and Unit strata, there does not appear to be excess variation associated with dryer runs. In contrast, there does appear to be washer run variation, making that stratum's PSE larger than the PSE that captures just the variability associated with individual garment units within washers and dryers.

Table 10.11. Lenth t statistics for Miller's (1997) laundry experiment

Stratum	Effects	Estimate	PSE	Lenth t
Block	AD=BE=CF=-PQ=-RS	-0.271		
Washer	A = BC = EF	0.344	0.090	3.82
	B = AC = DF	0.203	0.090	2.26
	C = AB = DE	0.026	0.090	0.29
	D = BF = CE	0.083	0.090	0.92
	E = AF = CD	-0.024	0.090	-0.27
	F = AE = BD	0.060	0.090	0.67
Dryer	P	-0.212	0.0497	-4.26
	Q	-0.034	0.0497	-0.69
	R	0.018	0.0497	0.36
	S	-0.063	0.0497	-1.27
	PR = QS	-0.033	0.0497	-0.67
	QR = PS	0.022	0.0497	0.44
Unit	AP = -DQ	0.051	0.0525	0.98
	AQ = -DP	-0.094	0.0525	-1.79
	AR = -DS	-0.098	0.0525	-1.86
	AS = -DR	0.161	0.0525	3.07
	BP = -EQ	-0.001	0.0525	-0.01
	BQ = -EP	0.064	0.0525	1.23
	BR = -ES	0.027	0.0525	0.51
	BC = -ER	0.056	0.0525	1.06
	CP = -FQ	0.041	0.0525	0.79
	CQ = -FP	0.019	0.0525	0.36
	CR = -FS	0.015	0.0525	0.29
	CS = -FR	0.104	0.0525	1.98
	APR = ...	0.010	0.0525	0.19
	AQR =-	0.030	0.0525	-0.57
	BPR = ...	0.026	0.0525	0.49
	BQR =-	0.007	0.0525	-0.13
	CPR = ...	0.058	0.0525	1.10
	CQR =-	0.035	0.0525	-0.67

Strip-block designs, such as the one employed by Miller (1997), provide for very efficient experimentation. For more regarding the construction of strip-block designs, see Butler (2004) and Vivacqua and Bisgaard (2009). For an example of a strip-strip-block design and how to construct such three-way split-unit designs in general, see Paniagua-Quiñones and Box (2008, 2009).

10.4 Sequences of Fractional Factorial Designs

We summarize literature pertaining to running sequences of fractional factorial designs. In Section 10.1 we discussed running a fractional factorial design in blocks, where the entire design was selected from the start and randomization was used to determine the order in which the blocks were performed, as well as the order of runs within blocks. Here, we consider sequences of two or more blocks, where a deliberate ordering of the blocks is followed. Chapter 9 discussed sequences of just two blocks. In Section 9.4, we introduced the concept of augmenting a resolution III fractional factorial design with an additional resolution III fraction in order to increase the resolution of the combined design to IV. This is the most common (and simplest) scenario involving a sequence of fractional factorial designs. In Section 9.5, we considered adding a block of runs to a resolution IV design. There, which set of foldover runs is best to perform is judged utilizing results of the initial experiment. What additional literature exists regarding running sequences of blocks?

Li and Jacroux (2007) contemplated how an experiment should be chosen when both the initial fraction as well as a possible follow-up foldover design are to be conducted in blocks. This differs from Sections 9.4 and 9.5 in that in Chapter 9 we assumed that the initial designs themselves were completely randomized.

Addelman (1969) investigated various sequences of fractional factorial designs for 4–10 factors and shows how many main effects and two-factor interactions are estimable after each block of runs. For instance, for 7 factors, Addelman considers 13 different sequences for 8 blocks of size 8. One sequence has eight resolution III blocks that together form the resolution VII half fraction. Other sequences use eight resolution II blocks to construct the half-fraction with resolution VI or V. Resolution I blocks are also considered. However, Addelman’s tables assume estimation of a model without any block effects. If the designs are run in blocks and the analysis accounts for the possibility of additive block effects, the number of estimable effects and the precision of estimable effects Addelman reported do not apply if his blocks have resolution less than III. Jacroux (2006) gave recommended sequences for experiments run in four to eight blocks for up to $k = 9$ factors. Practitioners should refer to Jacroux’s article first, since it uses familiar notation, it focuses on the more useful sequences of blocks, and it does not ignore the possibility of block effects. However, Jacroux (2006) only tabulated sequences up to 32 runs (except for the case of 6 factors). We now discuss one example for $k = 9$ factors in blocks of size 16 from Addelman (1969).

Addelman’s sequence 9.12B involves partitioning the resolution VI 2^{9-2} fraction in eight blocks and then sequencing these blocks in order to estimate all the two-factor interactions after just four blocks. Addelman defined the resolution VI fraction using $\mathbf{F} = \mathbf{ABCDE}$ and $\mathbf{J} = \mathbf{ABCGH}$ and then blocks on \mathbf{ABDH} , \mathbf{AEG} , and \mathbf{ABC} . Each block is a 2^{9-5} fraction with $A_3 = 6$ and $A_4 = 9$. The first block has the following aliasing:

$$\begin{aligned}
\mathbf{A} &= -\mathbf{BC} = -\mathbf{EG} \\
\mathbf{B} &= -\mathbf{AC} = -\mathbf{FJ} \\
\mathbf{C} &= -\mathbf{AB} = -\mathbf{DH} \\
\mathbf{D} &= -\mathbf{CH} = -\mathbf{EF} \\
\mathbf{E} &= -\mathbf{AG} = -\mathbf{DF} \\
\mathbf{F} &= -\mathbf{BJ} = -\mathbf{DE} \\
\mathbf{G} &= -\mathbf{AE} = -\mathbf{HJ} \\
\mathbf{H} &= -\mathbf{CD} = -\mathbf{GJ} \\
\mathbf{J} &= -\mathbf{BF} = -\mathbf{GH} \\
\mathbf{AD} &= \mathbf{BH} = \mathbf{GF} \\
\mathbf{AF} &= \mathbf{CJ} = \mathbf{DG} \\
\mathbf{AH} &= \mathbf{BD} = \mathbf{EJ} \\
\mathbf{AJ} &= \mathbf{CF} = \mathbf{EH} \\
\mathbf{BE} &= \mathbf{CG} = \mathbf{DJ} \\
\mathbf{BG} &= \mathbf{CE} = \mathbf{HF}
\end{aligned}$$

This is a special fraction in that, ignoring three-factor and higher-order interactions, all the chains have length 3. By contrast, the minimum aberration design 9-5.1 has one alias set involving five terms (see Section 6.2).

Addelman (1969) recommended the sequence of blocks shown in Table 10.12. After the second block, one has a 2^{9-5} in 2 blocks, and the 15 alias sets are reduced to:

$$\begin{aligned}
\mathbf{A} &= -\mathbf{EG} \\
\mathbf{B} &= -\mathbf{FJ} \\
\mathbf{AB} &= \mathbf{DH} \\
\mathbf{CH} &= \mathbf{EF} \\
\mathbf{E} &= -\mathbf{AG} \\
\mathbf{F} &= -\mathbf{BJ} \\
\mathbf{G} &= -\mathbf{AE} \\
\mathbf{CD} &= \mathbf{GJ} \\
\mathbf{J} &= -\mathbf{BF} \\
\mathbf{AD} &= \mathbf{BH} \\
\mathbf{CJ} &= \mathbf{DG} \\
\mathbf{AH} &= \mathbf{BD} \\
\mathbf{CF} &= \mathbf{EH} \\
\mathbf{CG} &= \mathbf{DJ} \\
\mathbf{CE} &= \mathbf{HF}
\end{aligned}$$

After the third block, we have a three-quarter fraction design that eliminates the aliasing with main effects but retains the aliasing of nine pairs of two-factor interactions due to the length-4 words **ABDH**, **CEFH**, and **CDGJ**. The fourth block changes the sign of these four contrasts, and so the first four blocks combine to make an irregular resolution V 2^{9-3} fraction. Its variance inflation factors are 1, 1.5, and 2.

Adding additional blocks improves the precision of these coefficients, so that after six blocks, we have a 3/4 fraction equivalent to the design proposed

in Section 8.3.4, with VIFs of 1 and 1.125. The Section 8.3.4 design was constructed from three resolution III 2^{9-4} with $A_3 = 2$ and $A_4 = 3$. Here, blocks 1 and 6 form such a fraction, and 2 and 5 (3 and 4) form an equivalent fraction from the same family. The advantage of Addelman's sequence is that we have a resolution III design in just 16 runs and an irregular resolution V design after 64 runs.

Table 10.12. One of Addelman's sequences of 2^{9-5} fractions; eight blocks together compose the 2^{9-2} fraction with $\mathbf{I} = \mathbf{ABCDEF} = \mathbf{ABCGHJ} = \mathbf{DEFGHJ}$

Block	ABDH	AEG	ABC	df for 2fi's
1	+	—	—	6
2	+	—	+	21
3	+	+	—	27
4	—	+	+	36
5	—	—	—	36
6	—	—	+	36
7	—	+	—	36
8	+	+	+	36

Sometimes blocks are defined by confounding several sources of variation rather than a single source. For instance, Holms and Sidik (1971) discussed an example involving a nuclear reactor, where blocks are identified by different fuel cycles and time within a fuel cycle. Equipment and instrument changes between cycles account for some differences, whereas within-cycle blocks account for radiation changes as the fuel is consumed. In such cases, Cheng, Wu, and Wu (2003) recommended that the shortest factorial effects confounded with blocks not be assigned to main effects of the confounding factors. To illustrate their idea, suppose we plan to investigate seven factors in two fuel cycles, and each fuel cycle will be divided into four blocks of size 4. As we saw in Section 10.1, design 7-2.1 is recommended with blocking on columns 5, 11, and 19. The full set of seven blocking contrasts and the corresponding factorial effects are given in Figure 10.7. Since no two-factor interactions are confounded with columns 11, 19, or 31, we use these to define the blocks displayed in Figure 10.8, so that the average fuel cycle difference and any linear effect within a fuel cycle will primarily bias the estimates for three-factor interactions but not any two-factor interactions. For more about protection from trends in the errors, see Section 13.5.

Figure 10.8 assumes that all 16 runs planned for each fuel cycle can be completed. Holms and Sidik (1971) commented that nuclear reactor experiments such as this often terminate a fuel cycle early. If this is the case, the treatment combinations from the canceled blocks should be performed using a third fuel cycle. One benefit of blocking is that such repairs to the experimen-

Column 5: **AC = BF = ...**
 Column 11: **ABD = CDF = ...** ← Block generator
 Column 14: **BCD = ADF = EG = ...**
 Column 19: **ABE = CEF = ...** ← Block generator
 Column 22: **BCE = AEF = DG = ...**
 Column 24: **DE = BCG = AFG = ...**
 Column 29: **ACDE = ABG = CFG = ...** ← Block generator

Fig. 10.7. All contrasts of optimal 2^{7-2} confounded with blocks of size 8

	Runs 1-4: ABD = -1 ABE = -1	Runs 5-8: ABD = -1 ABE = 1	Runs 9-12: ABD = 1 ABE = -1	Runs 13-16: ABD = 1 ABE = 1
Fuel Cycle 1: ACDE = -1	Block 1	Block 2	Block 3	Block 4
Fuel Cycle 2: ACDE = 1	Block 5	Block 6	Block 7	Block 8

Fig. 10.8. Preferred blocking assignment for 2^{7-2} example

tal design cause no complications for the analysis. In addition, by choosing a blocked fractional factorial with carefully ordered sequences of blocks [e.g., as given by Jacroux (2006) and Holms (1998)], early termination of a design may still permit estimation of most effects of interest.

More Fractional Factorial Design Examples

This chapter contains the analysis of four interesting experiments reported in the literature. The sections are as follows:

Section 11.1. A Mirror-Image Foldover with Unexpected Results

Section 11.2. Steepest Ascent with Constraints

Section 11.3. A Group Screening Experiment

Section 11.4. Nonorthogonal Blocking for a Fractional Factorial

11.1 A Mirror-Image Foldover with Unexpected Results

Ahuja, Ferreira, and Moreira (2004) described a sequence of experiments run to determine which component of a standard growth medium was limiting the growth of clumps of a marine bacterium. The standard growth medium with sucrose was composed of 18 components, in quantities defined by the lower levels in Table 11.1. Fifteen of these 18 were varied in an experiment, with the high level four times the lower level. According to the authors, the remaining three components (NaCl, HEPES, and Sucrose) were held constant at the standard concentration, in order to provide degrees of freedom for error. The 20-run orthogonal array from Plackett and Burman (1946) was utilized—the OA($20, 2^{19}, 2$) shown earlier in Table 6.26. The protease activity, measured as the micrograms of azocasein digested per hour over a 28-hour period, are presented in Table 11.2. The authors reported that each treatment combination was performed in duplicate and that each of these was measured twice. Thus, the values in Table 11.2 are averages of four measurements. Our estimates will differ slightly from those reported in the article due to rounding error in the y values reported in the article.

Table 11.1. Factor levels for Ahuja et al. (2004) bacterium experiment

Factors	Levels	
	-	+
A NaCl (g/L)	18.8	18.8
B KCl (g/L)	0.4	1.6
C MgSO ₄ ·7H ₂ O (g/L)	1.9	7.6
D MgCl ₂ ·6H ₂ O (g/L)	1.5	6.0
E CaCl ₂ ·2H ₂ O (g/L)	0.4	1.6
F HEPES (g/L)	4.9	4.9
G K ₂ HPO ₄ (mg/L)	15.3	61.1
H Na ₂ CO ₃ (mg/L)	10.0	40.0
J Sodium citrate·2H ₂ O (mg/L)	4.56	18.2
K Fe ₂ (SO ₄) ₃ ·H ₂ O (mg/L)	3.14	12.5
L H ₃ BO ₃ (mg/L)	2.9	11.6
M MnCl ₂ ·4H ₂ O (mg/L)	1.8	7.2
N ZnSO ₄ ·7H ₂ O (mg/L)	0.2	0.8
O Na ₂ MoO ₄ ·2H ₂ O (mg/L)	0.04	0.16
P CoSO ₄ ·6.5H ₂ O (mg/L)	0.0484	0.194
Q CuSO ₄ ·5H ₂ O (mg/L)	0.08	0.32
R Sucrose (g/L)	5.0	5.0
S NH ₄ Cl (g/L)	1.0	4.0

Table 11.2. Initial bacterium experiment by Ahuja et al. (2004)

t.c.	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q	R	S	T	y
1	+	+	-	-	+	+	+	+	-	+	-	-	-	-	-	+	+	+	+	0.048
2	-	+	+	-	-	+	+	+	+	-	+	-	-	-	-	-	+	-	-	0.039
3	+	-	+	+	-	-	+	+	+	-	+	-	+	-	-	-	-	-	-	0.039
4	+	+	-	+	+	-	-	+	+	+	-	+	-	+	-	-	-	+	0.029	
5	-	+	+	-	+	+	-	-	+	+	+	-	+	-	+	-	-	+	0.044	
6	-	-	+	+	-	+	-	-	+	+	+	+	-	+	-	+	-	+	0.041	
7	-	-	-	+	+	-	+	-	-	+	+	+	+	-	+	-	+	+	0.046	
8	-	-	-	-	+	+	-	-	+	+	+	+	+	-	+	-	-	-	0.041	
9	+	-	-	-	-	+	+	-	+	+	+	+	+	+	-	+	-	+	0.026	
10	-	+	-	-	-	-	+	+	-	+	+	+	+	+	+	-	-	-	0.038	
11	+	-	+	-	-	-	+	+	-	+	-	+	+	+	+	+	+	+	0.035	
12	-	+	-	+	-	-	-	+	+	-	+	+	-	+	+	+	+	-	0.031	
13	+	-	+	-	+	-	-	-	+	+	-	+	-	+	-	+	-	-	0.034	
14	+	+	-	+	-	+	-	-	-	+	+	-	+	+	-	-	+	-	0.037	
15	+	+	+	-	+	-	-	-	-	+	+	-	+	+	-	-	-	-	0.033	
16	+	+	+	+	-	+	-	-	-	-	+	+	-	+	+	-	+	-	0.036	
17	-	+	+	+	-	+	-	-	-	-	+	+	-	+	+	-	+	+	0.030	
18	-	-	+	+	+	-	+	-	-	-	-	-	+	+	-	+	-	-	0.034	
19	+	-	-	+	+	+	-	+	-	-	-	-	-	-	+	+	-	-	0.031	
20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	0.039	

Note: Four columns (A, F, R, T) were not assigned factors that varied.

11.1.1 Analysis of the initial experiment

There are two approaches to the analysis of this initial experiment. One is to fit an additive model, which would leave 4 df for error. The other is to fit a saturated model, using all 19 columns of the orthogonal array, and then use Lenth's PSE to test for statistical significance. Using Lenth's method, the estimates are presented in Table 11.3, sorted by magnitude. According to Lenth's t , only the estimate for $\text{MnCl}_2 \cdot 4\text{H}_2\text{O}$ ($b_M = 0.00295$) is statistically significant. Furthermore, the estimate is positive, which was the expected sign. This factor explains 27% of the variation in the response. If we had fit a model with only the 15 factors included, the root mean square error, with 4 df, is 0.005055 and the standard error for the estimates is $0.005055/20^{1/2} = 0.00113$; no resulting t statistics exceed $t_{4,0.025} = 2.776$. Ahuja et al. made an error in their analysis and wrongly concluded that the largest four estimates were statistically significant.

Table 11.3. Estimates and Lenth t statistics for initial bacterium experiment

Column	Estimate	PSE	Lenth t
M	0.00295	0.001275	2.31
P	-0.00215	0.001275	-1.69
J	-0.00205	0.001275	-1.61
H	0.00195	0.001275	1.53
A	-0.00175	0.001275	-1.37
F	0.00115	0.001275	0.90
Q	-0.00115	0.001275	-0.90
D	-0.00115	0.001275	-0.90
N	-0.00095	0.001275	-0.75
T	0.00085	0.001275	0.67
L	0.00085	0.001275	0.67
G	0.00055	0.001275	0.43
O	0.00055	0.001275	0.43
S	-0.00055	0.001275	-0.43
E	0.00045	0.001275	0.35
K	-0.00015	0.001275	-0.12
B	-0.00005	0.001275	-0.04
R	-0.00005	0.001275	-0.04
C	-0.00005	0.001275	-0.04

Might other effects be active? There is no conclusive evidence here. The next two largest estimates are both negative, contrary to our expectation for active effects, and have p -values $> .10$. The fifth and sixth largest estimates correspond to columns not assigned a factor. In addition to b_M , only

$b_H = 0.00195$ (for Na_2CO_3) has a positive estimate larger than the PSE. Analysis of the residuals and fitting models with interactions produced nothing useful. These results correspond to several possibilities. Perhaps **M** is the only active factor. Alternatively, there could be other active effects, but because the error variance is large (or interactions are missed, which inflate the variance estimate) we need more data to identify them.

11.1.2 Analysis of a mirror-image foldover fraction

Because of the possibility of interactions, Ahuja et al. decided to run a second 20-run design, obtained by reversing the signs of every column. The treatment combinations and data appear in Table 11.4. This experiment is a disappointment in one respect, in that no estimates are statistically significant. In another respect, this experiment appears promising, in that the mean response is 24% higher than in the first experiment (0.0454 versus 0.0366). We now investigate further to account for this surprising outcome.

Table 11.4. Foldover orthogonal array experiment by Ahuja et al. (2004)

t.c.	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q	R	S	T	<i>y</i>
21	-	-	+	+	-	-	-	-	+	-	+	-	+	+	+	+	-	-	-	0.044
22	+	-	-	+	+	-	-	-	-	+	-	+	-	+	+	+	+	-	+	0.038
23	-	+	-	-	+	+	-	-	-	-	+	-	+	-	+	+	+	+	+	0.039
24	-	-	+	-	-	+	+	-	-	-	+	-	+	-	+	+	+	+	-	0.040
25	+	-	-	+	-	-	+	+	-	-	-	+	-	+	-	+	+	-	0.039	
26	+	+	-	-	+	-	+	+	-	-	-	-	+	-	+	-	+	-	0.039	
27	+	+	+	-	-	+	-	+	-	-	-	-	-	+	-	+	-	-	0.043	
28	+	+	+	+	-	-	+	-	+	-	-	-	-	-	+	-	+	+	0.041	
29	-	+	+	+	+	-	+	-	+	+	-	-	-	-	-	+	-	-	0.039	
30	+	-	+	+	+	-	+	-	-	+	+	-	-	-	-	-	+	+	0.043	
31	-	+	-	+	+	+	-	-	+	-	+	+	-	-	-	-	-	-	0.050	
32	+	-	+	-	+	+	+	-	-	+	-	-	+	-	-	-	-	+	0.051	
33	-	+	-	+	-	+	+	+	-	-	+	-	-	+	-	-	+	-	0.051	
34	-	-	+	-	+	+	+	+	-	-	+	-	-	+	-	+	-	+	0.052	
35	-	-	-	+	-	+	+	+	+	-	-	+	-	-	+	+	+	0.049		
36	-	-	-	-	+	-	+	+	+	-	-	+	-	-	+	-	+	-	0.052	
37	+	-	-	-	+	-	+	-	+	+	+	-	-	+	-	-	-	-	0.048	
38	+	+	-	-	-	+	-	+	-	+	+	+	-	-	+	-	+	-	0.052	
39	-	+	+	-	-	-	+	-	+	-	+	+	+	-	-	+	+	0.048		
40	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	0.049	

Note: Four columns (**A**, **F**, **R**, **T**) were not assigned factors that varied.

Since the main effects can be estimated from each 20-run design, it is useful to compare the estimates from the two designs. Table 11.5 contains the

estimates from each design, as well as the average and the difference. The first scatterplot in Figure 11.1 displays estimates from Experiment 1 versus Experiment 2, and the lower figure plots the difference versus the combined (i.e., average) estimate, excluding the intercept. It is striking how little consistency exists in the estimates from one design to the next. Considering only estimates for the 15 factors, the correlation is .22. Pleasingly, the two factors with the largest positive estimates in Experiment 1 had positive estimates in Experiment 2. The third largest combined estimate is also positive, and it corresponds to column **G** (K_2HPO_4). Although these three estimates are not significantly different from zero, they are the only factors that show promise. In fact, the authors chose these three factors for another experiment—a central composite design—as will be discussed in Section 12.2.

Table 11.5. Main effect estimates from two bacterium experiments by Ahuja et al. (2004)

Column	Exp. 1 Estimate	Exp. 2 Estimate	Combined Estimate	Difference
Intercept	0.03655	0.04535	0.0410	0.0088
A (fixed)	-0.00175	-0.00105	-0.0014	0.0007
B	-0.00005	-0.00025	-0.0001	0.0002
C	-0.00005	-0.00035	-0.0002	0.0003
D	-0.00115	-0.00105	-0.0011	0.0001
E	0.00045	-0.00015	0.0002	0.0006
F (fixed)	0.00115	0.00095	0.0011	0.0002
G	0.00055	0.00235	0.0015	0.0018
H	0.00195	0.00115	0.0016	0.0008
J	-0.00205	0.00205	0.0000	0.0041
K	-0.00015	0.00165	0.0008	0.0018
L	0.00085	0.00105	0.0010	0.0002
M	0.00295	0.00065	0.0018	0.0023
N	-0.00095	0.00105	0.0001	0.0020
O	0.00055	0.00065	0.0006	0.0001
P	-0.00215	0.00005	-0.0010	0.0022
Q	-0.00115	-0.00125	-0.0012	0.0001
R (fixed)	-0.00005	-0.00135	-0.0007	0.0013
S	-0.00055	-0.00145	-0.0010	0.0009
T (empty)	0.00085	0.00105	0.0010	0.0002

One reason for conducting the foldover fraction was to increase the strength (resolution) of the design, so that two-factor interactions that are active would not bias the main effect estimates. Variation among the 40 observations for the combined design partitions into the following sets:

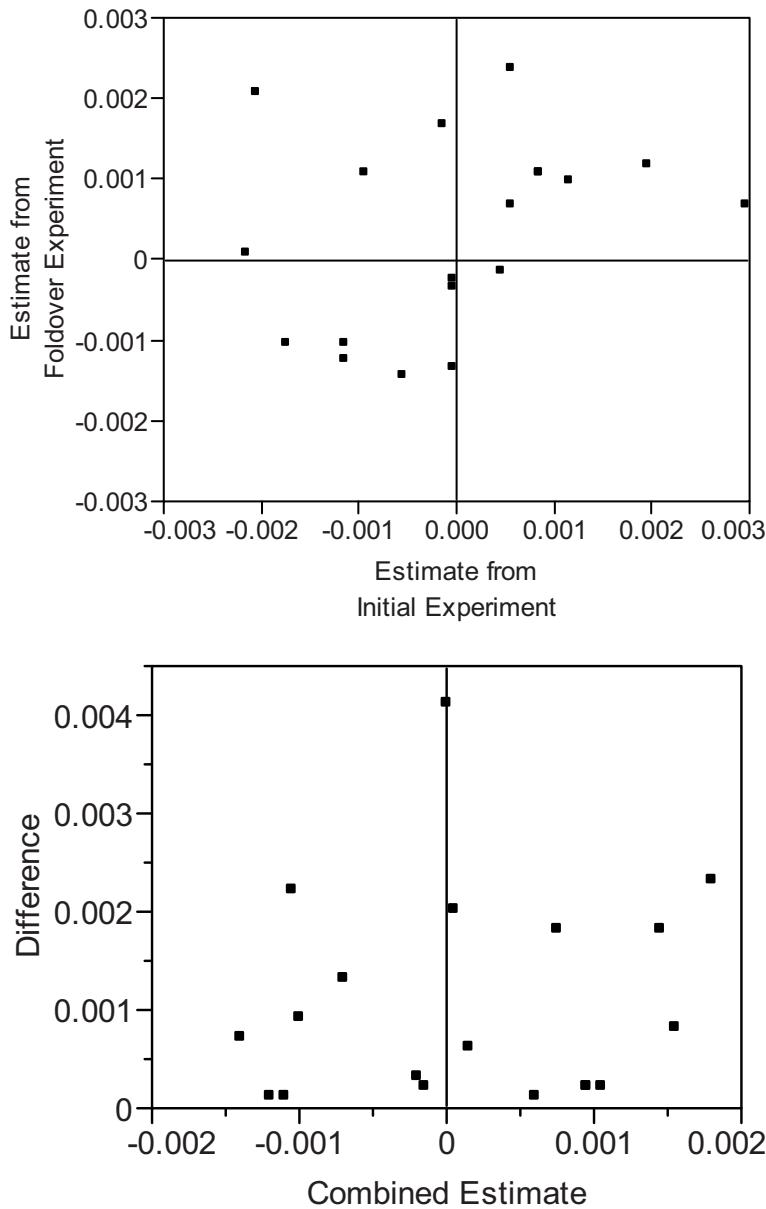


Fig. 11.1. Plots of estimates from Table 11.5, comparing results from initial 20-run orthogonal array and foldover orthogonal array

- 1 df for the difference in mean between first and second experiment. This is a main effect for blocking.
- 15 df for (combined) main effects.
- 4 df for combined estimates for unassigned columns. These correspond to error, or to three-factor and higher (odd-order) interactions.
- 19 df for differences among the main effects. These may be entered in the model as 19 orthogonal Block*Factor interactions, but these contrasts represent linear combinations of two-factor interactions. We would like to discover any prominent two-factor interactions.

Table 11.6 contains an ANOVA, partitioning the effects into these portions, and using the 4 df for unassigned columns as error. The only significant effect is the block effect. Not even the largest main effect estimate (**M**) is statistically significant ($F = 0.0001296/0.00004455 = 2.909$, with p -value = .16). The mean square for interactions, based on the differences in estimates from Experiment 1 to Experiment 2, is negligible. Thus, this ANOVA directs our attention to the question of what changed in the second experiment.

Table 11.6. ANOVA for bacterium experiments by Ahuja et al.

Source	df	SS	MS	F
Blocking	1	0.0007744	0.00077440	17.38
Main effects	15	0.0005764	0.00003843	0.86
Two-factor interactions	19	0.0004169	0.00002194	0.49
Error	4	0.0001782	0.00004455	
Total (corrected)	39	0.0019459		

To investigate the surprising change in mean from the initial orthogonal array to the foldover design, we plot the residuals (from the model with only a block effect) versus run order; see Figure 11.2. The results for the second experiment are disturbing. Runs 21–30 have negative residuals, whereas Runs 31–40 have positive residuals. If this were due to some main effect or a single two-factor interaction, these contrasts would have appeared significant in our previous analysis. Instead, the runs in the residual plot strongly suggest instability in the inputs for this experiment or the measurement process. Something changed after the 30th treatment combination. If one adds main effects for columns **H** and **M** to the model, the residual versus run order plot is slightly improved, but the p -value for positive autocorrelation is still .01.

Ahuja et al. commented that they did each treatment combination in duplicate. However, if there were a shift in the process after the 30th run, making consecutive duplicate runs of the same treatment combination and duplicate measurements offers no protection. Thus, rather than confirm that we have one or two possible main effects, the follow-up fraction here calls into question control of the experimental conditions. If it were found that, for example, there

was a change of technician after the 30th run, an additional blocking variable could be added to distinguish the two halves of the second experiment.

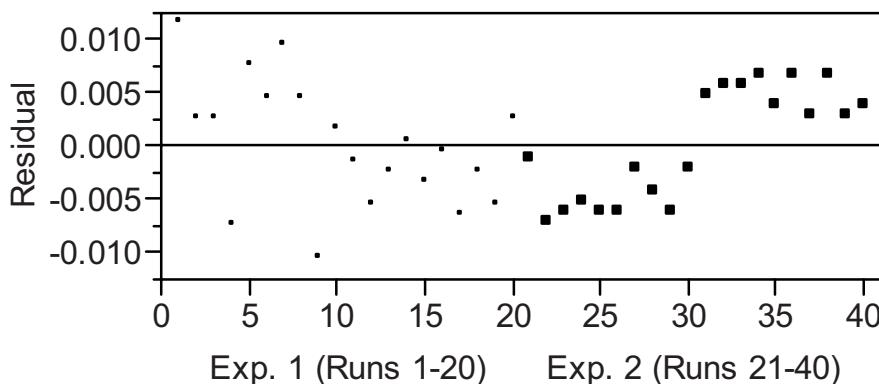


Fig. 11.2. Response versus run order, with runs for the second experiment highlighted

11.2 Steepest Ascent with Constraints

Petersson, Lundell, and Markides (1992) documented a sequence of experiments to improve the time and resolution for separating two enantiomers using supercritical fluid chromatography. Both resolution and retention time are important responses. Achieving satisfactory levels for both requires careful trade-offs in the inputs. The initial experiment involved the factors and levels indicated in Table 11.7. A center run was performed followed by a 2³⁻¹ fractional factorial with $\mathbf{C} = -\mathbf{AB}$. The results for chiral resolution (CR) and retention time (*Time*) for the second enantiomer are shown in Table 11.8. Because the coding changes from experiment to experiment, we use a subscript to distinguish the coded factors in the first versus second experiment. Note the strong positive correlation between resolution and time. All of these values had satisfactory resolution, but the time was considered excessive. The purpose of this first experiment was to determine a treatment combination that achieved the minimally acceptable resolution (CR = 0.274) and then to minimize time subject to achieving minimal resolution.

A first-order model for CR fitted to these data is

$$\widehat{\text{CR}} = 0.795 - 0.200\mathbf{A}_1 - 0.080\mathbf{B}_1 - 0.151\mathbf{C}_1$$

explains 99.6% of the variation in CR. The coefficients for both CR and *Time* are all negative, indicating that trade-offs are required to optimize both the

larger-the-better response CR and the smaller-the-better response *Time*. The authors chose the treatment combination $D_0 = 0.344$ ($\mathbf{A}_1 = 1.47$), $t = 52^\circ$ ($\mathbf{B}_1 = 0.67$), $D_g = 0.013$ ($\mathbf{C}_1 = 1$) as the location for the second experiment. This value is near to the steepest descent path for CR, and yields a predicted value of 0.29, slightly better than required.

Table 11.7. Factor levels for first experiment by Petersson et al. (1992)

Factors	Levels	
	-1	1
A Start density, D_0 (g/mL)	0.27	0.33
B Temperature, t ($^\circ\text{C}$)	47.0	53.0
C Density gradient, D_g (g/mL/min)	0.007	0.013

Table 11.8. First experiment by Petersson et al. (1992)

Run	D_0	t	D_g	\mathbf{A}_1	\mathbf{B}_1	\mathbf{C}_1	CR	<i>Time</i>
1	0.30	50	0.010	0	0	0	0.825	17.336
2	0.27	53	0.013	-1	1	1	0.756	16.270
3	0.33	47	0.013	1	-1	1	0.516	15.077
4	0.33	53	0.007	1	1	-1	0.659	15.943
5	0.27	47	0.007	-1	-1	-1	1.219	25.980

A second five-run experiment was conducted surrounding the new center-point; see Table 11.9. Spacing for the three factors was unchanged. Time for the second experiment centerpoint is 5 min shorter than the initial center. The purpose of this second experiment is to search for additional decreases in time. To determine the best direction as suggested by Mee and Xiao (2008b), we begin by fitting an additive model in the coded variables for each response:

$$\widehat{\text{Time}} = 12.54 - 1.54\mathbf{A}_2 - 1.19\mathbf{B}_2 - 1.08\mathbf{C}_2,$$

$$\widehat{\text{CR}} = 0.264 - 0.239\mathbf{A}_2 - 0.082\mathbf{B}_2 - 0.052\mathbf{C}_2.$$

Table 11.9. Second experiment by Petersson et al. (1992)

Run	D_0	t	D_g	\mathbf{A}_2	\mathbf{B}_2	\mathbf{C}_2	CR	<i>Time</i>
6	0.344	52	0.013	0	0	0	0.281	12.251
7	0.314	55	0.016	-1	1	1	0.365	11.882
8	0.374	49	0.016	1	-1	1	0.050	11.180
9	0.374	55	0.010	1	1	-1	-0.009	10.964
10	0.314	49	0.010	-1	-1	-1	0.632	16.416

To decrease *Time* without changing predicted CR, we take the *Time* steepest descent vector $b_{Time} = (1.54, 1.19, 1.08)$ and subtract the portion that is correlated with steepest descent for CR, $b_{CR} = (.239, .082, .052)$. First we standardize each vector to unit length,

$$\tilde{b}_{Time} = b_{Time} / \| b_{Time} \| = (0.6919, 0.5346, 0.4852)',$$

$$\tilde{b}_{CR} = b_{CR} / \| b_{CR} \| = (0.9265, 0.3179, 0.2016)'$$

and then compute

$$\tilde{b}_{Time} - [\tilde{b}'_{Time} \tilde{b}_{CR}] \tilde{b}_{CR} = (-0.1500, 0.2458, 0.3020)'.$$

Thus, by lowering the initial density and increasing the other factors, we achieve the predicted *Time* and CR indicated in Table 11.10. Petersson et al. (1992) took a path similar to that in Table 11.10. Their results for both CR and *time* were disappointing, in that, contrary to the predicted values, actual CR declined and *time* dropped less than expected. At $D_0 = 0.311$, $t = 58^\circ$, and $D_g = 0.019$, CR = 0.255 and *time* = 10.7. By centering another 2^{3-1} at this treatment combination and then once again following a modified steepest descent search, slight improvements in each response were obtained [CR = 0.265 and *time* = 10.3 at $D_0 = 0.276$, $t = 62^\circ$, $D_g = 0.022$] through 11 additional runs.

Table 11.10. Steepest descent for time, modified to keep \widehat{CR} unchanged

Step	A ₂	B ₂	C ₂	D_0	t	D_g	\widehat{CR}	\widehat{Time}
0	0.00	0.000	0.000	0.344	52.00	0.0130	0.264	12.54
1	-0.15	0.246	0.302	0.340	52.74	0.0139	0.264	12.15
2	-0.30	0.492	0.604	0.335	53.47	0.0148	0.264	11.76
4	-0.60	0.983	1.208	0.326	54.95	0.0166	0.264	10.99
6	-0.90	1.475	1.812	0.317	56.42	0.0184	0.264	10.22
8	-1.20	1.966	2.416	0.308	57.90	0.0202	0.264	9.44

In all, Petersson et al. (1992) performed 27 runs, 3 different 2^{3-1} fractions, each with 1 centerpoint run, and 2 modified steepest descent searches. A preferred sequential strategy would change the sign of the generator for **C** in each fraction; that is, rather than using $\mathbf{C} = -\mathbf{AB}$ for all three fractions, use $\mathbf{C} = \mathbf{AB}$ for the second half-fraction. The benefit of this change in the defining relation is recognized when the data from two or more fractions are combined for analysis. Mee (2001) showed how two 2^{k-f} designs with different centers and aliasing can be used to estimate a second-order model. In Section 12.3 we discover strong interactions involving D_0 from a combined analysis of the Petersson et al. data.

11.3 A Group Screening Experiment

Rooda and van der Schilden (1982) presented an investigation of a complex simulation model for the transport of coal to 13 northwest European ports in sea-going barges. The article detailed the advantage of tug–barge combinations versus conventional ships and listed the simulation inputs for each port (including channel characteristics, number of terminal jetties, number of cranes, etc.) and a listing of the transport and ballast routes, with the tonnage for each year. A total of 29 factors were investigated: the number of extra barges and the loading/discharging capacities at each of the 13 ports plus 3 port independent factors; see Table 11.11 for details.

Table 11.11. Factor levels for Rooda and van der Schilden (1982) simulation experiment

Factors		Levels	
		-1	1
A	Required annual transport capacity (10^9 ton-km)	40.6	63
B	Number of Tugs	8	12
C	Capacity of a barge (10^3 DWT)	30	40
P_i	Port i Loading/discharging capacity (ton/hr.):		
	Ports 8 (Emden), 11 (Wilhelmshafen)	1200	1600
	Port 2 (Asnaes)	1300	1800
	Ports 3 (Esbjerg), 10 (Weserport)	1500	2000
	Ports 4 (Hunterston), 13 (Oxelösund)	2000	2400
	Port 1 (Zeebrugge)	2400	2800
	Port 9 (Hamburg)	3000	3500
	Ports 5 (Amsterdam), 12 (Landskrona)	3500	4000
	Port 7 (Narvik)	5000	6000
	Port 6 (Rotterdam)	7000	8000
Q_i	Extra barges in Port i:		
	Ports 1–5, 8, 10–13	0	2
	Port 9	0	3
	Ports 6, 7	0	5

These 29 factors were investigated using a 2^{8-4} design by grouping factors $\mathbf{P}_1-\mathbf{P}_{13}$ into three groups and $\mathbf{Q}_1-\mathbf{Q}_{13}$ into two groups. The five group factors were defined as follows:

- **D:** $\mathbf{P}_6 = \mathbf{P}_7 = \mathbf{P}_9$
- **E:** $\mathbf{P}_3 = \mathbf{P}_{11} = \mathbf{P}_{12}$
- **F:** $\mathbf{P}_1 = \mathbf{P}_2 = \mathbf{P}_4 = \mathbf{P}_5 = \mathbf{P}_8 = \mathbf{P}_9 = \mathbf{P}_{13}$
- **G:** $\mathbf{Q}_3 = \mathbf{Q}_6 = \mathbf{Q}_7 = \mathbf{Q}_8 = \mathbf{Q}_9 = \mathbf{Q}_{11} = \mathbf{Q}_{12}$
- **H:** $\mathbf{Q}_1 = \mathbf{Q}_2 = \mathbf{Q}_4 = \mathbf{Q}_5 = \mathbf{Q}_{10} = \mathbf{Q}_{13}$

Note that in a group screening experiment, subsets of factors are aliased, so that all factors in a group are simultaneously set at the high level or the low level. By assigning the “ -1 ” and “ $+1$ ” levels so that all factors within a group are expected to have effects of like sign, we hope to have the aliased active effects within a group sum rather than cancel one another. Although it is logical to group together the loading capacity factors, we are not told why each particular subset of ports was chosen. Presumably \mathbf{Q}_1 – \mathbf{Q}_{13} were put into just two groups because these were deemed the least likely variables to affect the response.

Table 11.12 contains the resolution IV screening design that was used in the first stage. Since this is a simulation experiment, run order does not need to be randomized. The mean Net Present Value (NPV) Cost for each of the 16 runs is reported, along with $y = \ln(\text{NPV Cost})$, which the authors chose as the response variable.

Table 11.12. First group screening experiment by Rooda and van der Schilden (1982)

A	B	C	D	E	F	G	H	NPV		
								Cost	y	s^2
1	-1	-1	-1	1	1	1	-1	1043	6.95	0.015
1	1	-1	-1	-1	-1	1	1	2208	7.70	0.034
1	-1	1	-1	-1	1	-1	1	1422	7.26	0.015
1	1	1	-1	1	-1	-1	-1	2644	7.88	0.034
1	-1	-1	1	1	-1	-1	1	424	6.05	0.500
1	1	-1	1	-1	1	-1	-1	437	6.08	0.015
1	-1	1	1	-1	-1	1	-1	659	6.49	0.180
1	1	1	1	1	1	1	1	365	5.90	0.086
-1	1	1	1	-1	-1	-1	1	1339	7.20	0.068
-1	-1	1	1	1	1	-1	-1	337	5.82	0.043
-1	1	-1	1	1	-1	1	-1	871	6.77	0.018
-1	-1	-1	1	-1	1	1	1	302	5.71	0.048
-1	1	1	-1	-1	1	1	-1	1882	7.54	0.003
-1	-1	1	-1	1	-1	1	1	2208	7.70	0.025
-1	1	-1	-1	1	1	-1	1	1353	7.21	0.007
-1	-1	-1	-1	-1	-1	-1	-1	2039	7.62	0.007

In addition, each simulation run was partitioned into five subruns, and the variance of y across these subruns is reported in the last column of Table 11.12. This lack of constant variance shows the atypical pattern of larger variance associated with smaller means; see Figure 11.3, where a logarithmic scale is used for the variance to display the summary statistics more effectively. Rooda and van der Schilden (1982) used generalized least squares to estimate the coefficients for an additive model. Section 14.4 will discuss this method,

which improves the efficiency of the estimates when there exists heterogeneity in the true error variance that can be well estimated.

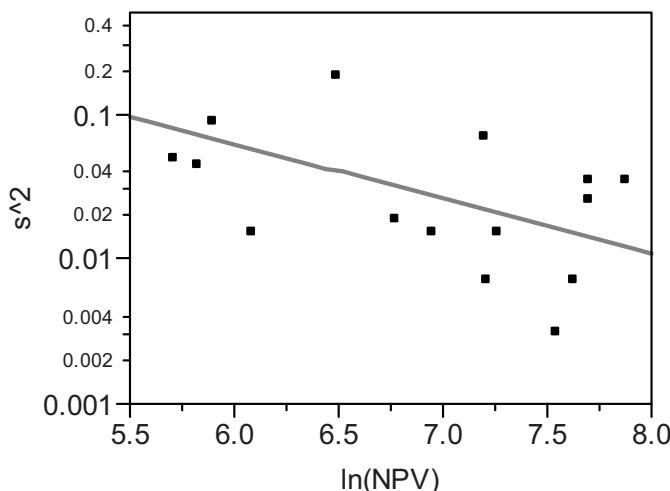


Fig. 11.3. Variance versus mean y for the first simulation experiment

Table 11.13. Estimates for Rooda and van der Schilden's (1982) first simulation experiment

Term	Estimate	Lenth t	p -Value
G	-0.615	-9.37	.0007
D	-0.309	-4.70	.0036
B	0.168	2.55	.0294
C	0.106	1.62	.1106
H	-0.082	-1.26	.1968
A	-0.079	-1.20	.2158
AD = ...	0.068	1.03	.2778
AB = ...	-0.066	-1.01	.2849
AG = ...	-0.044	-0.67	.5002
AF = ...	-0.035	-0.53	.6224
F	-0.026	-0.40	.7112
E	-0.023	-0.34	.7494
AC = ...	-0.013	-0.19	.8565
AH = ...	-0.011	-0.17	.8705
AE = ...	-0.006	-0.10	.9266

We analyze the data fitting a saturated model, in which case ordinary and generalized least squares estimates coincide. The results are in Table 11.13. According to Lenth's method, group factors **G** and **D** have the most impact on $\ln(\text{NPV})$, and number of tugs (**B**) is also statistically significant at $\alpha = .05$. All the interaction estimates are small. Rooda and van der Schilden (1982) used generalized least squares, utilizing the sample variances that appear in the last column of Table 11.12. However, according to Carroll and Cline (1988), these variances, with 4 df each, are not estimated with adequate precision to use generalized least squares. For further details, see Section 14.4.

The second stage for the group screening experiment ignored all factors corresponding to the insignificant estimates for **A**, **E**, **F**, and **H**. This leaves 12 factors to be estimated, since group factors **D** and **G** contain a total of 10 factors. A resolution III 2^{12-8} design was chosen, with generators $\mathbf{P}_9 = \mathbf{BC}$, $\mathbf{Q}_3 = \mathbf{P}_6\mathbf{P}_7$, $\mathbf{Q}_6 = \mathbf{BCP}_6$, $\mathbf{Q}_7 = \mathbf{BP}_6$, $\mathbf{Q}_8 = \mathbf{BCP}_7$, $\mathbf{Q}_9 = \mathbf{BP}_7$, $\mathbf{Q}_{11} = \mathbf{CP}_7$, and $\mathbf{Q}_{12} = \mathbf{BCP}_6\mathbf{P}_7$. The 16 treatment combinations and the simulation results are reported in Table 11.14.

Table 11.14. Second group screening experiment by Rooda and van der Schilden (1982)

B	C	P ₆	P ₇	P ₉	Q ₃	Q ₆	Q ₇	Q ₈	Q ₉	Q ₁₁	Q ₁₂	<i>y</i>	<i>s</i> ²
-1	1	-1	-1	-1	1	1	1	1	1	-1	-1	7.183	0.034
-1	1	1	-1	-1	-1	-1	-1	1	1	-1	1	7.169	0.046
-1	1	-1	1	-1	-1	1	1	-1	-1	1	1	7.255	0.056
-1	1	1	1	-1	1	-1	-1	-1	-1	1	-1	7.147	0.034
-1	-1	-1	-1	1	1	-1	1	-1	1	1	1	6.141	0.236
-1	-1	1	-1	1	-1	1	-1	-1	1	1	-1	6.855	0.032
-1	-1	-1	1	1	-1	-1	1	1	-1	-1	-1	7.211	0.069
-1	-1	1	1	1	1	1	-1	1	-1	-1	1	5.792	0.038
1	-1	-1	-1	1	1	-1	1	-1	1	-1	1	7.288	0.058
1	-1	1	-1	-1	-1	-1	1	1	-1	1	1	7.153	0.023
1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1	7.278	0.079
1	-1	1	1	-1	1	-1	1	-1	1	-1	-1	7.208	0.021
1	1	-1	-1	1	1	-1	-1	-1	-1	-1	1	7.063	0.070
1	1	1	-1	1	-1	1	1	-1	-1	-1	-1	7.389	0.004
1	1	-1	1	1	-1	-1	-1	1	1	1	-1	7.628	0.029
1	1	1	1	1	1	1	1	1	1	1	1	6.188	0.045

We analyze the second-stage data, just as we did the first experiment. Table 11.15 shows estimates for the saturated model, using Lenth's method to identify significant effects. The analysis shows that having two extra barges in Ports 3 and 12 is clearly beneficial. In addition, extra capacity in Ports 6 and 9 also lowers the cost, as does having eight tugs (**B** = -1) and smaller capacity barges (**C** = -1).

Noting that all the main effects for capacity (\mathbf{P}_j) and extra barges (\mathbf{Q}_j) have negative estimates, we suspect that the sparsity of effects assumption does not hold and that the PSE overestimates the true standard error. Rather than using Lenth's PSE = 0.0603, one can average the 16 variances in Table 11.14 to obtain a pooled variance of 0.0546. Since y in Table 11.14 is the mean of $n = 5$ runs, σ^2 for this response is estimated by $0.0546/5 = 0.0109$ and the corresponding standard error of the estimates in Table 11.15 is $(0.0109/16)^{1/2} = 0.026$. Thus, \mathbf{Q}_6 's estimate is also statistically significant.

Table 11.15. Estimates for Rooda and van der Schilden's (1982) second simulation experiment

Term	Estimate	PSE	Lenth t	p -Value
\mathbf{Q}_3	-0.245	0.0603	-4.07	.0062
\mathbf{Q}_{12}	-0.242	0.0603	-4.01	.0064
\mathbf{P}_9	-0.213	0.0603	-3.54	.0106
\mathbf{B}	0.153	0.0603	2.53	.0296
\mathbf{P}_6	-0.134	0.0603	-2.22	.0435
\mathbf{C}	0.131	0.0603	2.17	.0466
\mathbf{Q}_6	-0.093	0.0603	-1.55	.1211
\mathbf{Q}_8	-0.045	0.0603	-0.75	.4181
\mathbf{Q}_9	-0.040	0.0603	-0.67	.4735
\mathbf{Q}_{11}	-0.040	0.0603	-0.66	.5388
\mathbf{P}_7	-0.033	0.0603	-0.55	.6111
\mathbf{BQ}_3	0.033	0.0603	0.55	.6173
\mathbf{Q}_7	-0.031	0.0603	-0.51	.6396
\mathbf{BQ}_6	-0.020	0.0603	-0.34	.7532
\mathbf{BQ}_{12}	0.013	0.0603	0.22	.8411

11.4 Nonorthogonal Blocking for a Fractional Factorial

Broudiscou, Papon, and Broudiscou (2000) studied the effects of 13 plant extracts on fermentation, with the hope of identifying natural, affordable additives that would promote feed digestibility and so facilitate weight gain and milk production in livestock. They included a 14th factor as a control or null factor. The authors considered the trade-off between doing a large number of 1-day runs versus a small screening design of week-long runs. They concluded that short-term *in vitro* runs would not be sufficiently representative and so opted for week-long runs. Each run utilized a dual outflow fermenter, of which six were available for the experiment. For each week-long run, the first 5 days were for “adaptation”; data collected Days 6 and 7 were averaged into a single value. Broudiscou et al. blocked the design on weeks but thought

it safe to ignore possible differences between fermenters. Given the constraint of only six runs per week, a 2^{14-10} experiment with two centerpoint replicates was chosen, so that the run size was a multiple of 6. The full design required 3 weeks. If they had added a fourth week, they could have used orthogonal blocking, with four factorial runs per week, but this would have accommodated no additional factorial points. Instead, they adopted a nonorthogonal blocking scheme.

Table 11.16 lists the treatment combinations. For each factor, the low and high levels correspond to 0 and 0.5 g/day, respectively. Thus, except for the first run, which has no additives, every run has 3.5 or 4 g/day of additives. The reference factor was the additive *D. glomerata* (**D**), which was simply extracted from the basic dietary forage. The other 13 factors correspond to the 13 plant extracts being studied, whose effects may be either positive or negative. The 18 runs were assigned to weeks and fermenters as indicated in Table 11.17. Table 11.17 also lists the responses % acetate, % propionate, and % butyrate as a molar percentage of volatile fatty acids.

Table 11.16. Broudiscou et al.'s (2000) treatment combinations

Table 11.17. Broudiscou et al.'s (2000) blocking assignment for treatment combinations, with responses

Week	Fermenter	t.c.	Molar % of Volatile Fatty Acids		
			Acetate	Propionate	Butyrate
1	I	2	52.52	24.43	17.67
1	II	5	53.35	23.56	17.41
1	III	14	53.78	21.63	18.00
1	IV	12	53.85	21.67	17.54
1	V	9	49.79	27.23	17.58
1	VI	10	53.25	26.46	15.22
2	I	3	48.27	22.09	25.39
2	II	16	50.01	21.56	23.61
2	III	17	49.89	20.57	25.09
2	IV	18	49.93	23.49	22.56
2	V	8	47.98	21.39	25.69
2	VI	7	49.64	19.68	26.04
3	I	6	51.29	21.94	22.54
3	II	15	50.09	20.71	24.41
3	III	11	49.09	21.10	24.79
3	IV	4	50.78	20.79	23.75
3	V	13	49.02	22.52	24.04
3	VI	1	48.12	20.04	27.17

Prior to data collection, the six fermenters were thought to be identical, and the effect of week was expected to be negligible. If blocking is ignored, the analysis of these data is straightforward. The authors fitted a main effects model and computed *t* statistics based on 3 df for error (from 1 df for pure error and 2 df for lack-of-fit). For the three responses in Table 11.17, none of the 14 estimates were statistically significant at $\alpha = .05$; the smallest of the 42 *p*-values was .059. However, if one performs an ANOVA ignoring the factors and including only the unit structure (Weeks and Fermenters), Weeks is highly significant for each response, and Fermenters is not. A quick inspection of the data in Table 11.17 reveals that Week 1 was strikingly different than the other weeks, with low percentages for butyrate and high percentages for the other fatty acids. In private correspondence, one of the authors commented, "As a blocking factor, the reactor can be ignored but not the period. Unfortunately, at the time, we had not identified how significant the period variability was—probably linked to some characteristics of the rumen inoculum still unknown."

We now reanalyze the data, taking into account the period (i.e., Week) effect. Since the 14 factors utilize all but one of the df for the factorial, we first consider using the single contrast that distinguishes Week 1 from the other 2 weeks. For % butyrate, the main effects estimates, with and without the contrast for Weeks, are shown in Table 11.18. For the six factors that are

balanced in Week 1, with three levels high and three low, the effect contrasts are orthogonal to the Week contrast, and so these estimates are unchanged. The other eight factors have estimates that change, with factor **B** changing the most because its contrast is the most correlated with the contrast for Weeks. Thus, some unknown block effect provides an alternative explanation for the factor. Weeks accounts for most of the variation in butyrate %, and no other effect is statistically significant.

Table 11.18. Estimates for % Butyrate ignoring and accounting for the difference for Week 1

Term	(Ignoring Weeks) Estimate	(Including Weeks) Estimate	Difference
Intercept	22.14	22.14	0.00
Week contrast	.	-3.80	
A: <i>A.millefolium</i>	-1.43	-0.71	0.71
B: <i>A.chamissonis</i>	1.97	0.55	-1.43
C: <i>B.alba</i>	0.79	0.08	-0.71
D: <i>D.gloemerata</i>	-1.28	-0.57	0.71
E: <i>E.globulus</i>	-0.17	-0.17	0.00
F: <i>G.biloba</i>	-1.17	-0.46	0.71
G: <i>L.officinalis</i>	0.63	-0.08	-0.71
H: <i>L.capitata</i>	-0.24	-0.24	0.00
J: <i>H.perforatum</i>	0.04	0.04	0.00
K: <i>E.arvense</i>	-1.08	-0.36	0.71
L: <i>Propolis</i>	-0.20	-0.20	0.00
M: <i>vF.esculentum</i>	-0.13	-0.13	0.00
N: <i>S.officinalis</i>	-0.82	-0.11	0.71
O: <i>S.virgaurea</i>	-0.69	-0.69	0.00

What about the other two responses in Table 11.17? Here we consider stepwise regression as a tool for selecting a model, with Week included first as a nominal variable with three levels, followed by the largest factorial effects, after adjusting for Weeks. For acetate %, there is weak support for two active factors. Table 11.19 shows the fitted model. Given that we have selected the largest estimates and have based the standard error on pooling the rest into error, the *p*-values should be interpreted cautiously. Analysis of % Propionate is left to the reader.

Table 11.19. Forward selection model for % acetate

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	50.59	0.23	222.19	.000
Week[1]	1.76	0.34	5.10	.000
Week[2]	-1.13	0.33	-3.40	.005
Week[3]	-0.62	0.33	-1.88	.083
<i>A. millefolium</i>	0.71	0.25	2.80	.015
<i>G. biloba</i>	0.53	0.25	2.08	.058

Several articles present computational algorithms for optimally assigning treatment combinations to blocks for nonstandard block sizes. Trinca and Gilmour's (2000) algorithm takes a specified set of treatment combinations and finds the best blocking assignment. Cook and Nachtsheim's (1989) algorithm allows one to fix the set of treatment combinations or to search for the best treatment combinations and blocking simultaneously; see also the BLKL algorithm by Atkinson and Donev (1989).

The blocking assignment used by Broudiscou, Papon, and Broudiscou (2000) can be improved by assigning the 18 treatment combinations in Table 11.16 differently. One optimal assignment (that maximizes the determinant of $\mathbf{X}'\mathbf{X}$) is

- Block 1: Treatment combinations 2, 5, 8, 9, 12, 14
- Block 2: Treatment combinations 1, 3, 10, 15, 17, 18
- Block 3: Treatment combinations 4, 6, 7, 11, 13, 16

This assignment produces the following sums by block for each main effect contrast:

Block	A	B	C	D	E	F	G	H	J	K	L	M	N	O
1	0	0	-2	2	2	0	0	0	0	-2	-2	2	0	0
2	0	-2	0	-2	-2	0	-2	-2	0	2	2	-4	-2	0
3	0	2	2	0	0	0	2	2	0	0	0	2	2	0

Note that factors **A**, **F**, **J**, and **O** are orthogonal to blocks, in that their contrasts sum to zero for each block. Only one sum exceeds 2 in absolute value. The variance inflation factors for the 14 factors using this blocking assignment are, after sorting

$$1.0, 1.0, 1.0, 1.0, 1.111, 1.278, \dots, 1.278, 2.0.$$

By contrast, the assignment in Table 11.17 produced the following sums by blocks for each factor,

Block	A	B	C	D	E	F	G	H	J	K	L	M	N	O
1	2	-4	-2	2	0	2	-2	0	0	2	0	0	2	0
2	0	4	2	-2	0	-2	-2	-2	2	0	2	2	0	0
3	-2	0	0	0	0	0	4	2	-2	-2	-2	-2	-2	0

and larger VIFs for the factors

$$1.0, 1.0, 1.25, 1.25, 1.25, 1.313, \dots, 1.313, 1.813, 2.25.$$

Both designs place the two centerpoint replicates together in a block. The optimal design has four factors orthogonal to blocks (not just two), and it decreases the maximum variance inflation factor.

Part III

Additional Topics

Response Surface Methods and Second-Order Designs

Response surface methodology is a strategy for investigating, characterizing, or optimizing processes using empirical models. Often several experiments are performed sequentially. Estimation of a multifactor second-order polynomial model is one of the primary tools. This chapter provides a survey of the second-order model and its analysis, as well as the popular experimental designs used for estimation. The designs presented here are constructed by augmenting or adapting two-level factorial designs. The sections are as follows:

- Section 12.1. The Response Surface Methodology Strategy
- Section 12.2. Central Composite Designs
- Section 12.3. Other Composite Designs
- Section 12.4. Box–Behnken Designs
- Section 12.5. Analysis/Interpretation of the Fitted Second-Order Model

12.1 The Response Surface Methodology Strategy

The work of Box and Wilson (1951) is the seminal response surface methodology (RSM) article. The opening paragraph reads, “The work described is the result of a study extending over the past few years by a chemist and a statistician. Development has come about mainly in answer to problems of determining optimum conditions in chemical investigations, but we believe that the methods will be of value in other fields where experimentation is sequential and the error (variance) fairly small.” Their belief of usefulness was certainly warranted, as the subsequent RSM literature has grown to entail numerous books and hundreds of articles.

As with Box and Wilson’s investigations, RSM as espoused here, works best when the error variation is small. This permits one to use small experiments effectively. Furthermore, we will focus on the simplest case where all

the factors are quantitative. The sequence of questions to address is often the following:

1. Which of many possible factors affect the response?
2. Concerning the most influential factors, are we operating in a region containing the optimum response? If not, what direction should one investigate to search for improvement?
3. If we are operating in a region containing the optimum response, how can one construct a design to estimate an adequate model?
4. Given that we have fit a suitable model for the response(s) in a region containing an optimum, how can the optimum be identified?

Fractional factorial designs from Chapters 6 and 7 are particularly well suited to question 1. If the error variance is small, it is often advisable to choose a resolution III design with narrow spacing for the levels of each factor. In this case, a first-order model with linear effects is likely to explain most of the variation. In such cases, the initial experimental design is not expected to contain the optimum, but it can reliably point in a direction of improvement. Section 9.2 detailed how the fitted first-order model identifies the steepest ascent search direction for subsequent runs. Also, Example 11.2 showed how one may combine a sequence of small two-level designs, each followed by a steepest ascent search to progressively move closer to the region of the optimum.

What should be done when no clear direction of improvement is revealed by the two-level design? If this two-level design follows previous experiments with the same factors that showed active effects, then the conclusion is that the insignificant experiment is located in a region where the response surface is relatively flat—akin to a ridge or a trough rather than a steep hillside. The next section begins with such a situation. There we show how the two-level design can be augmented to fit a second-order polynomial model.

The first-order model was introduced in (1.2) and contains an intercept and linear main effect terms. The second-order polynomial multifactor model adds to this two types of second-order terms: linear-by-linear interactions as in the two-factor interaction model (1.3) and pure quadratic terms of the form $\beta_{i.i}x_i^2$. The second-order model is thus

$$y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{i=1}^k \beta_{i.i} x_i^2 + \sum_{i=1}^{k-1} \sum_{j=i+1}^k \beta_{i.j} x_i x_j + \epsilon. \quad (12.1)$$

As we know, any full factorial (or fractional factorial of resolution V or higher) permits estimation of the two-factor interaction model. However, a two-level design does not enable one to estimate any $\beta_{i.i}$ coefficients, since each x_i^2 is aliased with the intercept. Adding centerpoint runs, as discussed in Section 2.3, breaks the aliasing between the intercept and the pure quadratic coefficients, enabling one to estimate $\sum_{i=1}^k \beta_{i.i}$, but not the individual $\beta_{i.i}$. Sections 12.2–12.4 describe various designs suitable for estimating the second-order model, all of which incorporate two-level factorial designs in some form.

The second-order model (12.1) is very useful because it can take on many shapes and because these different shapes are easily characterized based on the second-order coefficients. To understand the shapes, we express the fitted second-order model as

$$\begin{aligned}\hat{y}(\mathbf{x}) &= b_0 + \sum_{i=1}^k b_i x_i + \sum_{i=1}^k b_{i,i} x_i^2 + \sum_{i=1}^{k-1} \sum_{j=i+1}^k b_{i,j} x_i x_j \\ &= b_0 + \mathbf{x}'\mathbf{b} + \mathbf{x}'\mathbf{B}\mathbf{x},\end{aligned}\quad (12.2)$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}, \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} b_{1,1} & 0.5b_{1,2} & \cdots & 0.5b_{1,k} \\ 0.5b_{2,1} & b_{2,2} & \cdots & 0.5b_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ 0.5b_{k,1} & 0.5b_{k,2} & \cdots & b_{k,k} \end{bmatrix}.$$

Assuming \mathbf{B} is nonsingular, the unique stationary point of the fitted surface occurs at

$$\mathbf{x}_S = -0.5\mathbf{B}^{-1}\mathbf{b}. \quad (12.3)$$

If all of the eigenvalues of \mathbf{B} are negative, then $\hat{y}(\mathbf{x}_S)$ is the global maximum. If all of the eigenvalues of \mathbf{B} are positive, then $\hat{y}(\mathbf{x}_S)$ is the global minimum. Finally, if \mathbf{B} has both positive and negative eigenvalues, then $\hat{y}(\mathbf{x}_S)$ is a saddle point, and the fitted surface is unbounded. In Section 12.5, several fitted models are examined, including cases where some eigenvalues lie much nearer zero than the rest. For the most thorough characterization of second-order models, see Box and Draper (2007).

12.2 Central Composite Designs

The most frequently employed second-order design is the central composite design, which is obtained by augmenting a two-level factorial design with centerpoints and axial points. For five or more factors, a fractional factorial may be used, provided it has resolution V or more. The following example illustrates the utility of these designs.

Box and Liu (1999) described a sequence of experiments involving paper helicopters, where the objective was to maximize flight time. The first experiment, a 2^{8-4} , was followed by a steepest ascent search in the direction of longer wings and a shorter, narrower body. A subsequent 2^4 led to a second steepest ascent search, again increasing the wing length while increasing the wing width and continuing to shorten the body. By this point, after 4 stages of experimentation and 42 treatment combinations, another 2^4 factorial was planned, using the factors and their ± 1 levels in Table 12.1. The treatment combinations are shown as Block 1 in Table 12.2. One copter was made corresponding to each treatment combination, and each was flown four times. The means and standard deviations are reported in Table 12.2. Note that these

Table 12.1. Factors and levels for Box and Liu central composite design

Factors	Levels				
	-2	-1	0	1	2
A Wing area (in ²)	11.20	11.80	12.40	13.00	13.60
B Wing length-to-width ratio	1.98	2.25	2.52	2.78	3.04
C Base width (in)	0.75	1.00	1.25	1.50	1.75
D Base length(in)	1.00	1.50	2.00	2.50	3.00

Table 12.2. Box and Liu (1999) central composite design for helicopter flight times

Block	A	B	C	D	Mean	Std Dev
1	-1	-1	-1	-1	3.67	0.052
1	1	-1	-1	-1	3.69	0.052
1	-1	1	-1	-1	3.74	0.055
1	1	1	-1	-1	3.70	0.062
1	-1	-1	1	-1	3.72	0.052
1	1	-1	1	-1	3.55	0.065
1	-1	1	1	-1	3.97	0.052
1	1	1	1	-1	3.77	0.098
1	-1	-1	-1	1	3.50	0.079
1	1	-1	-1	1	3.73	0.072
1	-1	1	-1	1	3.58	0.083
1	1	1	-1	1	3.63	0.132
1	-1	-1	1	1	3.44	0.058
1	1	-1	1	1	3.55	0.049
1	-1	1	1	1	3.70	0.081
1	1	1	1	1	3.62	0.051
1	0	0	0	0	3.77	0.032
1	0	0	0	0	3.75	0.055
2	-2	0	0	0	3.61	0.129
2	2	0	0	0	3.64	0.085
2	0	-2	0	0	3.55	0.100
2	0	2	0	0	3.73	0.063
2	0	0	-2	0	3.61	0.051
2	0	0	2	0	3.60	0.095
2	0	0	0	-2	3.80	0.049
2	0	0	0	2	3.60	0.055
2	0	0	0	0	3.70	0.072
2	0	0	0	0	3.68	0.055
2	0	0	0	0	3.69	0.078
2	0	0	0	0	3.66	0.058

standard deviations capture only flight-to-flight variation of the same copter and do not provide any indication about the reproducibility of the copter manufacturing step. With an overall average flight time of 3.7 sec for an 8-ft drop, the copters are falling at an average rate near 2.2 ft/sec.

Initially, only Block 1 was run. Table 12.3 shows the results of fitting a two-factor interaction model, plus a quadratic term to account for any pure quadratic curvature, to these 18 means. All higher-order interactions are assumed negligible, and this model has $R^2 = .976$. Two factors have significant main effects, meaning that further improvement in flight time is possible. Steepest ascent could be pursued; however, the first-order model has a low R^2 , and so Box and Liu (1999) chose to augment these 18 runs with a second, 12-run block (see Table 12.2).

The second block in Table 12.2 uses a wider spacing of levels for each factor, but only one factor at a time is moved from the center level. These are the axial points of a central composite design. The spacing for the axial points is determined by a parameter α . In Table 12.2, $\alpha = 2$, which means that the levels for the axial points are twice the spread of the factorial points for each factor. Note that when $\alpha = k^{1/2}$, as is the case here, the factorial and axial points all fall on the same hypersphere. As shown by Kiefer (1960), this is optimal for fitting second-order models in spherical regions; that is, given a spherical experimental region, it is best to place the design points on the boundary of the spherical region and at the center.

Note that the second block enables one to estimate the linear and quadratic coefficient for each factor, since each factor has three levels. Thus, from the combination of a 2^k , centerpoint replicates, and axial points, one can estimate the full second-order model. Estimates and t statistics for this fitted second-order model are given in Table 12.4.

Table 12.3. Fitted model for mean flight time in Table 12.2, Block 1 only

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	3.760	0.0223	168.86	.0000
A	-0.005	0.0079	-0.64	.5488
B	0.054	0.0079	6.83	.0005
C	0.005	0.0079	0.64	.5488
D	-0.066	0.0079	-8.42	.0002
AB	-0.029	0.0079	-3.65	.0107
AC	-0.038	0.0079	-4.76	.0031
BC	0.046	0.0079	5.87	.0011
AD	0.044	0.0079	5.56	.0014
BD	-0.015	0.0079	-1.91	.1054
CD	-0.021	0.0079	-2.70	.0356
A² + B² + C² + D²	-0.100	0.0236	-4.23	0.0055

Table 12.4. Fitted second-order model for mean flight time

Term	Estimate	Std Error	t-Ratio	p-Value
Intercept	3.758	0.0238	157.76	.0000
Block	-0.030	0.0121	-2.44	.0285
A	-0.001	0.0064	-0.13	.8977
B	0.051	0.0064	7.99	.0000
C	0.002	0.0064	0.39	.7004
D	-0.061	0.0064	-9.56	.0000
AB	-0.029	0.0078	-3.69	.0024
AC	-0.038	0.0078	-4.81	.0003
BC	0.046	0.0078	5.93	.0000
AD	0.044	0.0078	5.61	.0001
BD	-0.015	0.0078	-1.92	.0749
CD	-0.021	0.0078	-2.73	.0164
A²	-0.020	0.0060	-3.37	.0045
B²	-0.017	0.0060	-2.75	.0156
C²	-0.025	0.0060	-4.20	.0009
D²	-0.002	0.0060	-0.27	.7918

First note that the estimates for two-factor interactions are unchanged from Table 12.3 to Table 12.4. Axial and centerpoint replicates furnish no information about interactions, since the product of any two columns in Block 2 is zero. The slight change in the standard errors is simply due to a change in the estimate for σ ; in both Table 12.3 and 12.4, the estimated standard error for each $b_{i,j}$ is $\text{RMSE}/16^{1/2}$.

Estimates for the linear terms come from both blocks. For example, the estimate for the last linear term from block 1 is -0.066, based on the 16 factorial points; the corresponding estimate from block 2 is $(3.60 - 3.80)/4 = -0.050$. The combined estimator is a weighted average of these two results and has estimated standard error $\text{RMSE}/24^{1/2}$. The estimates for the pure quadratic coefficients comes almost entirely from the second block. Note that every pure quadratic term is estimated with good precision, which is due to having a sufficient number of centerpoint replicates and choosing $\alpha = 2$. Using a smaller α or fewer centerpoint replicates would produce a larger standard error for the $b_{i,i}$'s. We will interpret this fitted second-order model in Section 12.5. Here, we simply note that all four pure quadratic coefficients are negative, which is a necessary condition for the fitted model having a unique maximum. However, the quadratic coefficient for **D** is weak; we will see that this causes the fitted model to be unbounded, having neither a global maximum or minimum (refer to Section 12.5).

This is just one example of a central composite design. Common choices for various other cases are summarized in Table 12.5. For five or six factors, a half-fraction permits estimation of all two-factor interactions. In each case, $\alpha =$

$k^{1/2}$ is the optimal choice for spherical experiment regions. For some cases, a slightly smaller value [$(\text{No. of factorial points})^{1/4}$] is needed to achieve second-order rotatability; for a description of this property, see Box and Draper (2007, p. 471). Any α in the ranges listed in Table 12.5 produces a reasonable design. In Table 12.5, $n_0 = 3$ centerpoint replicates are recommended in each case. This is the integrated variance (IV) optimal n_0 for spherical regions except for $k = 4$, where $n_0 = 4$ is IV optimal. If the central composite design is run in stages, with the axial runs in a separate block, it is important to include centerpoint runs in the axial block, as this benefits the precision of the pure quadratic coefficients. For $k = 3, 4$, and 6 , the factorial designs in Table 12.5 can also be run in two blocks, if necessary, since this can be done by confounding three-factor or higher-order interactions.

Table 12.5. Common central composite designs

No. Factors	Factorial Portion	Spacing for Axial Points	Total No. Runs (with $n_0 = 3$)
2	2^2	$\alpha = 2^{1/2}$	$4 + 4 + 3 = 11$
3	2^3	$1.68 \leq \alpha \leq 3^{1/2}$	$8 + 6 + 3 = 17$
4	2^4	$\alpha = 2$	$16 + 8 + 3 = 27$
5	2^{5-1}	$2 \leq \alpha \leq 5^{1/2}$	$16 + 10 + 3 = 29$
6	2^{6-1}	$2.38 \leq \alpha \leq 6^{1/2}$	$32 + 12 + 3 = 47$

12.3 Other Composite Designs

This section gives an overview of two other lesser known composite designs used for estimating second-order models: noncentral composite designs (Mee 2001) and asymmetric composite designs (Lucas 1974, Mee 2001). Similar to central composite designs, these alternative designs are constructed by augmenting a two-level factorial design in order to estimate a second-order model. The central composite design augments the 2^k with an axial block centered about the coded treatment combination $(0, 0, \dots, 0)$. The noncentral and asymmetric composite designs augment with treatment combinations directed toward the best performing treatment combination in the original 2^k . The advantage of these strategies is that they attempt to provide experimental runs close to the perceived optimum, ensuring a better fit for one's model in the region of greatest interest. The noncentral composite design is simplest to describe. We use an example from Mee (2001) to illustrate this design approach. Subsequently, we discuss asymmetric composite designs.

The central composite design discussed in the previous section was preceded by four stages of experimentation by Box and Liu (1999). Prior to obtaining the data in Table 12.2, a fractional factorial and a 2^4 full factorial

were conducted, each followed by a steepest ascent search. Such experimentation does not fully utilize the information from previous experiments. The benefit of a noncentral composite design is that it experiments in the direction of improvement in a manner that immediately supports estimation of a second-order model. This usefulness is demonstrated by a sequence of 2^3 copter experiments in Mee (2001).

Motivated by Box and Liu's (1999) 2^4 (that preceded their central composite design), which showed main effects for base length and wing area, the 2^3 experiment with $n_0 = 3$ centerpoint replicates shown in Table 12.6 was performed. The three factors were L = Wing length, W = Wing width, and R = (Wing length)/(Base length). The ratio of lengths was selected as a factor, rather than Base length, because interactions between R and the other factors were expected to be negligible.

Table 12.6. Initial block of noncentral composite design

Copter	Order	Run			Mean	Std. Dev.
		L	W	R		
1	2	4	2.0	2.0	3.32	0.131
2	9	6	2.0	2.0	3.14	0.306
3	8	4	3.0	2.0	2.85	0.046
4	4	6	3.0	2.0	3.25	0.132
5	10	4	2.0	3.0	3.15	0.344
6	5	6	2.0	3.0	4.03	0.021
7	3	4	3.0	3.0	2.97	0.186
8	7	6	3.0	3.0	3.44	0.042
9	1	5	2.5	2.5	3.48	0.065
10	6	5	2.5	2.5	3.29	0.137
11	11	5	2.5	2.5	3.55	0.119

Fitting a saturated model, the largest t statistic is 4.14 for Wing length. However, the p -value exceeds .05 even for this estimate, because we only have 2 df for estimating the error variance, based on replication. Because we do not expect effect sparsity to hold, Lenth's method is not appropriate. Furthermore, only one copter (Copter 6) performed better than the average time at the center. Thus, we opt for a second 2^3 experiment, with the same factors and spacing as before but centered at the treatment combination for Copter 6. The design points and resulting times are reported in Table 12.7.

Table 12.7. Second block of noncentral composite design

Copter	Order	Run			Mean	Std. Dev.
		L	W	R		
12	21	5	1.5	2.5	3.18	0.182
13	13	7	1.5	2.5	2.10	0.283
14	18	5	2.5	2.5	3.45	0.125
15	16	7	2.5	2.5	2.63	0.038
16	20	5	1.5	3.5	4.09	0.055
17	15	7	1.5	3.5	2.17	0.267
18	19	5	2.5	3.5	3.50	0.237
19	14	7	2.5	3.5	3.84	0.123
20	17	6	2.0	3.0	3.83	0.101
21	22	6	2.0	3.0	3.36	0.053
22	12	6	2.0	3.0	3.43	0.196

The data from Tables 12.6 and 12.7 together support estimating a second-order model, with a block main effect. Fitting a model to the factors in natural units, the results are given in Figure 12.1. The fitted model is statistically significant (p -value = .024), whereas the lack-of-fit is not (p -value = .106). The systematic variation is attributable to three terms: a linear effect for the ratio R ($b_R = 0.409$), curvature for Wing length ($b_{L \cdot L} = -0.264$), and a Wing length-by-Width interaction ($b_{L \cdot W} = 0.347$). In Section 12.5, we will explore further the shape of this fitted model and find a predicted optimum time of approximately 4 sec for the treatment combination R = 3.5, L = 5.6 in., and W = 2.4 in., which is outside the initial factorial design region but inside the second.

Noncentral composite designs can also be constructed by combining fractional factorial designs with different centers. For instance, a fourth factor could have been incorporated into the 11-run experiments in Tables 12.6 and 12.7. Provided the second experiment is a foldover of the initial 2^{4-1} , the design supports estimation of the second-order model plus a block effect. In general, the two fractional factorials must, together, form a design with resolution V or higher. For more about these useful, underutilized designs, see Mee (2001, Sect. 1).

Box and Wilson (1951) also mentioned an alternative to the central composite design for cases where an optimum appeared to be in the vicinity of one of the factorial points. Lucas (1974) elaborated on this suggestion and named the design an *asymmetric composite design*. The 2^4 isatin yield experiment (see Tables 2.5 and 2.6) discussed previously presents just the situation in which an asymmetric composite design is appropriate. The initial 2^4 demonstrated that yield is improved at low Acid strength, low Reaction time, and high Reaction temperature. Due to the apparent importance of two-factor interactions, plus the likelihood of pure quadratic terms, one would prefer a follow-up design that permits estimation of a second-order model. Furthermore, given

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	10	4.1772374	0.417724	3.5874
Error	11	1.2808737	0.116443	Prob > F
C. Total	21	5.4581111		0.0235

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	7	1.1148663	0.159267	3.8376
Pure Error	4	0.1660074	0.041502	Prob > F
Total Error	11	1.2808737		0.1058

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	2.932	0.778	3.77	0.0031
Block	0.059	0.104	0.57	0.5810
R	0.409	0.171	2.40	0.0355
L	-0.119	0.085	-1.39	0.1907
W	0.094	0.171	0.55	0.5937
(R-2.75)*(R-2.75)	0.157	0.356	0.44	0.6664
(R-2.75)*(L-5.5)	0.167	0.162	1.03	0.3244
(L-5.5)*(L-5.5)	-0.264	0.089	-2.97	0.0127
(R-2.75)*(W-2.25)	-0.008	0.323	-0.03	0.9799
(L-5.5)*(W-2.25)	0.347	0.162	2.15	0.0546
(W-2.25)*(W-2.25)	-0.383	0.356	-1.08	0.3042

Fig. 12.1. Second-order model for copter data

the preference for the $(x_1, x_2, x_4) = (-1, -1, 1)$ treatment combination, it makes sense to experiment further in that direction. For this application, Lucas's asymmetric composite design would require augmentation with the three axial treatment combinations $(-1 - \alpha, -1, 1)$, $(-1, -1 - \alpha, 1)$, and $(-1, -1, 1 + \alpha)$. Lucas (1974) found the D-optimal choice for α in the context of a hypercube design region and a model with no block effect. However, the resulting optimal $\alpha = 2/k$ is too small to provide adequate precision for the pure quadratic terms.

Mee (2001, Sect. 3.1) generalized Lucas's asymmetric composite designs by adding a base point in addition to the k asymmetric axial points, so that a block effect can be included along with the terms of a second-order model. The base point can be either the attractive factorial treatment combination [here $(-1, -1, 1)$], or it can be nearer to the original design center, to make the augmenting block an orthogonal first-order design. We illustrate for the isatin experiment.

If we take $\alpha = 1.5$, the axial treatment combinations are $(-2.5, -1, 1)$, $(-1, -2.5, 1)$, and $(-1, -1, 2.5)$. Note that $\alpha < 2$ places the axial points nearer the attractive factorial treatment combination than any of the other original factorial points; this is generally desired. Mee's (2001) augmenting

design contains a base point in addition to the three axial points. The best option is to choose a base point between the attractive factorial treatment combination and the original design center. With $\alpha = 1.5$, the base point $(-0.5, -0.5, 0.5)$ creates a four-run orthogonal design here. Figure 12.2 shows this asymmetric composite design. The corners of the cube represent the initial block; the pyramid-shaped design, with treatment combinations denoted by circles, is the augmenting block that permits estimation of a second-order model. For more details, see Mee (2001, Sect. 3.1).

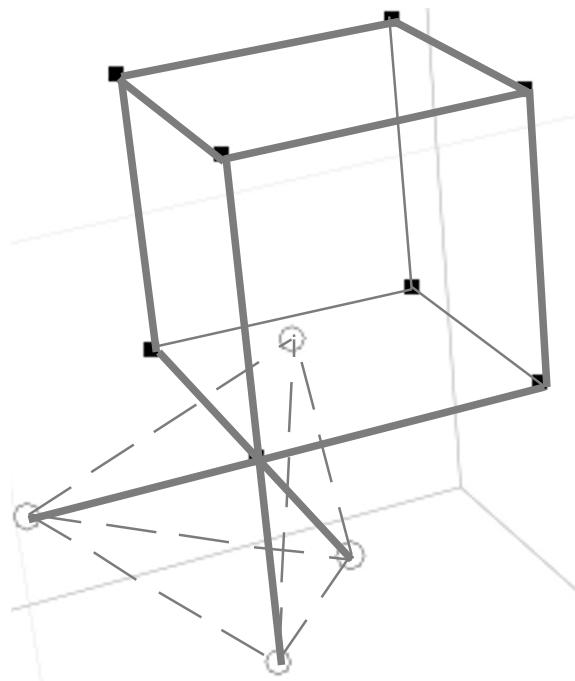


Fig. 12.2. Asymmetric composite design for three factors

12.4 Box–Behnken Designs

The most popular three-factor design for estimating a second-order model is due to Box and Behnken (1960). It consists of the 12 midpoints of the edges of a cube, plus typically 3 centerpoint runs. This design requires two fewer runs than a central composite design (with the same number of centerpoint replicates). Box–Behnken designs are also smaller than central composite designs for seven and nine factors, whereas Box–Behnken designs require more

runs for five, six, or eight factors. However, one consistent advantage of this and other Box–Behnken designs is that they require only three levels for each factor, whereas central composite designs with $\alpha \neq 1$ require five levels. Each Box–Behnken design is a collection of two-level designs in a subset of the factors. For instance, the design in Figure 12.3 is composed of three 2^2 factorial designs, each with a different factor held fixed at the center; see Table 12.8. Note that the runs should be performed in random order. The order in Table 12.8 simply emphasizes the structure of the design.

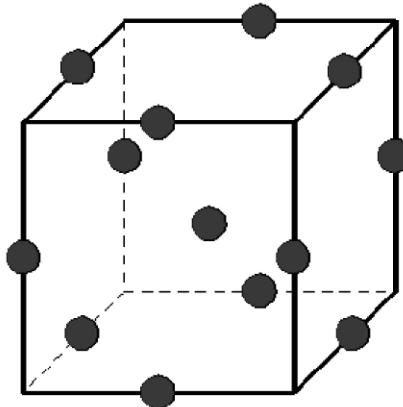


Fig. 12.3. Box–Behnken design for three factors

Table 12.8. Three-factor Box–Behnken design

x_1	x_2	x_3
-1	-1	0
1	-1	0
-1	1	0
1	1	0
0	0	0
-1	0	-1
1	0	-1
-1	0	1
1	0	1
0	0	0
0	-1	-1
0	1	-1
0	-1	1
0	1	1
0	0	0

Of the 27 points of a full 3^3 factorial, the Box–Behnken design consists of centerpoint replicates plus the 12 treatment combinations a distance $2^{1/2}$ from the center. The Box–Behnken designs for four and five factors are similar. For 4 factors, there are six pairs of factors and so $6(2^2) = 24$ treatment combinations a distance $2^{1/2}$ from the center. For 5 factors, there are 10 pairs of factors and $10(2^2) = 40$ such treatment combinations. Typically, the four- and five-factor Box–Behnken designs incorporate three to four centerpoint replicates.

From the 2^2 portion involving the pair of factors i and j , we compute estimates for the linear coefficients, β_i and β_j , and the two-factor interaction, $\beta_{i,j}$. In addition, the difference between the mean for this 2^2 portion and the centerpoint replicates estimates $\beta_{i,i} + \beta_{j,j}$. Combining this with pure quadratic curvature estimates for other pairs of factors enables the estimation of each individual $\beta_{i,i}$.

For six or more factors, Box–Behnken designs are no longer based on 2^2 designs for every pair of factors. For six to nine factors, they are based on combining 2^3 designs for various subsets of three factors. For details, see Mee (2007, Table 1). The Box–Behnken design for four factors can be run in three blocks, whereas the designs for five to nine factors can each be run in two blocks [see, e.g., Box and Behnken (1960) or Dean and Voss (1999, p. 578)].

12.5 Analysis/Interpretation of the Fitted Second-Order Model

Consider again the fitted second-order model displayed in Figure 12.1. The matrix of second-order coefficients \mathbf{B} defined below (12.2) is, for this model,

$$\mathbf{B} = \begin{bmatrix} 0.157 & 0.083 & -0.004 \\ 0.083 & -0.264 & 0.173 \\ -0.004 & 0.173 & -0.383 \end{bmatrix}.$$

Because the pure quadratic coefficients on the main diagonal of \mathbf{B} are both positive and negative, this is necessarily an indefinite matrix and the fitted surface is a saddle surface.

Most common statistical analysis software will solve for the stationary point (12.3) and perform an eigenanalysis (canonical analysis) of \mathbf{B} . Figure 12.4 shows that analysis for this second-order model fitted to the variables in natural units. First, the second-order and first-order coefficients are displayed. Note that the off-diagonal elements of \mathbf{B} are obtained by dividing interaction coefficients such as 0.1667 in half. Next, in Figure 12.4, the location of the stationary point ($R = 1.78$, $L = 4.87$, $W = 2.10$) is given, together with $\hat{y}(\mathbf{x}_S) = 3.44$ sec. This stationary point is not of particular use, however, since, at \mathbf{x}_S , the predicted flight time is unimpressive. For a saddle surface, the stationary point represents a flat region in the predicted surface. That this fitted model is a saddle surface is a consequence of \mathbf{B} having both positive

(0.175) and negative (-0.1536 and -0.5116) eigenvalues. The final information in Figure 12.4 is the eigenvector associated with each of the three eigenvalues. Negative eigenvalues define maxima (or ridges) in the surface, whereas positive values correspond to minima (or troughs). Eigenvalues close to zero indicate directions with little or no curvature.

Figure 12.5 shows the fitted model at two locations for (R, L, W) . The first is the stationary point $(1.78, 4.87, 2.10)$. From this plot it is evident that by increasing the ratio R we will improve predicted flight time. The plot also shows that changing only L or only W would result in lower predicted flight times than the 3.44-sec time at the stationary point.

That increasing R would result in better times is confirmed by the eigenanalysis of \mathbf{B} . For any larger-the-better response, we consider directions associated with positive eigenvalues. For this fitted model the only positive eigenvalue is 0.175, and the corresponding eigenvector is dominated by R , the ratio of wing length to base length. The stationary point is inside the region of experimentation except for R . By increasing R , predicted flight times will improve and we will move toward the design region. Since the coefficients for R , L , and W in the eigenvector $(0.98, 0.21, 0.06)$ are the same sign, as R is increased, L and W should also be increased slightly to stay on top of this rising ridge. The second part of Figure 12.5 shows the fitted model at the maximum predicted value inside the range of values for each factor; that is, the coordinates for L and W correspond to a maximum, given that R is set at the highest level from the experiment. The predicted flight time is 4.01 sec. Any larger predicted values would require extrapolation in R .

Contour plots are also useful for understanding a fitted model such as this, but they are limited to viewing two dimensions at a time. The eigenanalysis of \mathbf{B} is not limited in this way. Recall that the only important interaction in this fitted model is the L^*W interaction. Because of this interaction, one should view a wing length by wing width contour plot. Figure 12.6 shows such a plot for $R = 3.5$, the preferred value according to Figure 12.5. The ideal (L, W) combination appears in the ellipse centered at $(5.62, 2.42)$. To increase one wing dimension and decrease the other would produce a rapid decline in the predicted times. That this would be the case is evident in the eigenanalysis shown in Figure 12.4, where the most negative eigenvector has coefficients of opposite sign for L and W .

Note that we used natural units to fit the model and display the graphs for this analysis. Doing so facilitates the interpretability of profiler plots and contour plots such as Figures 12.5 and 12.6. It is important to recognize, however, that the choice of units affects the magnitudes of the eigenvalues and the coordinates of the eigenvectors of \mathbf{B} . It is generally advisable to consider first an eigenanalysis for coefficients from a model with coded factors. The analysis just completed involved few factors and so was easily understood without such coding. We now consider the second-order model from Box and Liu (1999), where we begin with a model fit to coded factors.

Response Surface				
Coef	R	L	W	Y
R	0.1575	0.1667	-0.0083	0.4088
L	.	-0.2644	0.3475	-0.1190
W	.	.	-0.3833	0.0938

Solution	
Variable	Critical Value
R	1.7820
L	4.8693
W	2.0969

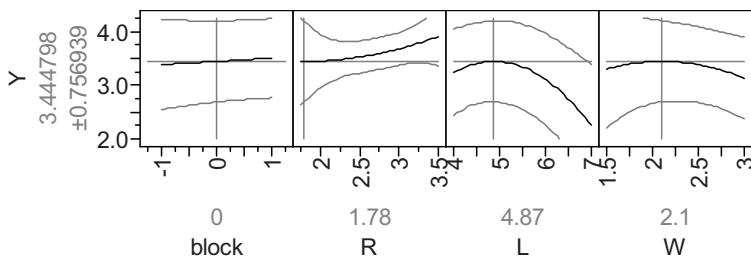
Solution is a SaddlePoint
 Critical values outside data range
 Predicted Value at Solution 3.4448

Canonical Curvature

Eigenvalue	Eigenvalues and Eigenvectors		
	0.1750	-0.1536	-0.5116
R	0.97646	-0.20086	0.07859
L	0.20790	0.77955	-0.59083
W	0.05741	0.59326	0.80296

Fig. 12.4. Canonical analysis of three-factor, second-order model from Figure 12.1

Prediction Profiler at Stationary Point



Prediction Profiler at Maximum Y within the Experimental Region

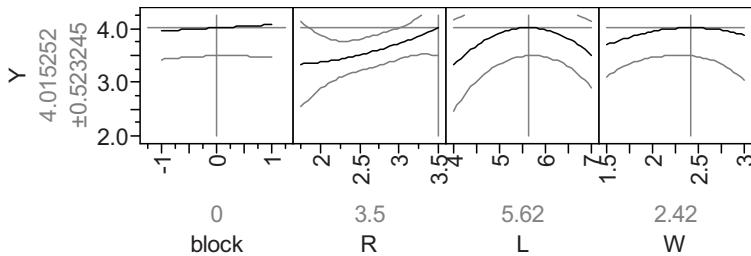


Fig. 12.5. Fitted second-order model at stationary point and at maximum within the experimental region

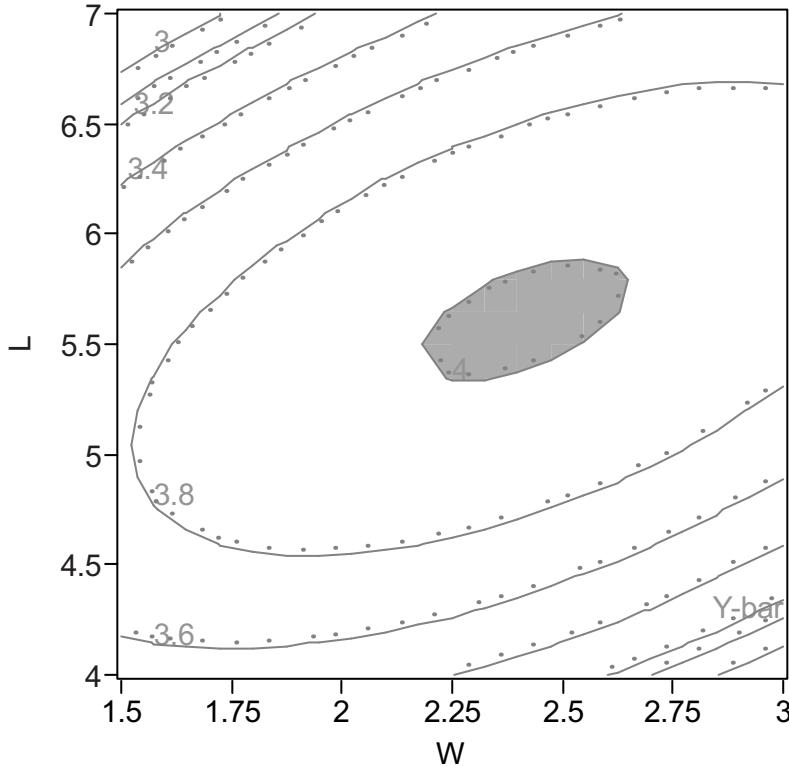


Fig. 12.6. Contour plot for L and W; R = 3.5 in

Table 12.4 shows a fitted second-order model with four factors. Here, all of the pure quadratic coefficients are negative. This is a necessary condition for the stationary point to be a maximum. However, that is not a sufficient condition, as Figure 12.7 reveals. Box and Liu's (1999) model is also a saddle surface, because the eigenvalues are both positive and negative. Since the stationary point is well inside the experimental region, one may search for improved responses in two directions: $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) = \pm(0.52, -0.45, -0.45, 0.57)$. Since moving in the direction $-(0.52, -0.45, -0.45, 0.57)$ from the stationary point initially moves us nearer the center than does the opposite direction, larger predicted values are obtained in this direction for a given distance from the design center. At $\mathbf{x}_S - 3(0.52, -0.45, -0.45, 0.57) = (-0.69, 1.02, 0.52, -1.83)$, the predicted value reaches 4 sec; see Figure 12.8. This point is inside the range of ± 2 for each factor but is slightly outside the sphere of radius 2 that contains the factorial and axial points of the central composite design.

Box and Liu (1999), using a method of steepest ascent for second-order models, found their best copter performance with a time of 4.16 sec slightly

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	15	0.292335	0.019489	20.0401
Error	14	0.013615	0.000972	Prob > F
C. Total	29	0.305950		<.0001

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	10	0.012540	0.001254	4.6660
Pure Error	4	0.001075	0.000269	Prob > F
Total Error	14	0.013615		0.0755

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	3.7575	0.0238	157.76	<.0001
Block	-0.0295	0.0121	-2.44	0.0285
A	-0.0008	0.0064	-0.13	0.8977
B	0.0508	0.0064	7.99	<.0001
C	0.0025	0.0064	0.39	0.7004
D	-0.0608	0.0064	-9.56	<.0001
A*A	-0.0204	0.0060	-3.37	0.0045
A*B	-0.0288	0.0078	-3.69	0.0024
B*B	-0.0166	0.0060	-2.75	0.0156
A*C	-0.0375	0.0078	-4.81	0.0003
B*C	0.0462	0.0078	5.93	<.0001
C*C	-0.0254	0.0060	-4.20	0.0009
A*D	0.0438	0.0078	5.61	<.0001
B*D	-0.0150	0.0078	-1.92	0.0749
C*D	-0.0213	0.0078	-2.73	0.0164
D*D	-0.0016	0.0060	-0.27	0.7918

Response Surface

Coef	A	B	C	D	Mean
A	-0.0204	-0.0288	-0.0375	0.0438	-0.0008
B	.	-0.0166	0.0462	-0.0150	0.0508
C	.	.	-0.0254	-0.0213	0.0025
D	.	.	.	-0.0016	-0.0608

Solution

Variable	Critical Value
A	0.861
B	-0.331
C	-0.839
D	-0.116

Solution is a SaddlePoint

Predicted Value at Solution 3.71

Canonical Curvature (Eigenvalues and Eigenvectors)

Eigenvalue	0.0326	-0.0120	-0.0381	-0.0465
A	0.518	0.041	0.761	0.389
B	-0.450	0.582	0.506	-0.451
C	-0.452	0.376	-0.122	0.800
D	0.570	0.720	-0.388	-0.076

Fig. 12.7. Eigenanalysis for Box and Liu's fitted second-order model

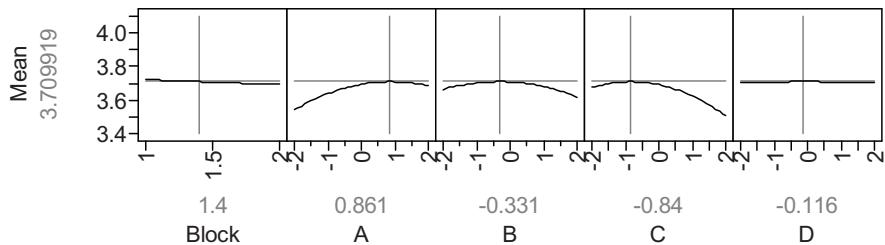
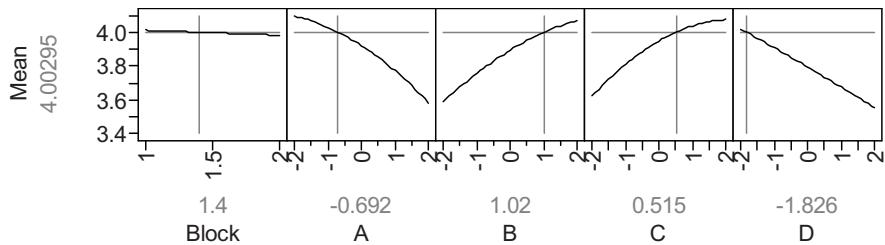
Prediction Profiler at Stationary Point for Box and Liu's second-order model**Prediction Profiler at one point along rising ridge**

Fig. 12.8. Box and Liu's fitted second-order model at stationary point and at point along a ridge

further out in this direction. The method of exploration shown here is simpler and is adequate when only one eigenvalue is positive. For the more sophisticated method, see Box and Draper (2007, Ch. 12).

13

Special Topics Regarding the Design

This chapter addresses many practical issues that were skipped earlier. We discuss the choice of levels for quantitative factors and the choice of design size. Several specialized requirements influencing the choice of run order and treatment combinations are addressed briefly. The sections are as follows:

Section 13.1. Power and the Choice of Sample Size

Section 13.2. Choice of Factor Levels

Section 13.3. Tips for Studying Variation

Section 13.4. Accommodating Factors with More Levels

Section 13.5. Special Requirements for Run Order and Treatment Combinations

13.1 Power and the Choice of Sample Size

13.1.1 Continuous responses

Sample size determination is often posed in the context of achieving sufficient power for hypothesis testing. Such calculations involve the relationship among the following three quantities:

- The probability of a Type I error, α . This probability is also named the “significance level of the test.”
- The true magnitude of the effect being tested. The larger the magnitude of a true effect, the greater the likelihood of rejecting the null hypothesis and declaring the effect active. Generally, the size of the effect is specified as the regression coefficient divided by the error standard deviation, σ .
- The power of the test (i.e., the desired probability of concluding that the effect is active, given its true magnitude). Note that power = $1 - \beta$, where β is the probability of a Type II error.

Two types of t -tests have been used throughout this book: ordinary t statistics, where a mean square error is used in the calculation of the standard error, and Lenth t statistics, which are calculated using the PSE. Under the alternative hypothesis that $\beta_i \neq 0$, ordinary t statistics follow a noncentral t distribution, and most statistical packages calculate probabilities under this distribution. The noncentral t distribution $t_\nu(\lambda)$ is characterized by its degrees of freedom ν and its noncentrality parameter λ . For any equally replicated, two-level full factorial design, the noncentrality parameter for the i^{th} regression coefficient is

$$\lambda = N^{1/2}(\beta_i/\sigma),$$

where N is the number of factorial points.

To illustrate the calculations, consider the replicated 2^3 factorial in Table 1.3. Using the pure error mean square with 8 df as the estimator for σ^2 , the critical value from Appendix A for two-sided t -tests at $\alpha = .05$ is 2.306. Thus,

$$\text{Power} = P[|t_8(4\beta_i/\sigma)| > 2.306].$$

For instance, when $\beta/\sigma = 1$,

$$\begin{aligned} \text{Power} &= P[t_8(4) < -2.306] + P[t_8(4) > 2.306] \\ &= 0.000 + 0.937 = 0.937. \end{aligned} \tag{13.1}$$

Software facilitates this computation. In JMP, one begins by fitting a model to (artificial or actual) data. Here, we fit a saturated model for a replicated 2^3 ; this step fixes the model degrees of freedom for the subsequent calculation. Next, one requests the calculation of power for a particular effect and specifies α , σ , N , and δ . For two-level designs, δ equals the regression coefficient when the factor is coded ± 1 . Output for this example is shown in Figure 13.1. If the pure error mean square of 270.7 is an accurate indication of σ^2 , then $\sigma \approx 16$ and we have nearly a 75% chance of getting a statistically significant estimate for an effect with $\beta = 12$. By requesting power calculations for a range of N and true coefficient value combinations, one can better choose an appropriate sample size. Lynch (1993) tabulated values of $2\beta_i/\sigma$ corresponding to various α , β , and number of factorial points N . However, those tables provide no flexibility to choose the appropriate degrees of freedom.

For unreplicated designs, we have routinely relied on Lenth's method for estimating the standard errors and constructing tests. Calculating power for Lenth's method is not straightforward, since the power for one active effect depends on how many other effects are active (and on the magnitudes of all the active effects). Haaland and O'Connell (1995) described a power study for $N = 16$ run designs with one to eight active effects with true effects of various sizes. With only one large active effect equal to σ , the power equals 0.904, which is only slightly smaller than the power (0.937) calculated in (13.1) based on the Student's t distribution with 8 df. However, as the number of active effects increases, Lenth's power declines. Hamada and Balakrishnan

Power Details Dialog

	α	σ	δ	Number
From:	0.050	16.0	2	16
To:	.	.	12	32
By	.	.	2	8

Calculations will be done on all combinations of sequences.

Power

α	σ	δ	Number	Power
0.0500	16	2	16	0.0728
0.0500	16	2	24	0.0888
0.0500	16	2	32	0.1043
0.0500	16	4	16	0.1433
0.0500	16	4	24	0.2105
0.0500	16	4	32	0.2740
.
.
0.0500	16	12	16	0.7480
0.0500	16	12	24	0.9314
0.0500	16	12	32	0.9825

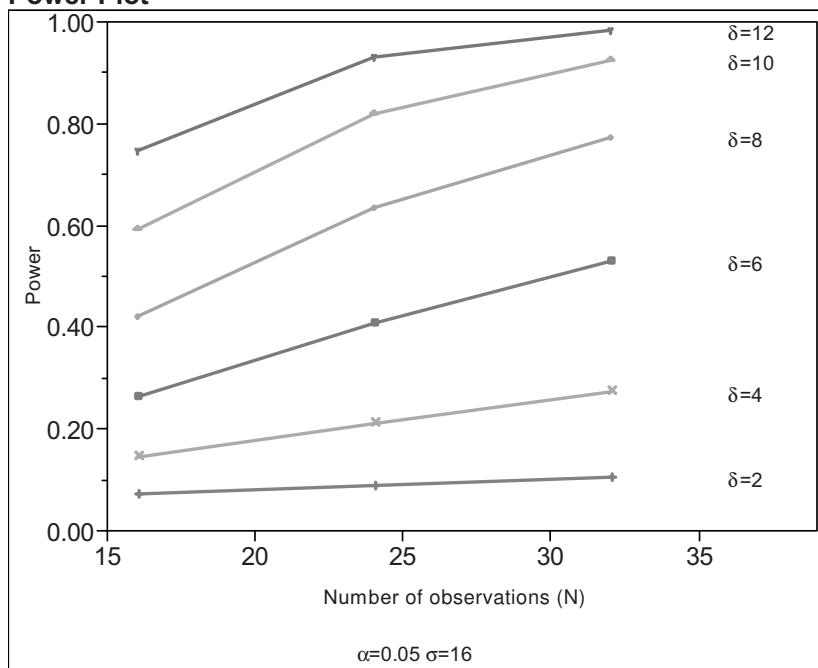
Power Plot

Fig. 13.1. Power calculations for t -tests with $N - 8$ df for error

(1998) also described a power study for Lenth's method and many alternative procedures, based on a 16-run orthogonal design and up to six active effects. Repeating their simulations for $\beta_i = \sigma$, one may verify that the power is approximately 0.88, 0.78, and 0.49, respectively, for two, four, and six active effects. Thus, for tests based on Lenth's method, the power of the test certainly depends on the validity of the sparsity of effects assumption.

To determine power for any particularly case when using Lenth's method, one can specify values for N , σ , and the true β_i 's, simulate the least squares estimates, implement Lenth's method for each simulated vector of b_i 's, and, for a particular active effect, determine the proportion of Lenth t statistics that exceed the appropriate critical value from Appendix C. To illustrate this calculation, suppose we have an unreplicated 2^3 with a single active effect of size $\beta_1 = \sigma$. (For convenience, set $\beta_1 = \sigma = 1$.) With the critical value $c_{0.05}^{\text{IER}} = 2.297$ from Appendix C, we find that the Lenth t for the active effect exceeds the critical value for only 50.5% of the simulated samples. This lower power (0.505 vs. 0.904) is primarily a consequence of the standard error of b_1 being larger for $N = 8$ versus $N = 16$.

Simulations to determine power may be performed not only for Lenth's method but also for any adaptive or model-dependent test procedure, such as those discussed in Section 14.2. Alternatively, if you are concerned with achieving a specified power, regardless of the values of other coefficients, it is best to use a pure error estimate for σ obtained by replication; then the determination of power becomes a straightforward calculation based on the noncentral t distribution. Even when one does not formally assess power, one can compute the square root of the diagonals of $(\mathbf{X}'\mathbf{X})^{-1}$ to determine the standard error for each b_i/σ . This is true whether the design is orthogonal or not.

13.1.2 Count responses

For Binomial and Poisson distributions, the variance is a function of the mean, so power calculations must be performed differently. Bisgaard and Fuller (1995) considered the case of factorial and fractional factorial designs with N treatment combinations, where the data for each treatment combination is a binomial sample of size n . Bisgaard and Fuller also assume that the analysis will be performed using the arcsin transformation shown in Figure 2.11 or the Freeman–Tukey transformation (2.12). Power depends not only on N , n , α , and the regression coefficient (Bisgaard and Fuller's 0.5Δ) but also on the average proportion in the experiment, p_0 . Bisgaard and Fuller's approximate sample size formula is

$$n = (z_{\alpha/2} + z_\beta)^2 / (N\delta^2),$$

where $\delta = \arcsin[(p_0 + .5\Delta)^{1/2}] - \arcsin[(p_0 - 0.5\Delta)^{1/2}]$. Values of n are tabulated for $\alpha = .05$ and Power = 90% (i.e., $\beta = 0.1$) for $N = 8$, 16, and 32.

Bell, Ledolter, and Swersey (2006) presented a 19-factor, 20-treatment combination design, analyzed in Section 6.3.3, where 5000 test mailings were sent for each treatment combination. The average response was 1.3%. For $\alpha = .05$ and $\beta = 0.1$, the minimum detectable δ is

$$\delta = [1.96 + 1.282]/(100000)^{1/2} = 0.01025.$$

By iteration we find that

$$\arcsin[(0.013 + 0.00116)^{1/2}] - \arcsin[(0.013 - 0.00116)^{1/2}] = 0.01025.$$

Thus, true regression coefficients larger than 0.116% for the arcsin transformation [or larger than 0.232% for the Freeman–Tukey transformation (2.12)] have at least a 90% chance of being identified as active. Bell, Ledolter, and Swersey (2006) made a similar observation and then stated, “Thus with a sample size (nN) of 100,000, the authors and marketing team were very confident of being able to detect very small, but economically meaningful differences.” To avoid the possibility of confusion, note that Bisgaard and Fuller’s $\Delta/2 = 0.00116$ corresponds to a regression coefficient for p , whereas $\delta/2 = 0.01025/2 = 0.00512$ corresponds to regression coefficients for the $\arcsin(p^{1/2})$ scale. Since $f_{FT}(p)$ is the sum of two arcsin functions [see (2.12)], the minimum detectable regression coefficient for $f_{FT}(\hat{p})$ with the specified power is $\delta = 0.01025$.

For Poisson counts c , $\text{Var}(\sqrt{c})$ approaches 0.25 as $E(c)$ increases, but can be as large as 0.4 for $E(c)$ near 2 (refer to Figure 2.15). Thus, for large $E(c)$, if the expected count is c_+ and c_- at the low and high levels, respectively, of a factor, then the approximate power for a two-level design with N runs is

$$\text{Power} = P(Z < \sqrt{N}|\sqrt{c_+} - \sqrt{c_-}| - z_{\alpha/2}). \quad (13.2)$$

This approximate formula based on a normal distribution approximation is a useful guide even when the expected Poisson counts are less than five. For instance, the median for one response variable in Hsieh and Goodwin’s (1986) 2^{9–5} experiment was three defects. What likelihood would we have of detecting the effect of a change that lowered the expected count to two? Using (13.2), the power is approximately

$$P(Z < \sqrt{16}|\sqrt{3} - \sqrt{2}| - 1.96) = P(Z < -0.689) = 0.2455.$$

As an alternative, one can simulate data from a Poisson distribution to determine the power. For instance, we simulate c_1, \dots, c_{16} from a Poisson distribution with mean 2.5 for all 16 observations; from these we compute $y_i = \sqrt{c_i}$ and then compute the estimated regression coefficient

$$b_1 = [(y_1 + \dots + y_8) - (y_9 + \dots + y_{16})]/16.$$

The upper 0.05 percentile of this null distribution for $|b_1|$ is found to be 0.295. Then we repeat the simulation using a mean of 3 for the first eight

observations and a mean of 2 for the last eight; approximately 23% of these $|b_1|$ values exceed 0.295. Thus, the power is only slightly less than the value indicated by the normal approximation (0.2455). This simple simulation is based on there being only one effect present. If there are other active effects, then the Poisson means will take on more than two values, and the distribution of the coefficient b_1 will be affected. However, the fact that the simple normal approximation gives a rough idea of power is the key point here. If this power is not sufficient, then sampling must be enlarged to increase the Poisson means and thereby raise the power.

13.2 Choice of Factor Levels

The success of two-level factorial experiments depends heavily both on the choice of factors and the choice of levels for quantitative factors. Here are some tips to guide you in choosing the levels.

If the experiment involves a process with standard operating conditions, one option is to choose levels that straddle the standard level. When the low and high levels for each factor are centered about the standard level, then the centerpoint corresponds to current practice, and every factorial treatment combination involves changes from this standard for every factor. This choice makes the inclusion of centerpoint replicates particularly relevant. Alternatively, one may choose the standard level as the low (or high) level for a factor and the other factorial level in the direction believed in advance to be the more promising search direction. If for each factor the standard level is assigned to one of the factorial levels, then this treatment combination should usually be included in the design. This is automatic if a full factorial design is conducted. For a fractional factorial, it can be determined as follows. Suppose we begin with the standard 2^{7-4} fraction as shown in Table 6.4 containing the treatment combination with every factor at the high level. Suppose the factor levels have been assigned so that $(1, 1, -1, -1, -1, -1, -1)$ is the standard operating condition. Although this treatment combination is not in Table 6.4, the fourth row agrees for all factors except **D**. Thus, if one reverses column **D**, then one obtains a $1/16^{\text{th}}$ fraction containing the standard operating condition.

For the initial Huhtamaki example (Table 1.1), these ideas were blended: current operating speed of 18 was selected as the low level for speed, whereas current operating temperature of 90° was straddled asymmetrically by the factorial temperature levels ($75^\circ, 145^\circ$). Such a choice of levels may make the most sense, even though the current operating condition is not a level for the design. For the Huhtamaki situation, current operating conditions were giving unsatisfactory results, and there was no need to replicate this fact.

Another common situation arises where a zero level is used for the low level. For instance, Example 3.1 involved five possible additives (see Table

3.1). Setting the low level at zero makes practical sense in that the experiment then investigates the absence or presence of each additive. Example 6.4 (Table 6.13) is similar, in that 12 nutrients in the media each are assigned a low level of zero. A consequence of this choice was that one treatment combination contained no nutrients and yielded no measurable growth. Every other treatment combination in Table 6.14 contained eight factors at the high level (i.e., eight nutrients present).

Even when one level is chosen to be zero, or to coincide with the standard operating condition, one must still choose the second level. Similarly, when the levels straddle a standard level, one must choose how wide the spacing must be. Expert knowledge is essential for this step. Choosing a narrow spacing risks failing to see any active effect. Choosing a wide spacing involves other risks:

- If the effect is monotonic, then excessively wide levels will make its effect dominate all the rest; interaction effects with this factor will also be more likely.
- If the factor produces good results only in an intermediate range, one risks missing this range entirely if the levels are chosen too widely. For instance, if we seek to maximize the response and a factor's effect is quadratic, then too wide a spacing may result in uniformly poor results.

Knowledge—or suspicions—about the likely effect of each factor should be explored in the planning stage. The more one understands about the “science” of the process, the better one can choose levels that are appropriate. Documenting the expected effects is an important part of the planning process; see Irvine, Clark, and Recupero (1996, p. 349) for a good example. Trial runs are strongly recommended by Coleman and Montgomery (1993). One may guess in advance which treatment combinations are likely to produce the most extreme outcomes. If these are run, they can either confirm that the level choices as a whole appear satisfactory or not. If little difference is seen in the response at these supposedly extreme treatment combinations, then some levels should be widened. Alternatively, if one or both are found to be too extreme (or unsafe), some levels should be moved closer together. Vindevogel and Sandra (1991) commented that originally a surfactant concentration of 60 mM was contemplated, but that this was found to produce excessively long analysis time (refer to quote before Table 6.1). Running a couple of trial treatment combinations likely to produce extreme results is a quick means of discovering the need for changes to the levels.

Ruggedness studies and robust parameter design applications involve varying factors to see whether they transmit variation to the response. In such cases, levels are either based on engineering specifications or on observed natural variation. Bafna and Beall (1997) chose barrel temperature levels for a melt index measurement experiment to match the high and low specifications of the relevant American Society for Testing and Materials (ASTM[®]) standard. Levels for the other factors were similarly based on the ASTM spec-

ifications. Snee (1985) presented a 2^{5-1} experiment intended to identify which of five process inputs were contributing to color variation in a chemical product. Prior data collection should have confirmed the variation in these inputs, such as reactant pH and reactant purity; otherwise, they would not be suspected as causes of the color variation. The same prior data would document the extent of variation in these inputs. Thus, one could use the range for pH in the observational data as a basis for choosing the low and high levels for pH in the experiment.

If the error variance is small, one may choose levels closer together and still see a statistically significant effect. This fact is exploited for experiments intended to identify a path of steepest ascent (see Section 9.3). Factors known to be critical should be varied less aggressively. By contrast, if the effect is likely monotone and relatively weak, wide spacing is appropriate. Example 6.6 illustrates this situation, where water temperature and exposure time are varied widely. Exposure time was varied from 10 to 120 min and water temperature was varied from 15°C to 60°C, with no recognizable effect (see Tables 6.21 and 6.23). Clearly, expert knowledge and/or trial runs are essential to choosing levels appropriately.

13.3 Tips for Studying Variation

The Huhtamaki experiments described in Chapter 1 were performed to increase the average for dry crush, a numerical response variable. Experience showed that the mean was too low for some applications, and the desired improvement was a larger mean. Other experiments focus on variability of a numerical response. Planning for variability experiments requires more prior information because variation generally has multiple sources. For example, in a variation study, the objective may be as follows:

- To determine which factors affect within-run (short-term) variance, or to determine which combination of factor levels minimizes this variation;
- To determine which factors affect long-term variation;
- To determine factors that make a response insensitive to variation in hard-to-control inputs.

Prior to conducting any experiment to reduce variability, one should first quantify the magnitude of variation attributable to different sources. For instance, Lynch and Markle (1997) collected data from 27 runs of a plasma etch process for silicon wafers. Each run of the process involved etching 18 wafers, arranged in 3 rows on each of 6 faces of a hexode. Within each run, six wafers were selected for measurement, one from each face of the hexode. The sampling of wafers was systematic so that each of the 3×6 wafer locations was sampled in 9 of the 27 runs. Finally, each of the 27×6 wafers was measured at the same 9 sites in a “box and star” pattern. Thus, for each run, we have 54 etch thickness measurements. Such systematic sampling of

wafers and sites is almost always more informative than random sampling. Here, the data were summarized as a mean and standard deviation for each run. Further partitioning of the variation would be useful to separate within wafer variation (using the average of six within-wafer standard deviations) from between-wafer variation (summarized by the standard deviation of the six wafer means). This brief summary highlights the following lessons:

- Sufficient sampling within runs is critical for experiments that focus on variation.
- Systematic samples allow one to partition the variance into specific components. Inadequate or poorly conceived sampling plans render many variation studies useless.
- Simple summary statistics may be used as the response, and each component of variation may require its own separate analysis. Fitting a single model to the data will not answer all the relevant questions. Rather, begin by plotting the data and then analyze different statistics that address particular questions. Section 14.3 illustrates such an analysis.

Recognizing which sources of variation are most prominent provides a basis for prioritizing variance-reduction efforts. Furthermore, understanding well the existing variation patterns helps one to understand its source and potential remedies. Diagnosing the problem is often harder than finding the cure. For instance, if the Lynch and Markle (1997) data had shown between-wafer variation to be the largest, one would then explore whether these differences were primarily row-to-row or face-to-face. Face-to-face variation should be eliminated by rotating the hexode. Row-to-row variation would require other solutions.

Studying between-run variability is more difficult, since it is more difficult to isolate the sources of variation. If the source or variation can be linked to variables that are measurable—or even controllable—then studying how to reduce variation becomes simpler. A variable linked to variation in the response is termed a noise variable or noise factor. Variation in the noise variable is transmitting some variation to the response of interest. The objective in such cases is to alter the relationship between the noise variable and the response, making the process more robust to variation in the noise variable. Interactions between controllable factors and the noise are of such interest because they are the key to reducing the response's sensitivity to the noise variable. This was the objective in Engel's experiment analyzed in Section 10.3.2. It is this systematic imposition (or sampling) of variation that makes robust parameter design experiments successful. For two very helpful summaries of the strategy and analysis of robust parameter design experiments, see Abraham and MacKay (1993) and Steinberg and Bursztyn (1994).

13.4 Accommodating Factors with More Levels

Four-level factors are easily constructed from two-level factorial designs by the method of replacement. To understand the method, consider the simple 2^2 factorial with factors denoted by x_1 and x_2 . In Table 13.1 we display the contrasts for these main effects and their interaction. In addition, we construct the column $\mathbf{A} = x_1 + 2x_2$. The three factorial effects x_1 , x_2 , and $x_1 * x_2$ can be replaced by a single four-level factor. If the levels of \mathbf{A} correspond to a quantitative factor's levels, then the last three columns of Table 13.1 correspond to the orthogonal linear, quadratic, and cubic contrasts for \mathbf{A} . Thus, the sum of squares for \mathbf{A} can be partitioned into three 1-df pieces using either the first three columns or the last three columns. Note that one contrast is identical in each set; the quadratic contrast [= $0.25(A^2 - 5)$] corresponds exactly to the $x_1 * x_2$ interaction. The linear (cubic) contrast for \mathbf{A} is most correlated with x_2 (x_1), with a correlation of $.8944 = .8^{1/2}$.

Table 13.1. Construction of a four-level factor by replacement

x_1	x_2	x_1x_2	\mathbf{A}	Orthogonal Polynomial Contrasts		
				Linear	Quadratic	Cubic
-1	-1	1	-3	-3	1	-1
1	-1	-1	-1	-1	-1	3
-1	1	-1	1	1	-1	-3
1	1	1	3	3	1	1

For resolution III fractional factorial designs, any pair of two-level contrasts can be replaced by a four-level column, provided they and their interaction are not aliased with any main effect. Consider the 2^{15-11} factorial design with factors x_1-x_{15} numbered to match the Yates column order (i.e., $x_3 = x_1x_2$). Saturated main effect fractional factorial designs can be constructed for each of the following cases:

- Replace x_1 , x_2 , and x_3 with $\mathbf{A} = x_1 + 2x_2$. The remaining 12 columns form a 2^{12-8} .
- Also replace x_4 , x_8 , and x_{12} with $\mathbf{B} = x_4 + 2x_8$. The remaining nine columns form a 2^{9-5} .
- Also replace x_5 , x_{10} , and x_{15} with $\mathbf{C} = x_5 + 2x_{10}$. The remaining six columns form a resolution III 2^{6-2} .
- Also replace x_6 , x_{11} , and x_{13} with $\mathbf{D} = x_6 + 2x_{11}$. The remaining three columns form a 2^{3-1} .
- Finally, replace x_7 , x_9 , and x_{14} with $\mathbf{E} = x_7 + 2x_9$. Factors $\mathbf{A-E}$ form a 4^{5-3} .

Note that this replacement method can also be applied to some orthogonal arrays. In particular, any OA($16, 2^{15}, 2$) can be utilized to create a single four-level factor, since each of the orthogonal arrays discussed in Section 6.3.2 have

some complete aliasing between main effects and two-factor interactions. For instance, $\mathbf{1} = \mathbf{2} * \mathbf{3}$ for the design in Table 6.17. Thus, we may replace the first three columns with $\mathbf{A} = x_1 + 2x_2$ and obtain the following strength-2 fraction of a $4^1 \times 2^{12}$:

A	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
3	1	1	1	1	1	1	1	1	1	1	1	1
1	1	-1	1	-1	1	1	-1	-1	-1	-1	1	1
-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1
-3	1	-1	-1	1	1	-1	1	-1	-1	1	-1	1
3	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1
1	-1	1	-1	1	1	-1	-1	1	-1	1	1	-1
-1	-1	-1	1	1	1	-1	1	-1	1	-1	1	-1
-3	-1	1	1	-1	1	-1	-1	1	1	-1	-1	1
3	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1
1	1	-1	1	-1	-1	-1	1	1	1	1	-1	-1
-1	1	1	-1	-1	-1	1	1	-1	-1	1	1	1
-3	1	-1	-1	1	-1	1	-1	1	1	-1	1	-1
3	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1
1	-1	1	-1	1	-1	1	1	-1	1	-1	-1	1
-1	-1	-1	1	1	-1	1	-1	1	-1	1	-1	1
-3	-1	1	1	-1	-1	1	-1	-1	1	1	1	-1

This orthogonal array cannot accommodate a second four-level factor without creating some loss of orthogonality. This is because the other six length-3 words that create complete aliasing between main effects and two-factor interactions all involve one of the first three columns of Table 6.17. For instance, one could replace x_4 and x_5 with the four-level factor $\mathbf{B} = x_1 + 2x_4$, but the design only has 8 of the 16 combinations for \mathbf{A} and \mathbf{B} .

Whenever the run size N is a power of 4, it is possible to create an orthogonal saturated main effect design with $(N - 1)/3$ four-level factors, such as 21 factors in 64 runs. When N is an odd power of 2, up to $(N - 5)/3$ four-level factors can be constructed (Mukerjee and Wu 1995); for instance, for $N = 2^5$, we can have up to $(32 - 5)/3 = 9$ orthogonal four-level factors. Edmondson (1991) and Gilmour (2004) discussed the construction and analysis of four-level designs for quantitative factors. Consumer preference experimentation is one context where many factors with more than two levels is common. Green and Devita (1975) performed a conjoint study involving four, four-level qualitative factors.

Generally we are interested in having just a few factors with more than two levels. For example, Schoen's (1997) cheese experiment analyzed in Section 10.3.3 involved a single four-level factor along with nine two-level factors. Wu and Zhang (1993) present 16- and 32-run minimum aberration fractional factorial designs with just one or two four-level factors; Li, Liu, and Zhang (2007) provided additional designs with one four-level factor. Tables 13.2 and

13.3 list their minimum aberration designs. For larger run sizes, see Mukerjee and Wu (2001), Zhang and Shao (2001), and Li, Liu, and Zhang (2007). For blocking of designs with two- and four-level factors, see Ai, Yang, and Zhang (2006).

When the four-level factors are quantitative, it matters how one constructs the four-level factors, assuming the cubic polynomial is less important. Consider, for example, the minimum aberration 16-run fraction of a $2^2 \cdot 4^2$. The two-level factors are based on columns 6 and 9. The choice involves how to construct the four-level factors from the sets of columns $\{C_1, C_2, C_3\}$ and $\{C_4, C_8, C_{12}\}$. Choosing $\mathbf{A} = C_1 + 2C_3$ and $\mathbf{B} = C_{12} + 2C_4$, one can estimate the 12 terms of the second-order model (i.e., 4 linear terms, 2 pure quadratic terms, and the 6 bilinear terms). The largest variance inflation factor is 1.5625. By contrast, if column C_{12} is most correlated with the linear effect of \mathbf{B} , the largest variance inflation factors = 6.25; if column C_{12} coincides with \mathbf{B}^2 , then the model is not even estimable.

Table 13.2. Minimum aberration 4×2^k fractional factorial designs, with one four-level factor constructed from columns {1, 2, 3}

<i>N</i>	<i>k</i>	Two-Level Factor Columns
16	3	4, 8, 13
	4	4, 6, 8, 13
	5	4, 6, 8, 10, 13
	6	4, 6, 8, 10, 13, 15
	7	4-6, 8-11
	8	4-11
	9	4-8, 12-15
	10	4-9, 12-15
	11	4-10, 12-15
	12	4-15
	4	4, 8, 16, 29
	5	4, 8, 16, 26, 29
32	6	4, 8, 16, 22, 26, 29
	7	4, 8, 14, 16, 22, 26, 29
	8	4, 5, 8, 14, 16, 22, 25, 31
	9	4, 5, 8, 9, 14, 16, 22, 26, 29
	10	4, 8, 11, 14, 16, 19, 21, 25, 26, 28
	11	4, 7, 8, 11, 14, 16, 19, 21, 25, 26, 28
	12	4, 7, 8, 11, 13, 14, 16, 19, 21, 25, 26, 28
	13	4, 7, 8, 11, 13, 14, 16, 19, 21, 22, 25, 26, 28
	14	4, 7, 8, 11, 13, 14, 16, 19, 21, 22, 25, 26, 28, 31
	15	4-11, 16-19, 28, 30, 31

Note: Numbers identify columns in Yates's order; see Appendix F.

Table 13.3. Minimum aberration $4^2 \times 2^k$ fractional factorial designs, with four-level factors constructed from columns $\{1, 2, 3\}$ and $\{4, 8, 12\}$

<i>N</i>	<i>k</i>	Two-Level Factor Columns
16	1	9
	2	6, 9
	3	6, 9, 14
	4	6, 9, 11, 14
	5	6, 9-11, 14
	6	5-6, 9-10, 13-14
	7	5-6, 9-11, 13-14
	8	5-6, 9-11, 13-15
	9	5-7, 9-11, 13-15
32	2	16, 22
	3	16, 22, 27
	4	16, 22, 27, 29
	5	9, 16, 22, 27, 29
	6	9, 14, 16, 22, 27, 29
	7	5, 9, 14, 16, 22, 27, 29
	8	5, 9, 14-16, 22, 27, 29
	9	5, 9, 14-17, 22, 27, 29
	10	5, 9, 14-17, 22, 27-29

Note: Numbers identify columns in Yates's order; see Appendix F.

The extension to eight-level factors is straightforward, although the need is rare. Choose three columns $\{x_1, x_2, x_3\}$ that form a 2^3 and compute $\mathbf{A} = x_1 + 2x_2 + 4x_4$ (or add 7 and divide by 2 to create eight levels from 0 to 7). The 7 df for \mathbf{A} 's main effect include the four interactions involving $\{x_1, x_2, x_3\}$, all of which must not be aliased with other main effects if the design is to have orthogonal main effects.

Factors for which the number of levels is not a power of 2 may be constructed by collapsing a factor with a larger number of levels (Addelman 1962, Wu and Hamada 2000, Sect. 7.8). For example, if we need a nominal factor with three levels, one can take a factor with levels $\{0, 1, 2, 3\}$ and assign two levels (e.g., 0 and 1) to the same nominal level. Such collapsing always results in some levels having twice the number of runs as other levels. This choice should be made judiciously; for example, the level(s) for which we have a priori more interest should receive the higher frequency. Designs constructed from two-level designs by the combination of replacement and collapsing make sense only if there are many two-level factors, since there is a loss of efficiency resulting from the imbalance. One alternative is to use a mixed-level orthogonal array. A popular small mixed-level design is the strength-2 orthogonal array in 18 runs, which can accommodate 7 three-level factors and 1 two-level

factor. See Wu and Hamada (2000, Sect. 7.5) or other sources for a list of additional orthogonal, mixed-level designs.

13.5 Special Requirements for Run Order and Treatment Combinations

This section discusses the choice of run order and special conditions on the sets of treatment combinations comprising the design. The three questions addressed are as follows:

- When is it better to utilize a systematic run order?
- When is it necessary to “reset” factor levels?
- How can one purposefully exclude (or include) particular treatment combinations?

13.5.1 Systematic run orders robust to trend or autocorrelation in the errors

Daniel and Wilcoxon (1966) explained how to construct designs robust to linear or quadratic trends in the errors. The general advice regarding randomization in their introduction is particularly insightful:

Only a small fraction of all experimental work in the physical and engineering sciences meets the orthodox design statistician’s requirements for objective randomization. How does it happen then that a considerable part of all this work produces useful, even valid, results? It happens because randomization, while generally sufficient, is not always necessary... almost as soon as (randomization) is urged, sometimes dogmatically, the demand is modified and for obvious reasons. Full randomization would in many situations guarantee results which, while entirely valid, would not detect any effects.... Randomization is used, then, after we have exhausted our knowledge of the behavior of the system under study and have taken serious steps to control what can be controlled.

Their point is that trend in the errors can often be anticipated and its effect on the results controlled by restricting the design’s run order. Such designs are much more efficient than ignoring the trend and relying on randomization alone to produce a valid experiment.

When linear or quadratic trend is considered likely, rather than using blocking as discussed in Sections 3.3 and 10.1, it is more efficient to choose a run order that permits estimation of this trend along with the factorial effects of interest. Although subsequent articles [e.g., Cheng and Jacroux (1988), John (1990), Wang and Jan (1995)] provide other construction methods and

examples of trend-robust designs, the insights in Daniel and Wilcoxon (1966) are the most helpful for developing a general approach.

For any unreplicated 2^{k-f} factorial design, we may sort the treatment combinations based on $k-f$ independent contrasts. Doing so ensures that all contrasts orthogonal to these $k-f$ contrasts will be orthogonal to linear trend in the errors. For instance, for a 2^4 we may sort the treatment combinations based on **ABC**, then **ABD**, then **ACD**, and finally **BCD**. The resulting order is shown in Table 13.4. If a trend in the errors is suspected, the model may include run number as an explanatory variable. The column “Run” will be correlated with the 4 columns used to sort the treatment combinations, but it will be orthogonal to the remaining 11 factorial effect columns. Thus, the two-factor interaction model augmented with a column for Run (properly centered by subtracting the mean of 8.5) will have a diagonal $\mathbf{X}'\mathbf{X}$ matrix.

Table 13.4. Trend-robust run orders for a 2^4 factorial

Run	A	B	C	D	Alternate Run Order
1	-1	-1	-1	-1	5
2	-1	1	1	1	6
3	1	-1	1	1	7
4	1	1	-1	-1	8
5	1	1	-1	1	1
6	1	-1	1	-1	2
7	-1	1	1	-1	3
8	-1	-1	-1	1	4
9	1	1	1	-1	9
10	1	-1	-1	1	10
11	-1	1	-1	1	11
12	-1	-1	1	-1	12
13	-1	-1	1	1	13
14	-1	1	-1	-1	14
15	1	-1	-1	-1	15
16	1	1	1	1	16

Note: Run = $8.5 + 0.5\mathbf{BCD} + \mathbf{ACD} + 2\mathbf{ABD} + 4\mathbf{ABC}$

Alternate Run Order = $8.5 + 0.5\mathbf{BCD} + \mathbf{ACD} + 2\mathbf{CD} + 4\mathbf{ABC}$

What if the trend in the errors is not linear? If the treatment combinations of a 2^{k-f} are sorted by $k-f$ factorial effects, then $0.5(k-f)(k-f-1)$ factorial effects will be correlated with any quadratic effect in the errors. In particular, all of the two-factor interactions among the sorted columns will be correlated with the quadratic component of trend. Thus, for 16-run designs, four factorial effect contrasts will be correlated with the linear trend component, six will be correlated with the quadratic trend component, and the remaining five will be

orthogonal to both. For the 2^4 , products of the four three-factor interaction columns are two-factor interactions [e.g., $\mathbf{ABC}(\mathbf{ABD}) = \mathbf{CD}$].

For the run order in Table 13.4, **ABC** has the highest correlation [$8/(85)^{1/2} = .87$] with the linear trend, the other three-factor interaction columns used to sort the treatment combinations in Table 13.4 have correlations with the linear trend of $4/(85)^{1/2}$, $2/(85)^{1/2}$, and $1/(85)^{1/2}$, respectively. Similarly, the six interactions that are correlated with any quadratic trend component have differing correlations, with **CD** having the largest correlation.

The following general approach is recommended to create designs that are efficient in the presence of linear and quadratic trend.

- Given k and f , select a blocking scheme to create a factorial design in four blocks from Table E for full factorials or Table H for fractional factorials.
 - Of the three factorial effects confounded with blocks, choose the effect of least (most) interest as the first (second) column for sorting the treatment combinations.
 - Now choose $k - f - 2$ additional contrasts other than the block generators to sort the treatment combinations.
 - Fit a full factorial model, except with Order and Order² replacing the two factorial effects most correlated with Order and Order². The remaining $N - 3$ factorial effects will be estimated with variances of $1.25\sigma^2/N$ or less (Daniel and Wilcoxon 1966, p. 264).

For instance, for the 2^4 we choose to confound **ABC**, **ABD**, and **CD** with blocks (refer to Appendix E). For Step 2, we would sort first on one of the three-factor interactions (we choose **ABC**) and then on **CD**. For Step 3, we choose to sort on **ACD** and **BCD**. This results in the alternative run order

which matches the last column of Table 13.4. For Step 4, we fit a full factorial model, except with Order and Order² replacing **ABC** and **ABD**. The **CD** interaction and one other (**BD**) will be estimated with variance $\sigma^2/12.8$ (i.e., with an efficiency of $12.8/16 = 80\%$). All other coefficients are estimated with efficiencies of 16/17 or higher.

Daniel and Wilcoxon (1966) presented three other examples, which we discuss briefly, to make the method proposed here clear.

- 2^{6-2} with generators **E** = **ABD** and **F** = **ACD**: If this design were run in four blocks, the interactions **ABC** and **BCD** would be confounded with blocks, along with their generalized interaction **AD** (and its aliases). Thus, we sort on **ABC**, then **AD**, then two other interactions; Daniel and Wilcoxon (1966, Table 8) choose **AE** and **AF**. In the analysis, we would fit a model with the six main effects, seven two-factor interactions, and the linear and quadratic order contrasts. The two three-factor interactions **ABC** and **BCD** not confounded with main effects are the contrasts most correlated with linear and quadratic trend, respectively.
- 2^5 : From Appendix E, we would block on **ABC**, **ADE**, and **BCDE**. Daniel and Wilcoxon (1966, Table 10) sorted the 32 treatment combinations first by **ABC** and then **BCDE** so that **ADE** is most correlated with quadratic trend. Subsequently, they sorted on **AE**, **BE**, and **CE**. In addition to **ADE**, nine effects are correlated with the quadratic trend: **BCE**, **ABCD**, **ACE**, **ABE**, **CD**, **BD**, **AB**, **AC**, and **BC**. If one fits the two-factor interaction model, plus Order and Order² terms, main effects and three two-factor interactions are estimated with full precision. If, instead, we sort sequentially on **BCDE**, **ABC**, **ACD**, **ABD**, and **AD**, then 10 main effects and two-factor interactions are estimated with full precision, resulting in a better design. For details, see Mee and Romanova (2009).
- 2^{14-9} : Akin to the 2^{6-2} , the 2^{14-9} fraction has 2 df for three-factor interactions not aliased with main effects. Daniel and Wilcoxon (1966, Table 11) sorted on one of these interactions (**ABC**) and then on four two-factor interactions (**AG**, **AD**, **AM**, **AN**). As a result, 3 main effects are among the 10 contrasts correlated with a quadratic trend.

If a saturated model is fit, Lenth's method may be used for the least squares estimates standardized to have a common variance. With correlations so small, the null distribution of Lenth *t* statistics is little affected. Alternatively, one may compute *p*-values by simulation using the given model matrix (Edwards and Mee 2008). If the factorial design is replicated, or if there are many degrees of freedom for higher-order interactions that may be safely assumed to be negligible, then *t* statistics can be based on the root mean square error rather than Lenth's method.

In conclusion, Daniel and Wilcoxon (1966) remarked,

If no linear or quadratic trend appears, little precision will have been lost. If any large linear or quadratic drift does occur, we will have a

large gain in precision over a fully randomized ordering of the runs, as well as a good estimate of the trend. In the only plant experiment so far completed (a 2^{9-4}), the linear and quadratic effects, measuring age of the equipment, were so large as to be the principal findings.

These designs are recommended whenever wear or drift is likely to cause such patterns in the errors, since they permit estimation of trends with little disruption to the orthogonality of the factorial effect contrasts of interest.

The designs discussed above are robust to systematic time trends in the errors that can be explained by a quadratic polynomial model. Another possibility of violation of the usual assumptions is to have autocorrelation among the errors. Cheng and Steinberg (1991) showed that run orders robust to linear or quadratic trend are not necessarily good if the errors are positively correlated. The designs they construct are trend-robust and, in the presence of positive autocorrelation, efficient for the main effect contrasts (only).

Zhou (2001) proposed the use of run orders that are robust to autocorrelation, in the sense that the variance is little affected by the presence of positive or negative autocorrelation. For 2^k and 2^{k-f} factorials, this is achieved by utilizing run orders for which each contrast of interest has $N/2$ or $N/2 - 1$ switches in sign. Zhou presented the 2^5 factorial with run order shown in Table 13.5. Each main effect column changes sign 15 or 16 times, which is ideal given Zhou's criterion, and each interaction in addition to **ACD** and **BCDE** changes sign 13-18 times. However, **ACD** changes sign 23 times, as is indicated in Table 13.5, whereas **BCDE** changes 21 times. These two interactions would be estimated with higher (lower) precision than the others, in the presence of positive (negative) autocorrelation. Zhou (2001) used a simulated annealing algorithm to search for designs with the best number of switches. Although the literature discusses designs that minimize or maximize the number of level changes for factors, there appears to be little literature for designs with an intermediate number of changes for all the contrasts.

Table 13.5. Autocorrelation-robust run order for a 2^5 factorial

A	B	C	D	E	ACD	Switch
-1	-1	-1	-1	-1	-1	
1	-1	1	1	1	1	1
1	1	-1	1	-1	-1	1
1	1	-1	-1	1	1	1
-1	1	-1	-1	-1	-1	1
-1	-1	1	-1	-1	1	1
1	1	1	1	1	1	0
-1	1	1	1	1	-1	1
1	1	1	1	-1	1	1
1	1	1	-1	-1	-1	1
1	-1	1	1	-1	1	1
1	-1	-1	1	1	-1	1
1	1	1	-1	1	-1	0
-1	-1	1	1	1	-1	0
-1	-1	1	1	-1	-1	0
-1	1	-1	1	1	1	1
-1	1	1	-1	-1	1	0
1	1	-1	1	1	-1	1
-1	-1	-1	1	1	1	1
-1	-1	-1	-1	1	-1	1
1	1	-1	-1	-1	1	1
-1	-1	1	-1	1	1	0
-1	1	1	1	-1	-1	1
1	-1	-1	1	-1	1	1
1	-1	-1	-1	1	-1	1
-1	1	-1	-1	1	-1	1
-1	1	1	-1	1	1	1
-1	1	1	-1	-1	1	1
-1	1	1	1	-1	-1	1

13.5.2 Designs without resetting of factor levels

Draper and Stoneman (1968) thoroughly studied eight-run designs for four to seven factors in terms of the effect of linear trend and in terms of the number of level changes required for each factor. Their interest in run orders that minimized the number of level changes was motivated as follows:

Such a desire might be important in situations where it is physically very difficult to change factors, or the amount of time required for apparatus to return to steady state after changes have been made is considerable and increases with the number of factors changed. (p. 301)

Dickinson (1974) considered the same issues (minimizing the number of level changes while also providing some robustness against linear trends) for 16-run and 32-run designs. See also Tack and Vandebroek (2002).

Designs with infrequent level changes, although convenient, raise questions about the proper analysis. Indeed, even Dickinson (1974, p. 36) commented, “Some caution must be exercised in performing tests of significance because the experimental plans take on some of the aspects of split-plot designs in the presence of setup errors in fixing the levels of the factors.” Ganju and Lucas (1997), Ju and Lucas (2002), and Webb, Lucas, and Borkowski (2004) discussed the difference between a true randomized design and a *random run order design*. For example,

Experimental factors, especially when they are hard-to-change, are often not independently reset on each run. For example, it would be very costly to let a mold cool down between runs and then reheat it when the same mold temperature is required on successive runs. Therefore, even when the experiment is run using a random run order, a ‘completely randomized design’ requiring a single error term is not obtained. There is a restriction on randomization that causes the experiment to require more than one error term in the model. (Ju and Lucas 2002, p. 411)

So when does it matter whether a factor is reset between runs? When there are no likely setup errors or questions about the equipment achieving a steady state, then resetting is clearly not necessary. This debate is not relevant for factors such as time. However, what about factors such as pressure or temperature that are set? Even when setup errors are likely in practice, these errors are generally small relative to the difference between the intended low and high levels of the factor, and so have little practical consequence. This was the conclusion of Langhans, Goos, and Vandebroek (2005), who stated, “While this type of reset error may be fairly common, at the same time it is unlikely that the deviations are large with respect to the total range in the factor... with only minor effect on type I and type II error rates” (p. 13). The primary focus of Langhans, Goos, and Vandebroek (2005) is a different type of error, “where there is a source of random error coinciding with the level change of a factor but of external origin, in the sense that the factor levels themselves remain unaffected.” This is the type of error assumed by Webb, Lucas, and Borkowski (2004) and Langhans, Goos, and Vandebroek (2005) where they discuss an analysis of random run order designs using mixed models. Such analysis will rarely be needed. If the experiment is small, there will not be enough precision to usefully estimate the variance components. Furthermore, if it is recognized in advance that a factor is likely prone to such errors, then intentional split-unit designs taking that factor as a whole-unit factor should be used; that is, by fixing that factor within blocks, we increase the precision of all other estimates; this is the message in Anbari and Lucas (2008) and Goos and Vandebroek (2004).

13.5.3 Including / excluding particular treatment combinations

Cheng and Li (1993) presented fractional factorial designs that are created to avoid certain combinations of factor levels. For instance, consider an experiment involving four stages, with time factors **A–D** and temperature factors **E–H**. The following constraints are imposed:

1. $\mathbf{A} = \mathbf{B} = \mathbf{C} = \mathbf{D} = 1$ is disallowed, since the time required for these runs would be excessive. Furthermore, this treatment combination is unlikely to be of any practical utility because it would severely lower throughput.
2. The high-time, high-temperature combination in the first two stages ($\mathbf{A} = \mathbf{B} = \mathbf{E} = \mathbf{F} = 1$) tends to cause fatigue to the material and so is excluded.
3. Similarly, the high-time, high-temperature combination in the last two stages ($\mathbf{C} = \mathbf{D} = \mathbf{G} = \mathbf{H} = 1$) is known to excessively harden the material and so is excluded.

What fractional factorial designs are compatible with these restrictions? Constraint 1 is satisfied if $\mathbf{ABCD} = -1$, since this ensures that one or three of the times will be at the low level. Furthermore, $\mathbf{ABEF} = -1$ and $\mathbf{CDGH} = -1$ guarantee the second and third constraints. Since these constraints involve independent interactions, the 2^{8-3} fraction with defining relation $\mathbf{I} = -\mathbf{ABCD} = -\mathbf{ABEF} = \mathbf{CDEF} = -\mathbf{CDGH} = \mathbf{ABGH} = \mathbf{ABCDEFGH} = -\mathbf{EFGH}$ is permissible. If a smaller design is desired, then one may impose an additional constraint (e.g., $\mathbf{BCEG} = \pm 1$), producing a resolution IV, 2^{8-4} fraction satisfying all of the constraints.

Cheng and Li (1993) introduced a methodology for determining whether the set of “debarred combinations” is compatible with any fractional factorial, and if so, how to suitably choose any remaining design generators, so as to achieve a design with minimum aberration, subject to the constraints imposed.

Similarly, if there are certain treatment combinations that are of particular interest, by careful selection of the generators and their sign one can obtain a fractional factorial with the treatment combination(s) of interest.

Sometimes the restriction is not that certain treatment combinations are to be avoided but rather that the design is more convenient if we restrict the number of treatment combinations required for particular subsets of factors. This situation arises frequently when the factors naturally correspond to distinct groups. Miller’s (1997) washer–dryer example discussed in Section 10.3.4 was one such example. Six of the factors involved the washer and four involved the dryer. By restricting the design to only eight treatment combinations for the washing factors, the experiment was simpler to conduct. Bisgaard’s (1997) motor assembly experiment and Lopez-Alvarez and Aguirre-Torres’s (1997) paint experiment are additional examples with groups of factors. Yates and Mee (2000) showed how to identify designs that restrict the number of treatment combinations for each subset of factors of size 4 or more. They also considered how the choice of foldover fractions is affected by these constraints.

Special Topics Regarding the Analysis

Throughout this book we have used Lenth t statistics to conduct individual significance tests for unreplicated designs. Here we discuss modifications to Lenth's method when there is minimal replication. We also discuss alternatives to Lenth's method, especially for controlling the experimentwise error. In the previous chapter, we discussed structured sampling within runs in order to study different variance components. Here we illustrate the analysis of such data. This final chapter also discusses analysis for weighted least squares, mixed-model methods, and applications with too many response variables to analyze individually. We conclude with four tips to help one avoid common data analysis errors. The sections are as follows:

- Section 14.1. Minimal Replication and Lenth's Method
- Section 14.2. Alternatives to Lenth t -Tests for Unreplicated Designs
- Section 14.3. Analyzing the Variation in Structured Samples
- Section 14.4. Generalized Least Squares Analysis When Variances Are Unequal
- Section 14.5. Mixed-Model Analysis
- Section 14.6. Highly Multivariate Response Data
- Section 14.7. Analysis Blunders to Avoid

14.1 Minimal Replication and Lenth's Method

When there is no replication and effect sparsity is reasonable, Lenth's PSE, as introduced in Section 2.4.2, provides a simple and intuitive means of estimating the standard error of effect estimates. Alternatively, if one has sufficient replication, the pure error mean square should be used to estimate σ and the

standard errors. Most software will utilize the root mean square error as an estimate for σ , no matter how few the degrees of freedom for error. Is such a choice appropriate if we have only 1 or 2 df for error? Does it not make more sense that the information about σ contained in the PSE should be utilized if, for instance, the only replication is two or three centerpoint runs?

Reconsider Example 7.3, a 2^{6-2} design with two centerpoint runs. When $y = \ln(\text{SD})$ was analyzed in Chapter 7, pure quadratic curvature as well as three-factor and higher-order interactions were assumed negligible. We now consider two alternative means of estimating σ and constructing t -tests. If one fits a saturated model, the root mean square (pure) error is 0.196, based on a single degree of freedom. The t -tests and p -values are shown in Table 14.1. While several t statistics appear large, they are not exceptional for a t distribution with only 1 df.

Table 14.1. Estimates and three sets of t statistics

Term	Estimate	Using MSE (1 df)		Using PSE ($m = 17$)			t^*
		t -Ratio	p -Value	Lenth t	p -Value		
AB = DE = CF	-0.49668	-10.15	.063	-5.38	.001	-5.03	
AF = BC	-0.44002	-8.99	.070	-4.77	.002	-4.45	
C	-0.38024	-7.77	.081	-4.12	.004	-3.85	
AC = BF	0.35327	7.22	.088	3.83	.006	3.57	
CD = EF	0.16825	3.44	.180	1.82	.081	1.70	
E	0.11113	2.27	.264	1.20	.212	1.12	
A	0.10257	2.10	.283	1.11	.246	1.04	
F	-0.08150	-1.67	.344	-0.88	.348	-0.82	
CE = DF	0.07947	1.62	.351	0.86	.359	0.80	
AD = BE	0.07926	1.62	.352	0.86	.361	0.80	
AEF = ...	0.06150	-1.26	.428	-0.67	.485	-0.62	
B	-0.05773	-1.18	.448	-0.63	.555	-0.58	
BD = AE	-0.04375	-0.89	.535	-0.47	.651	-0.44	
A² = ...	0.08781	0.60	.657	0.32	.757	0.30	
D	-0.01948	-0.40	.759	-0.21	.838	-0.20	
BCD = ...	0.01732	0.35	.783	0.19	.856	0.18	

Suppose instead that one were to use Lenth's method to estimate σ . How many estimates do we include in the calculation of the PSE?

- $m = 15$: Excluding \mathbf{A}^2 and using only the 15 factorial effect estimates in Table 14.1, the PSE = $1.5(0.079) = 0.119$. The corresponding estimate for σ is $0.119(16)^{1/2} = 0.476$.
- $m = 16$: If one includes the pure quadratic term and computes the PSE from all 16 estimates in Table 14.1, we first sort the estimates in terms of the magnitude of the t statistics, since $b_{\mathbf{A}^2}$ has a different standard error than the other effects. The resulting PSE is $1.5 \times 0.070 = 0.105$.

- $m = 17$: JMP's Screening analysis includes the 1 df for pure error as a 17th estimate and computes a PSE = 0.087. However, since the experiment involves 18 observations, JMP scales all the estimates to have standard errors equal to $\sigma/(18)^{1/2}$. The equivalent PSE for estimates in Table 14.1 [with standard errors $\sigma/(16)^{1/2}$] is

$$PSE = (18/16)^{1/2} \cdot 0.087 = 0.092,$$

which is $1.5(0.0615)$. This is the PSE used to compute Lenth t statistics in Table 14.1.

The p -values reported in Table 14.1 for Lenth t were provided by JMP via simulation. Depending on the circumstances, using Lenth's method with $m = 17$ may be preferred to using the RMSE based on 4 df. The PSE is based on an assumption of effect sparsity (i.e., that no more than three or four effects are active; the validity of the RMSE as an estimate for σ is based on the assumption that **AEF**, **BCD**, **A²**, and all of their aliases are zero. The estimator for σ here that seems least attractive is the pure error root mean square, since with only 1 df, the power would be very poor.

Edwards and Mee (2008) proposed another means of combining Lenth's PSE and pure error estimates for σ . It is based on the logic that pure error contrasts should be given more weight than other estimates and that both the pruning of large estimates and the final estimate of σ should take the mean square pure error into account. The steps are as follows:

1. Fit a saturated model. From this we obtain the mean square pure error and the corresponding t statistics for each term.
2. Compute the median of the absolute values of these t statistics and multiply this by 1.5; denote this quantity t_0 .
3. Compute $t_0^c = [\omega t_0^2 + 1 - \omega]^{1/2}$ using the recommended weight $\omega = m/(m + 15 \text{ df}_{\text{PE}})$, where m is the number of contrasts used computing the PSE, and df_{PE} is the pure error degrees of freedom.
4. Exclude estimates with t statistics exceeding $2.5t_0^c$ in absolute value and compute PSE_t , the PSE based on the median of the remaining $|t|$ statistics.
5. Compute $r(\pi) = [\pi \text{PSE}_t^2 + 1 - \pi]^{1/2}$ using the recommended weight $\pi = m/(m + 3\text{df}_{\text{PE}})$.

The quantity r_π is the ratio of a combined estimator for σ and one based only on replication. If $\text{df}_{\text{PE}} \leq 3$, generally a combined estimator will provide more power than using only the pure error mean square.

For the example in Table 14.1, the median $|t|$ is 1.645, and so $t_0 = 2.47$. With $m = 16$ and $\text{df}_{\text{PE}} = 1$, $\omega = 0.484$ and $t_0^c = 1.86$. In Step 3, four estimates are pruned, and the median of the remaining $|t|$ statistics is 1.44. Thus, $\text{PSE}_t = 2.16$ and

$$r(\pi) = [(16/19)2.16^2 + (3/19)]^{1/2} = 2.02.$$

Since $r(\pi) = 2.02$, the combined estimator for σ is $2.02(0.196) = 0.395$, double the root mean square pure error, and the t^* statistics computed using the

combined standard error are half the size of the ordinary t statistics, as shown in the last column of Table 14.1. For these data, Lenth's method from the 17 estimates (16 least squares estimates and 1 null effect) produced a slightly smaller estimate (0.369) for σ , but this is not generally the case. Edwards and Mee (2008) explained how to obtain p -values by simulating the null distribution of these modified t statistics. The combined estimator method is preferred to Lenth's method, especially if there is more than 1 df for pure error.

14.2 Alternatives to Lenth t -Tests for Unreplicated Designs

As mentioned in Section 2.4.3, many other procedures have been proposed for analyzing unreplicated factorial experiments. Hamada and Balakrishnan (1998) used simulation to evaluate the power for 24 procedures after calibrating them to have an individual error rate of approximately 0.044 under the no-effects scenario. Their Figures 2–5 show the power when either one, two, four, or six effects were nonzero, and of the same size. For effects large enough to be identified 50% of the time by the best procedure, Lenth's method was found to be an overall good choice. In particular, although never the best, it was in the top third for all four cases. Thus, given its overall good performance and simplicity, Lenth's method has been emphasized in this book.

Dong (1993) proposed a procedure identical to Lenth's first stage; that is, all estimates exceeding $2.5s_0$ are excluded. However, rather than using the median of the remaining contrasts, Dong (1993) proposed using the root mean square of the contrasts. The acronym ASE for adaptive standard error is used to refer to this procedure. Haaland and O'Connell (1995) verified what is intuitively reasonable—that Dong's method is preferred to Lenth's method if there are very few true effects, since the mean square error is a more efficient estimator, but less robust, than the median. Haaland and O'Connell's (1995) recommendations are as follows:

- “If only one estimator can be used, the PSE-based test as defined by Lenth (1989) consistently provides high overall power”—the only exception being when the number of active effects is near 50%, such as 7 out of 15.
- Dong's (1993) ASE is recommended if fewer than 20% of the effects are likely to be active. “In practical terms, this case corresponds to early screening experiments in which there are many factors and only a few are thought to be significant. The ASE-based test should not be used routinely except in this case.”
- Lenth's (1989) method is recommended “if there is a priori reason to suspect that there are a moderate number of effects.” By moderate, Haaland and O'Connell mean 20%–40% active, “which would be relevant to intermediate screening experiments in which there are fewer factors and there are likely to be important interactions.”

- If more than 40% of effects are deemed likely, a modification to the first step, which will prune more severely, is recommended. “This corresponds roughly to the prior (belief that the proportion of active effects) > 40%, which is relevant to late screening experiments in which there are few factors, most of which are thought to be important, and there are important interactions.”

Schoen and Kaul (2000), motivated by these recommendations, provide extensive tables for three procedures that correspond to these recommendations. For their version of the Dong and Lenth procedures, Schoen and Kaul (2000) use a cutoff of 3.707 times the median of the absolute values of the standardized contrasts, rather than $2.5(1.5) = 3.75$ as suggested by Lenth (1989) and followed by Dong (1993). This slightly more aggressive pruning makes the necessary critical values listed by Schoen and Kaul modestly larger than those in Ye and Hamada (2000) and this book’s Appendix C. For instance, for $m = 7$ and an individual error rate of 0.05, both Appendix C and Ye and Hamada list the margin of error to be $2.297 \times \text{PSE}$, which is $2.297 \times 1.5 = 3.445$ times the median. Schoen and Kaul list $2.40(1.446) = 3.470$ times the median to achieve the same 0.05 level. For the version of Lenth’s method suitable with a large proportion of effects active, Schoen and Kaul (2000) proposed using the ordered effect closest to the 45th percentile. For practitioners who wish to follow Haaland and O’Connell’s (1995) advice and use Dong’s method when extreme sparsity is expected and use a more robust version of Lenth’s method when many effects are expected, Schoen and Kaul’s tables, with $m = 7\text{--}127$ and $\alpha = .01, .05, .1, .15$, and $.2$, are a great resource. We briefly re-analyze two examples to illustrate the calculations.

We illustrate Dong’s analysis using Example 6.3. Table 6.12 provided Lenth t statistics for this 2^{13-9} design. Here is a situation where perhaps few effects were expected, and so Deng’s method might be more powerful. Lenth’s PSE = 0.08344 for Tear and the smallest p -value = .078. Thus, these data appear entirely consistent with the hypothesis that no effects are active. (This comment is not to suggest that one may choose the method of analysis based on an inspection of the data.) The steps for Dong’s method, as implemented by Schoen and Kaul (2000), are as follows:

1. Determine the median magnitude estimate and multiply by 3.707. In Table 14.2, the median estimate is 0.056.
2. Since no estimates exceed $3.707 \times 0.056 = 0.208$, we compute the sum of squares of the 15 estimates, divide by 15, and take the square root, $0.006013^{1/2} = 0.07754$. When no estimates are excluded, this simply equals the standard deviation of the original $N y$ values, divided by $N^{1/2}$. Multiplying by Schoen and Kaul’s $cc_2 = 1.095$ provides an unbiased estimator of the standard error: $1.095(0.07754) = 0.085$.
3. Estimates exceeding 1.89 (1.60) times the standard error are statistically significant at $\alpha = .05$ ($.10$).

The largest estimate divided by the standard error is $0.157/0.085 = 1.847$. Thus, using the ASE rather than the PSE, we obtain a similar t -ratio and p -value. Even this procedure that performs best when very few effects are active does not identify any estimates as statistically significant here.

Table 14.2. Lenth's analysis for Example 6.3 (Irvine et al. 1996); sorted estimates for $y = \text{Tear}$

Term	Estimate	t -Ratio	p -Value
C	0.157	1.880	.078
N	0.122	1.461	.149
J	-0.112	-1.341	.181
B	-0.101	-1.206	.224
A	0.093	1.116	.253
D	-0.078	-0.936	.330
CD	0.061	0.727	.446
H	-0.056	-0.667	.500
E	-0.044	-0.532	.630
M	-0.039	-0.472	.668
L	0.038	0.457	.679
G	0.033	0.397	.719
K	-0.021	-0.247	.820
JH	0.007	0.082	.942
F	-0.002	-0.022	.982

We consider Example 7.2 to illustrate Schoen and Kaul's (2000) method that is robust to having a larger proportion of the effects active. Table 7.6 contains results based on fitting the two-factor interaction model. For the 2^{6-2} design, this leaves 2 df for error, based on omitting all three-factor and higher-order interactions. If instead we were to use Lenth's method, the results would be as given in Table 14.3. The PSE = 0.2156, and four main effect estimates have p -values $< .05$. If for this resolution IV design we had expected most of the main effects to be active as well as some of the interactions, then the more robust version of Lenth's method is appropriate. Apply the steps from Schoen and Kaul:

1. For $m = 15$ contrasts, we find the seventh smallest estimate (here, 0.1604), and multiply by 2.182. In general, round $0.45m + 0.5$ to the nearest integer to choose which order statistic, and use the multiplier in their Table 3.
2. Six estimates are larger than $2.182(0.1604) = 0.350$. The median of the remaining nine is 0.1104. Multiplying this by $cc_2 = 1.800$ yields $1.8(0.1104) = 0.199$ as the standard error for these estimates.

Table 14.3. Lenth's analysis for Example 7.2 (Bafna and Beall 1997)

Term	Estimate	t -Ratio	p -Value
A	2.1354	9.90	.0003
C	1.0521	4.88	.0026
D	0.6521	3.02	.0198
F	0.4854	2.25	.0434
B	0.4604	2.14	.0511
BD = CF	-0.3688	-1.71	.0958
AE = BC = DF	-0.1854	-0.86	.3629
CD = BF	0.1729	0.80	.3945
AB = CE	-0.1604	-0.74	.4295
AD = EF	-0.1271	-0.59	.5902
ACF = ...	0.1104	0.51	.6401
AC = BE	0.1062	0.49	.6508
E	0.0937	0.43	.6872
AF = DE	0.0313	0.14	.8948
ACD = ...	0.0062	-0.03	.9787

3. Estimates exceeding 2.70 (2.01) times the standard error are statistically significant at $\alpha = .05 (.10)$.

Only three estimates in Table 14.3 exceed $2.70(0.199) = 0.5373$, and just two more exceed $2.01(0.199) = 0.40$. Thus, even though this procedure is more robust to violations of effect sparsity and provides a slightly smaller standard error estimate, the resulting p -values are less extreme. Even here, the ordinary Lenth method seems quite satisfactory. Perhaps with larger m , the gain in robustness would exceed the increased variance that comes from using a smaller order statistic.

We now turn to methods that control the experimentwise error rate, rather than only the Type I error risk for individual tests.

14.2.1 Controlling experimentwise risk of Type I errors

If the number of tests is large and the sparsity of effects assumption is true, methods that control only the risk of Type I errors for each test individually are prone to declaring many negligible effects to be active. In Section 2.4.2, we illustrated the performance of Lenth's method for a 16-run orthogonal design when none of the factors has any effect. If we control the individual error rate (IER) to 0.05, then there is a 39.5% chance of declaring at least one effect active. Often, such a risk is acceptable, given the importance of finding the truly active effects. However, for larger experiments, the experimentwise error rate can be near 1, if we only control IER to a typical 0.05 level.

Here we discuss briefly methods for controlling the experimentwise error rate (EER) for unreplicated, orthogonal designs; such methods determine

critical values such that the probability of making one or more Type I errors across all the tests to be conducted does not exceed a specified α . When these methods are employed, it is typical to allow for much larger α than .05. Miller (2005) commented that allowing an experimentwise error rate risk of .20 or higher is common, and Daniel (1959) even contemplated screening situations with the EER α set to .4 or .8. For a helpful discussion of the difference between controlling IER versus EER, see Miller (2005, Sect. 3).

Lenth's method provides a simple means of controlling EER, which may be adequate if there are very few active effects. The needed critical values c_{α}^{EER} are in the second part of Appendix C, and JMP's Modeling Screening calculates the corresponding simultaneous-test p -values. However, other multistep procedures are more powerful than this simple method. We now present recently developed step-up and step-down tests. Step-down tests begin with the largest estimate and continue until the first nonsignificant effect is reached. Conversely, step-up tests begin with a small number of effects assumed insignificant and proceed by adding other negligible estimates until the first significant estimate is identified. Hence, step-up tests are akin to backward elimination regression methods, whereas step down tests are akin to forward selection.

Venter and Steel (1998) made the general comment that step-down procedures tend to be simpler computationally, while step-up procedures tend to have greater power. Ye, Hamada and Wu (2001) propose a step-down version for Lenth's method. The computations are illustrated for the Bafna and Beall (1997) data in Table 14.4, where the estimates are sorted from most to least significant. Using an EER $\alpha = .05$, we compare the largest Lenth t (9.903) with $c_{.05}^{\text{EER}} = 4.24$, and conclude A's effect is not zero. Next, we apply Lenth's method to the remaining 14 estimates; that is, we recalculate s_0 and the PSE. For our example, s_0 changes but the PSE does not change because it is still based on the smallest 12 estimates. The Lenth t statistic 4.879 for **C** is compared with $c_{.05}^{\text{EER}} = 4.33$ (the .05 critical value for 14 contrasts given by Ye, Hamada, and Wu 2001). Once again, we reject the null hypothesis and proceed to consider the next largest estimate. The next Lenth t is 3.024, which does not exceed the .05 critical value for 13 contrasts. Hence, the step-down test stops, having concluded that two effects are active. If we had used a larger α (e.g., .3), we would continue until reaching 2.07 for **CF**, which does not exceed the .30 critical value for 10 contrasts. Thus, controlling the risk of making any Type I errors to .30, we would conclude that five main effects are active.

Table 14.4. Ye et al. (2001) Step-Down Lenth Test for Bafna and Beall data

Term	Estimate	Lenth t	Ye et al. Sequential		
			s_0	PSE	t
A	2.1354	9.903	0.6484	0.2156	9.903
C	1.0521	4.879	0.6250	0.2156	4.879
D	0.6521	3.024	0.6016	0.2156	3.024
F	0.4854	2.251	0.5391	0.2156	2.251
B	0.4604	2.135	0.4766	0.1906	2.415
CF	-0.3688	-1.710	0.4453	0.1781	-2.070
DF	-0.1854	-0.860	0.4141	0.1656	-1.119
CD	0.1729	0.802	0.4063	0.1625	1.064
AB	-0.1604	-0.744	0.3984	0.1594	-1.007
AD	-0.1271	-0.589	0.3750	0.1500	-0.847
ACF	0.1104	0.512	0.3516	0.1406	0.785
AC	0.1062	0.493	0.2344	0.0937	1.133
E	0.0937	0.435	0.1172	0.0469	2.000
AF	0.0313	0.145	0.0703	0.0281	1.111
ACD	-0.0062	-0.029	0.0234	0.0094	-0.667

As noted by Voss and Wang (2006), the step-down method proposed by Ye, Hamada, and Wu (2001) has not been proven to control EER under all possible scenarios. Voss and Wang (2006) proposed an adaptive step-down test that for orthogonal designs does guarantee control of the EER under any configuration of the true coefficients. One hindrance to application is that, to date, the user must generate by simulation the necessary critical values. Thus, at present, the Ye et al. step-down test is easier for practitioners to implement. Another simple method discussed later in this section is due to Miller (2005), who furnished critical values for designs of sizes 8, 12, and 16.

We now consider a step-up test, since Venter and Steel (1998) indicated that step-up tests tend to have greater power. To apply a step-up test, one must begin by assuming that some specified number ν of true effects are zero. Note that this is not the same as stating which ν terms have zero coefficients; one only specifies at least ν of the m coefficients are zero. Wu and Wang (2008) proposed the intuitive procedure that begins with the model containing terms corresponding to the $m - \nu$ largest (standardized) estimates; thus, the MSE is based on the ν least significant terms. From this point, one uses the F (or t) statistics from backward elimination regression to sequentially drop additional terms. The process continues until the first test statistic exceeds the appropriate critical value. When this happens, one essentially concludes that all remaining terms in the model are statistically significant. This is named a step-up sequentially-scaled (SUS) test because we are considering larger estimates at each step, and the estimates are divided by a mean square error with more degrees of freedom as terms are dropped.

Wu and Wang (2008) illustrated their SUS test procedure for a 16-run design, assuming at least $\nu = 7$ of the 15 terms are negligible and using $\alpha = .05$. We use the same critical values to analyze the Bafna and Beal (1997) 2^{6-2} data. Table 14.5 shows the sequence of estimates from smallest to largest. The last column shows the .05 critical value from Wu and Wang (2008). Only the last F statistic, 28.31, exceeds the corresponding critical value. Thus, at $\alpha = .05$, we can declare only that the **A** effect is active. Note that the Lenth step-down test declared more (2) effects significant at $\alpha = .05$. At $\alpha = .3$, the Lenth step-down test identified five active factors, whereas Wu and Wang's step-up procedure identifies six.

Table 14.5. Step-up test for Bafna and Beall

No.	Last Term				$\alpha = .05$
Inactive	Removed	Estimate ²	MSE	F-Ratio	Critical value
1	ACD	0.0000	.	.	.
2	AF	0.0010	0.0000	.	.
3	E	0.0088	0.0005	.	.
4	AC	0.0113	0.0033	.	.
5	ACF	0.0122	0.0053	.	.
6	AD	0.0162	0.0067	.	.
7	AB	0.0257	0.0082	.	.
8	CD	0.0299	0.0107	2.78	14.9
9	DF	0.0344	0.0131	2.62	16.4
10	CF	0.1360	0.0155	8.78	16.0
11	B	0.2120	0.0275	7.70	15.5
12	F	0.2356	0.0443	5.32	15.1
13	D	0.4252	0.0603	7.06	14.6
14	C	1.1069	0.0883	12.53	14.3
15	A	4.5600	0.1611	28.31	14.0

Wu, Mee, and Ford (2009) proposed a simplification to this step-up test that utilizes a single critical value for a given m , ν , and α , rather than a sequence of values as shown in the last column of Table 14.5. For $m = 15$, $\nu = 7$, and $\alpha = .05 (.30)$, that value is 15.9 (8.34). For more details, see their technical report.

Finally, we mention the all possible comparisons (APC) procedure by Miller (2005), which like step-up procedures, requires that one specify the maximum number of active effects to be considered. For each model of a given number of terms, Miller's test is the ratio of a constant, divided by $1 - R^2$. The model with the maximum score is chosen and all of its terms are declared statistically significant. This procedure is exceedingly simple for orthogonal designs of size $N = 8, 12$, and 16 , where Miller provided the required constants that strictly control either the EER or the IER for orthogonal designs of size.

The challenge of computing constants with sufficient precision for larger N currently hinders the more widespread use of the APC method. For the Bafna and Beall data, the model with only **A** is chosen for EER controlled at .05. Only two terms are declared active for EER $\alpha = .3$, but this is not necessarily indicative of lower power, since the test statistic was very nearly maximized for the six-term model. See Table 14.6, where we show the results for EER at .05 and .30, as well as for IER at .05. Whereas only four main effects had Lenth *t* statistics below $c_{.05}^{\text{IER}}$, Miller's method declares six effects to be active, including five main effects and one interaction. Miller's (2005) power simulations for the $N = 12$ case reveal the huge advantage for the APC procedure when the sparsity of effects assumption does not hold. APC is recommended for such cases, provided the necessary constants are available.

Table 14.6. Miller's (2005) APC tests for strict control of EER and IER, with Bafna and Beall data ($\nu = 7$ terms assumed negligible)

No.	Term	EER $\alpha = .05$		EER $\alpha = .30$		IER $\alpha = .05$	
		Terms Added	$1 - R^2$	Constant	Score	Constant	Score
0	-	1.0000	1	1.00	1	1.00	1
1	A	0.3309	0.5246	<u>1.59</u>	0.6578	1.99	0.6828
2	C	0.1685	0.2639	1.57	0.4217	<u>2.50</u>	0.4524
3	D	0.1061	0.1264	1.19	0.2625	2.47	0.2893
4	F	0.0715	0.05721	0.80	0.1579	2.21	0.1775
5	B	0.0404	0.0242	0.60	0.09128	2.26	0.1037
6	CF	0.0205	0.009436	0.46	0.05045	2.46	0.05709
7	DF	0.0154	0.003334	0.22	0.02657	1.72	0.02907
8	CD	0.0110	0.001068	0.10	0.01372	1.24	0.01617
							1.47

Note: For each α , the maximum score is underlined, indicating the last term to be included.

14.2.2 Controlling false discovery rate

Haaland (1998, p. 34) argued against controlling EER tightly and claimed that “an experimenter is seldom going to care very much about the EER.” This comment has merit, since a small EER α seeks complete avoidance of Type I errors. An alternative to controlling either the individual or experimentwise error rate is to control the false discovery rate (FDR), as proposed by Benjamini and Hotchberg (1995). Kimel, Benjamini, and Steinberg (2008) applied this methodology to factorial experiments and commented that FDR is “a more relevant quantity to control than the probability of a single erroneous declaration... The idea behind FDR is to control the proportion of false discoveries among all contrasts identified as active. Thus, when all of the contrasts are inert, the FDR will be very cautious about identifying any effects.

As the number of active effects increases, the FDR will be tolerant of some false identifications, with a corresponding increase in power" (p. 33). Define V as the number of misidentified inert effects (i.e., as the number of Type I errors) and R as the number of effects declared active; m denotes the total number of effects tested. Whereas IER controls the $E(V) \leq m\alpha$ and EER controls $P(V > 0) \leq \alpha$, FDR controls $E(V/R)$ below some specified level.

For unreplicated experiments, Kimel, Benjamini, and Steinberg (2008) proposed FDR procedures based on the test procedures of Lenth (1989) or Dong (1993). Lenth's method seems more appropriate, since Dong's method tends to have lower power, unless the number of active effects is very small. We now illustrate the FDR procedure using the Lenth analysis of the Bafna and Beall (1997) data. Table 14.3 contains approximate p -values obtained by simulation from the null distribution for Lenth t statistics. The Benjamini and Hochberg (1995) FDR procedure utilizes these p -values, denoting their ordered values as $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$. Determine

$$l = \max\{i : (m/i)p_{(i)} \leq q\}, \quad (14.1)$$

where q is the false discovery rate allowed. Table 14.7 contains the sorted $p_{(i)}$'s and the corresponding $(m/i)p_{(i)}$. If we allow up to $q = 10\%$ of the effects declared active to be misidentified inert effects, then the maximum defined by (14.1) is 3; hence, we would conclude that three effects are active. If instead, one allowed $q = 0.20$, then five effects would be identified as active.

Table 14.7. False discovery rate analysis of Bafna and Beall (1997) data

i	Term	$p_{(i)}$	$(m/i)p_{(i)}$	$[(m-l)/i]p_{(i)}$
1	A	0.0003	0.004	0.003
2	C	0.0026	0.019	0.013
3	D	0.0198	0.099	0.066
4	F	0.0434	0.163	0.108
5	B	0.0511	0.153	0.102
6	CF	0.0958	0.239	0.160
7	DF	0.3629	0.778	0.518
8	CD	0.3945	0.740	0.493
9	AB	0.4295	0.716	0.477
10	AD	0.5902	0.885	0.590
11	ACF	0.6401	0.873	0.582
12	AC	0.6508	0.814	0.542
13	E	0.6872	0.793	0.529
14	AF	0.8948	0.959	0.639
15	ACD	0.9787	0.979	0.652

The FDR procedure just described tends to be too conservative, since it provides protection assuming that all m of the effects are inert. Benjamini,

Krieger, and Yekutieli (2006) proposed an adaptive procedure that adds the following step. Given the results of (14.1), with l of the contrasts having been declared active, one now determines

$$l' = \max\{i : [(m-l)/i]p_{(i)} \leq q/(1+q)\} \quad (14.2)$$

and declares the largest l' estimates to be statistically significant. For $q = 0.1$, this adaptive step makes no difference for this example. However, for $q = 0.2$, $l' > l$, and we illustrate the calculations. The last column of Table 14.7 gives $[(m-l)/i]p_{(i)}$ for $l = 5$, as was obtained when $q = 0.2$. Since $q/(1+q) = 0.167$, which exceeds the first $l' = 6$ values for $[10/i]p_{(i)}$, six estimates are declared active.

Equation (14.2) is the simplest adaptive procedure for FDR. An alternative adaptive FDR procedure proposed by Benjamini and Hotchberg (2000) is detailed by Kimel et al. (2008). Which procedure declares more effects active is example dependent.

In Section 4.2.1, we analyzed data from a 2^9 factorial, using an EER $\alpha = .20$ to determine which estimates to declare statistically significant. Rather than controlling the risk of making any Type I errors, suppose we were to require that at most 5% of our effects declared active were false positives. Multiplying the p -values in the last column of Table 4.3 by $511/i$, the first $l = 24$ rows have values < 0.05 , whereas for $x_4 * x_6$, we get $0.0025198(511/25) = 0.0515$. Computing (14.2), we get $l' = l = 24$. Thus, we declare the largest 24 estimates to be statistically significant while assuring that at least 95% are in fact active. Thus, controlling the FDR at no more than 5% here identifies more effects than when we limit EER to 20%.

14.3 Analyzing the Variation in Structured Samples

Section 13.3 described the benefits of systematic sampling within a run, rather than random sampling. Czitrom, Mohammadi, Flemming, and Dyas (1998) described a three-factor experiment to reduce the within-run variability for oxide thickness in a batch process for silicon wafers. The horizontal furnace was designed to hold 200 wafers. However, one end of the furnace could not be used under present conditions due to the excess variability there. The three factors investigated were H_2 flow, H_2/O_2 ratio, and whether a heat plug was inserted after the wafers. Twelve runs were performed, in the order and under the conditions shown in Table 14.8. For each run, 12 wafers were measured at 13 predetermined sites on the wafer surface. Consistent with the advice of Section 13.4, the 12 wafers were selected using a systematic sample, taking every 18th wafer. Data for the first two runs are displayed in Figure 14.1, where each box-plot summarizes the 13 thickness measurements on a single wafer. The first run represents current operating conditions with its excess variability in thickness, especially for low wafer position numbers (which correspond to

the furnace door end). The second run, with higher H_2 flow and lower H_2/O_2 ratio produced marked improvement.

Table 14.8. Czitrom et al. (1998) oxide thickness experiment

Run	H_2	H_2/O_2	Heat	Plug	Site-to-site		Wafer Means	Wafer-to-wafer Component
					Mean	Std. Dev.		
1	13	1.8		Yes	204.44	4.55	4.25	4.06
2	15	1.2		Yes	198.60	2.28	1.27	1.10
3	15	1.8		Yes	206.44	3.05	1.91	1.71
4	11	1.8		Yes	206.36	5.53	5.62	5.40
5	11	1.2		Yes	199.20	2.28	1.12	0.92
6	13	1.5		Yes	203.29	2.67	1.81	1.65
7	15	1.2		No	199.56	2.83	0.69	0.00
8	15	1.8		No	207.04	3.83	2.29	2.02
9	11	1.8		No	203.22	6.39	5.41	5.11
10	11	1.2		No	199.63	2.54	1.13	0.88
11	13	1.5		No	203.65	2.83	1.27	1.00
12	13	1.8		Yes	205.05	4.02	3.12	2.91

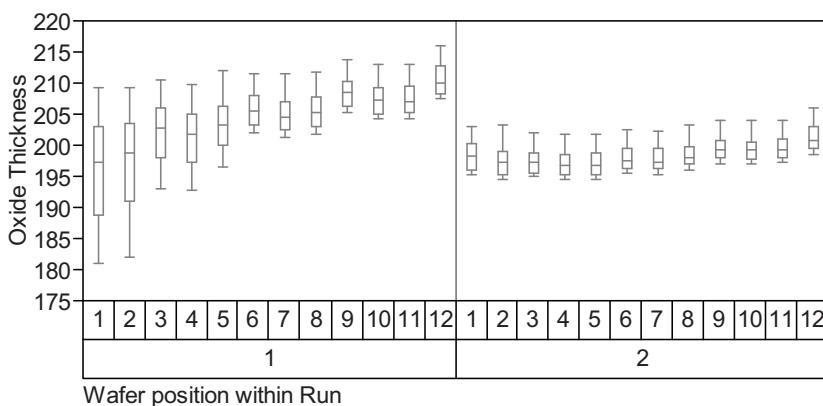


Fig. 14.1. Oxide thickness box-plots for runs 1 and 2, by wafer position

Here, the effects are so dramatic that simply plotting the data for all 12 runs would provide the essential insights needed. However, in general, one would need to compute summary statistics for each run, to be used as the response variables. Czitrom et al. (1998) used three summary statistics to summarize each run:

- Mean thickness

- Site-to-site variability within a wafer
- Wafer-to-wafer variability within a run

We illustrate the computations, using the data for run 1, which are given in Table 14.9. The mean thickness for run 1 is the average of all 156 measurements or, equivalently, the average of the 12 wafer means, $(196.58 + \dots + 210.64)/12 = 204.44$. To capture the site-to-site variability, Czitrom et al. squared each standard deviation, averaged these, and then took the square root:

$$s_{\text{site}} = [8.14^2 + \dots + 2.62^2]^{1/2} = 4.55.$$

To capture wafer-to-wafer variability, we compute the standard deviation of the 12 means. For run 1, this is 4.251. Czitrom et al. used the more complicated measure of wafer-to-wafer variability, $s_{\text{wafer}} = [4.251^2 - 4.55^2/13]^{1/2} = 4.06$. This extra calculation is not necessary here. In a random sampling setting, such a correction for within-wafer variability makes sense in the estimation of the between wafer variance component. However, with systematic sampling, it is less well motivated.

Table 14.9. Czitrom et al. (1998) oxide thickness data for run 1

Site	Wafer Position										
	1	2	3	4	5	6	...	11	12		
1	196.79	197.60	201.89	200.81	202.37	205.10	...	206.36	210.04		
2	187.20	188.88	196.75	195.96	198.61	202.47	...	204.70	207.84		
3	197.30	198.87	202.71	201.75	203.14	205.54	...	207.73	210.95		
4	202.77	202.84	205.52	204.52	205.47	207.51	...	208.87	212.16		
5	195.20	196.32	201.25	200.26	202.22	204.98	...	205.44	208.66		
6	190.20	192.65	198.94	198.23	200.68	203.87	...	206.33	209.42		
7	203.84	204.36	206.58	205.66	206.23	207.97	...	210.02	213.20		
8	202.77	202.72	205.59	204.53	206.10	208.01	...	208.34	211.41		
9	187.16	189.30	197.17	196.45	199.32	202.78	...	204.26	207.40		
10	181.12	181.97	192.99	192.85	196.51	201.88	...	204.87	207.55		
11	202.99	205.36	207.54	207.07	207.76	209.27	...	212.02	213.83		
12	209.23	209.30	210.39	209.68	211.94	211.57	...	213.05	215.96		
13	198.94	200.71	204.33	203.82	205.75	207.77	...	207.00	209.94		
Wafer mean	196.58	197.76	202.43	201.66	203.55	206.06	...	207.61	210.64		
Std. Dev.	8.14	7.76	4.94	4.85	4.22	2.91	...	2.78	2.62		

The same calculation was performed for each of the 12 runs. The results are shown as the last four columns of Table 14.8. We now proceed to analyze each of these summary statistics. First, however, one should consider the design more carefully. The 12 runs consist of 11 distinct treatment combinations. Runs 1 and 12 correspond to the current condition and were purposely placed at the start and end of the design to check for consistency. The remaining

runs correspond to a 2^3 factorial, plus two runs that serve as centerpoints—one with the heat plug and one without. Finally, the run order was not completely randomized. In addition to starting and ending with the standard operating conditions, for the sake of convenience the level of heat plug was only changed once. Thus, in effect we have blocked on heat plug. Since the heat plug main effect is negligible in all our analyses, this restriction seems to have had little effect on the outcome, and we ignore it.

For runs 2–11, a saturated model consists of the seven terms for a full factorial model, plus a term for curvature in the quantitative factors and the interaction of this curvature and heat plug. We fit this saturated model for the mean, $\ln(\text{Site-to-site Std. Dev.})$ and $\ln(\text{Std. Dev. of Wafer Means})$. The Lenth t statistics are shown in Table 14.10. We use the logarithmic transformation for two reasons when analyzing the standard deviations. First, if we had random sampling from a normal distribution, the log transformation stabilizes the variance (see Section 2.8.3). Second, the log transformation has the relative effect of spreading out the smaller standard deviations and reducing the differences among the larger standard deviations. Without this transformation, models tend to differentiate well what causes the worst variability while missing effects that distinguish the best treatment combinations from the moderately good ones. Note that for every response, the H_2/O_2 effect is the largest and is positive, so that low H_2/O_2 ratios minimize the variability. Note also that the H_2 main effect and the $H_2 * (H_2/O_2)$ interaction are similar for each response. Thus, the H_2 effect shows up most strongly at high ratios. At low H_2/O_2 , the effect of H_2 is negligible, as is evident in the profile plot in Figure 14.2.

Table 14.10. Lenth t statistics for Czitrom et al. (1998) oxide thickness data, runs 2–11

Term	Mean	Site-to-site $\ln(\text{Std. Dev.})$	$\ln(\text{Std. Dev.})$ of Wafer Means
H_2	0.65	-4.35	-2.08
H_2/O_2	5.19	10.43	4.30
$H_2 * (H_2/O_2)$	0.91	-5.27	-1.42
HeatPlug	-0.14	-0.51	0.64
$H_2 * \text{HeatPlug}$	-0.85	-0.83	0.35
$(H_2/O_2) * \text{HeatPlug}$	0.78	-0.22	-0.66
$H_2 * (H_2/O_2) * \text{HeatPlug}$	-0.64	0.10	-0.75
$(H_2/O_2)^2$	-0.69	3.02	0.67
$(H_2/O_2) * \text{HeatPlug} = \dots$	0.23	-0.90	-0.39

The preceding analysis ignored the first and last runs. Including these runs makes it possible to estimate both the H_2^2 and $(H_2/O_2)^2$ terms. However, it results in more correlated estimates and so a slightly more complicated analysis. Using all 12 runs confirms that the possible curvature evident in

Figure 14.2 is likely due to the ratio and not to H_2 . Furthermore, analyzing all 12 $\ln(\text{Site-to-site Std. Dev.})$ values gives some indication that including the heat plug does in fact lower the Site-to-site variability. This would become more apparent if we focused only on the Wafer positions at the heat plug end of the furnace.

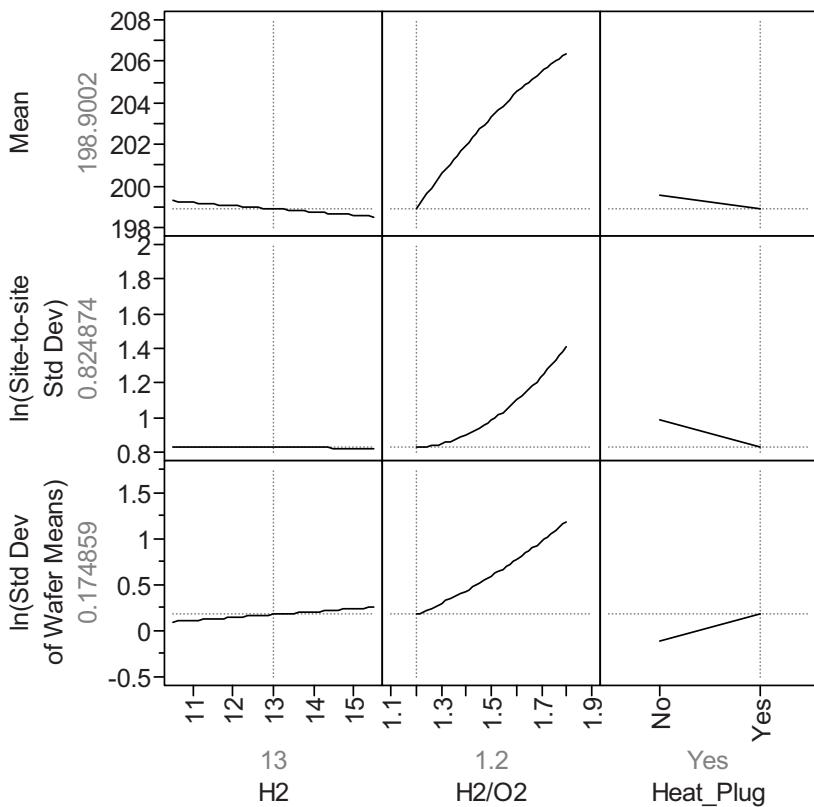


Fig. 14.2. Profile plot corresponding to models in Table 14.10

The central lesson from this example is that when structured samples are taken within each run, the analysis begins by calculating relevant summary statistics for each run. These summary statistics (means, standard deviations, correlations or regression coefficients) are then used as separate response variables. It is common to find factors that affect some but not all of the responses.

14.4 Generalized Least Squares Analysis When Variances Are Unequal

If the error variance is not constant, this impacts the analysis of the mean, since ordinary least squares is no longer fully efficient when fitting reduced models. An alternative is to use generalized least squares. Suppose that the random terms $\epsilon_1, \dots, \epsilon_N$ in our model (1.5) do not have the same variance. Let σ_i^2 denote the true variance for the i^{th} error. Assuming that the errors are distributed independently of one another, define

$$\Sigma = \text{Var} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_N^2 \end{bmatrix}.$$

The variance–covariance matrix for the ordinary least squares estimator (1.6) is thus

$$\text{Var}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\Sigma\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}. \quad (14.3)$$

For the case of diagonal Σ , the matrix (14.3) has equal diagonal elements (provided $\mathbf{X}'\mathbf{X} = N\mathbf{I}$), but the off-diagonal elements are not zero; that is, the regression coefficient estimators have the same standard error, but they are correlated. Furthermore, except for the special case of fitting a saturated model, a better estimator based on knowledge of the σ_i 's is possible.

The generalized least squares (GLS) estimator is given by

$$\mathbf{b}_{\text{gls}} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{Y}, \quad (14.4)$$

and its variance is

$$\text{Var}(\mathbf{b}_{\text{gls}}) = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}. \quad (14.5)$$

The diagonal elements of (14.5) are always less than or equal to the diagonal elements of (14.3). Note that this promise of a smaller variance requires knowledge of Σ , whereas we generally do not know Σ and must use an estimate. Thus, in practice, GLS may not be better because of uncertainty in the estimate of Σ . We discuss this more fully below, after considering an example.

For diagonal Σ , the GLS estimator is equivalent to the weighted least squares estimator, weighting by the reciprocal of the variances—although weighted least squares does not utilize (14.5) to calculate the standard errors of the estimates. However, many statistical software packages support weighted least squares, which does make calculation of the GLS estimates convenient. Observations with larger variances are weighted less in the GLS estimator (14.4), since their values are less reliable.

Starzec and Andersson (2002) investigated a fracture network model using a 2^3 factorial design. Their three factors were Orientation (**A**), Termination

fraction (**B**), and Fracture radii (**C**). At each of the eight treatment combinations, a computer model was run 10 times to compute the total unstable rock volume (in cubic meters). The mean and standard deviation for each set of 10 values are reported in Table 14.11. The response of interest is the mean volume. The squared standard deviation divided by 10 is an unbiased estimator for each σ_i^2 ; these estimates are reported in the last column of our table. Note that the first variance is nearly eight times the second variance. Thus, we have the equivalent of one-eighth as much data about the true mean at the first treatment combination versus the second.

Table 14.11. Starzec and Andersson (2002) 2^3 factorial for simulated unstable rock volume

A	B	C	Mean	Std. Dev.	$\hat{\sigma}_i^2$
-1	-1	-1	144	95	902.5
1	-1	-1	118	34	115.6
-1	1	-1	154	59	348.1
1	1	-1	132	54	291.6
-1	-1	1	142	47	220.9
1	-1	1	125	56	313.6
-1	1	1	157	48	230.4
1	1	1	179	51	260.1

We now fit a main effects model using ordinary least squares and generalized least squares. Ignoring the $\hat{\sigma}_i^2$ values, we compute the ordinary least squares estimates shown in Table 14.12. The MSE for the main effects model is 247.1, which is used by the software to compute the standard errors $(\text{MSE}/8)^{1/2} = 5.558$. For ordinary least squares, the standard errors based on (14.3) equal $[\bar{\sigma}^2/N]^{1/2}$, where $\bar{\sigma}^2$ is the average of the diagonal elements of Σ . Estimating this with the average of the last column of Table 14.11, one obtains a standard error of $(335.35/8)^{1/2} = 6.474$. This slightly larger standard error of 6.474 is preferred, since it is estimated with more precision than is the standard error based on the MSE. Thus, the t statistic for **B** is $11.625/6.474 = 1.80$, which is inconclusive about the effect of Termination fraction.

Since the software computing ordinary least squares estimates assumes equal variances, it does not recognize the loss of efficiency and the correlations among the estimators for the unequal variance case. Using our estimated variances in place of the unknown true variances, we compute the GLS estimates and their standard errors using (14.4) and (14.5); see Table 14.12. The standard errors are the square root of the diagonals of the covariance matrix

$$\text{Var}(\mathbf{b}_{\text{gls}}) = (X' \hat{\Sigma}^{-1} X)^{-1} = \begin{bmatrix} 32.7829 & -6.7089 & 2.0412 & -2.1716 \\ -6.7089 & 36.9546 & 5.1472 & 10.8189 \\ 2.0412 & 5.1472 & 32.7645 & -2.4145 \\ -2.1716 & 10.8189 & -2.4145 & 34.9358 \end{bmatrix}.$$

Note that these standard errors are all smaller than the standard error of 6.474 for the ordinary least squares estimates. More will be said momentarily about this apparent gain in efficiency.

Table 14.12. Starzec and Andersson (2002) 2^3 factorial for simulated unstable rock volume

Term	Ordinary Least Squares			Generalized Least Squares		
	Est.	Std Error	<i>t</i>	Est.	Std Error	<i>t</i>
Intercept	143.875	5.558	25.886	142.184	5.726	24.833
A	-5.375	5.558	-0.967	-2.817	6.079	-0.463
B	11.625	5.558	2.092	13.018	5.724	2.274
C	6.875	5.558	1.237	8.739	5.911	1.479

Std Error = 5.558 is based on MSE; replace with 6.474

If we fit a saturated model, the ordinary least squares estimators and the generalized least squares estimators will be identical, and each would have estimated standard error of 6.474; that is, as we add terms, the standard errors for the ordinary least squares estimators do not change, since the columns of \mathbf{X} are orthogonal, but the standard errors for the generalized least squares estimators will increase.

If we used software to compute weighted least squares estimates here, with weights equal to the reciprocal of the last column of Table 14.11, we obtain the GLS estimates in Table 14.12, but the reported standard errors are based on the MSE. The GLS standard errors are preferred because they are based on replication. By contrast, the MSE has only 4 df and its validity depends on the assumption of an additive model.

For this example, we have utilized sample variances, each estimated from 10 computer runs, in the GLS estimation. Ten replicates is marginally sufficient to use GLS. Carroll and Cline (1988) reported that using $n \geq 6$ replicates to estimate the variances results in a proportional increase of $(n-3)/(n-5)$ in the variance of the regression coefficients, relative to generalized least squares with known σ_i^2 , if the data are normally distributed. Thus, for the Starzec and Andersson (2002) example with $n = 10$, the estimated standard errors should be multiplied by $(7/5)^{1/2} = 1.18$ (i.e., increased by 18%). In light of this, we are not sure that using weighted least squares with estimated weights has increased the efficiency of our estimator here versus simply using ordinary least squares. The lesson is that unless one has more than 10 replicates (or variances that differ more dramatically), weighting produces little gain in precision. In Section 11.3 we mentioned the generalized least squares analysis by Rooda and van der Schijf (1982), which utilized sample variances computed from only 5 replicates. According to Carroll and Cline, it is pointless to use GLS based on sample variances from so few replicates. For more insight, see Carroll and Cline (1988) and Davidian and Haaland (1990).

14.5 Mixed-Model Analysis

Designs with randomization restrictions lead to multiple sources of variation, which, in turn, affect the analysis. Models that account for the factorial effects as well as these multiple sources of variation are referred to as *mixed models*. The factorial effects are represented as *fixed effects* (i.e., as unknown regression coefficients to be estimated), whereas each source of variation is represented as a *random effect* distributed with a mean zero and some unknown variance. We now revisit the various designs described in Chapter 3 and provide a unifying summary.

Consider the randomized block design. For this situation, we consider the mixed linear model

$$y_{ijk} = \mu_{ij} + \zeta_i + \eta_{ij}, \quad (14.6)$$

where μ_{ij} denotes the expected response, which will depend on the treatment combination for this run and the unknown β 's. In addition, we have two random terms representing two components. Each is assumed to be independently distributed with mean 0; their variances are denoted σ_ζ^2 and σ_η^2 , respectively. The first component ζ_i represents block-to-block differences. The second component η_{ij} is the additional within block variation associated with the individual units. For a completely randomized design, the error ϵ would be the sum of these two components and $\sigma^2 = \sigma_\zeta^2 + \sigma_\eta^2$. However, by grouping the experimental units into homogeneous blocks, we produce an arrangement where $\sigma_\eta^2 \ll \sigma^2$.

For a randomized complete block design, all of the factorial effect contrasts are orthogonal to blocks, and the block variance component does not affect the treatment comparisons. For the randomized incomplete block designs described in Section 3.3, some factorial effect contrasts are orthogonal to blocks, whereas others are completely confounded with blocks. For the model (14.6), the precision of effects orthogonal to blocks is a function of σ_η^2 and the number of runs, whereas effects confounded with blocks have a larger variance that depends on both σ_ζ^2 and σ_η^2 .

For the analysis of Example 3.1 in Section 3.4.1, the three factorial effects confounded with blocks (**ABC**, **ADE**, **BCDE**) had estimates smaller than most of the 31 factorial effects, so we reasoned that σ_ζ was negligible, so that the blocking restriction had little effect on the outcome.

For Example 3.2 analyzed in Section 3.4.2, **ABC** was confounded with blocks (i.e., with raw material lots). This design with four lots and four runs per lot results in 3 df between lots and 12 df within lots. Table 3.8 provides an ANOVA that partitions the 12 df within lots into the effects due to the six factorial effects orthogonal to blocks, and the remaining 6 df used to estimate σ_η^2 (using the notation of (14.6)). Assuming no three-factor interaction, the mean square for lots estimates $\sigma_\eta^2 + 4\sigma_\zeta^2$, since this is the variance of the sample mean for each lot. Thus, $(MS_{\text{lots}} - MSE)/4 = (182.062 - 7.396)/4 = 43.7$ can be taken as an estimate for σ_ζ^2 . This is a method of moments estimator,

which is based on the expected value of the mean squares. Statistical software that permits declaring Lots “random” may provide this variance component estimator. It may also compute the restricted maximum likelihood (REML) estimate, which in this case would be identical. REML estimates for variances are similar to maximum likelihood estimates but are less biased due to adjusting for the “degrees of freedom” used to estimate regression coefficients. For a useful reference, see Harville (1977) or Searle, Casella, and McCulloch (2006). Note that the tests for effects orthogonal to blocks are unaffected by the lot-to-lot variance component. Only the **ABC** estimate is affected. Here, with only 3 df for between lots (and with no two-factor interactions appearing important), we do not attempt to estimate **ABC**. One could in fact include **ABC** in the model and still estimate σ_ζ and use this to compute a standard error for b_{ABC} . To do this reliably would require more between-lot degrees of freedom (i.e., more lots). Instead, the important practical question here is to begin an investigation as to why the yield varies so much from lot to lot.

Examples 3.3 and 3.4 in Section 3.4 both involve eight blocks, and so block-to-block differences might be useful for estimating effects confounded with blocks. For Example 3.3, there are four replicates, and a different interaction is confounded with blocks in each replicate. The analysis in Section 3.4.3 treated blocks as fixed effects, and so each effect was estimated only using replicates for which it was not confounded with blocks. If we had fit a model containing interactions among the factors and had declared a model with random block effects, the estimates for interaction effects would be slightly different, since then the estimation would even incorporate the data where each interaction was confounded with blocks. Such an analysis would offer little or no gain in precision for this example, and so the analysis with random blocks was not considered. However, for Example 3.4, one interaction (Wire*Machine) is confounded with blocks (Days) in each replicate. Thus, only by treating Days as random can we obtain an estimate for this interaction. In Table 3.13, we fit a full factorial model in Replicate, Wire, and Machine and omitted Days. This allowed the 7 df for between Days to be partitioned into three sources, with the three-factor interaction mean square serving as an estimate for $\sigma_\eta^2 + 2\sigma_\zeta^2$, which is the appropriate error mean square for judging the significance of Wire*Machine. Note that the multiplier of σ_ζ^2 here is 2, because the Wire*Machine interaction is computed from the Day means and there are two runs per day. As noted in Section 3.4.4, the precision of the main effects, whose contrasts are orthogonal to Days, should be better than the precision for the Wire*Machine interaction. Here we have no evidence for Day-to-Day variation, since the error mean square involving Day-to-Day variation (2.78) is actually smaller than the error mean square based on within-day differences (6.65). Figure 14.3 shows a mixed-model analysis of these data. The fitted model is

$$y_{ijk} = \text{Rep}_i + \mu_{ijk} + \text{Block}_{j(i)} + \eta_{ijk}, \quad (14.7)$$

where $i = 1, 2, 3, 4$ denotes the replicate, $j = 1, 2$ denotes the Day within each Replicate, and $k = 1, 2$ denotes the two runs each Day. A full factorial model in Wire and Machine is assumed for the fixed effects μ_{ijk} . The analysis in Figure 14.3 allowed for unbounded REML variance component estimates, which in this case produced negative estimates for the variances of Day and Replicate. These match the expected mean square estimates. If instead we had specified bounded REML estimates, which precludes negative variance estimates, the Replicate and Day variance estimates would be zero and the analysis would match that of a completely randomized design, with the mean square error ($df = 12$) used to test each effect.

Response Sqrt(c)

Mean of Response	12.6132
Observations	16

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	12.6132	0.3194	39.49	<.0001
Type[Control]	2.4220	0.6445	3.76	0.0094
Machine[High Speed]	1.0532	0.6445	1.63	0.1534
Type[Control]*Machine[High Speed]	1.2691	0.4171	3.04	0.0557

REML Variance Component Estimates

Random Effect	Var Component	Std Error	95% Lower	95% Upper
Replicate	-0.2878	0.6587	-1.5789	1.0033
Day[Replicate]	-1.9315	2.2300	-6.3024	2.4393
Residual	6.6467	3.8375	2.7600	32.2306
Total	4.4274			

-2 LogLikelihood = 61.051

Fixed Effect Tests

Source	DF	DFDen	F Ratio	Prob > F
Type	1	6	14.1206	0.0094
Machine	1	6	2.6703	0.1534
Type*Machine	1	3	9.2578	0.0557

Fig. 14.3. Mixed-model analysis of Example 3.4

Model (14.7) is also appropriate for split-unit designs (i.e., for designs where one or more factors are confounded with blocks). Here we typically refer to the units confounded with certain main effects as whole units rather than blocks. For split-unit designs conducted in several replicates, we write our mixed model as

$$y_{ijk} = \mu_{ijk} + \text{Rep}_i + \zeta_{ij} + \eta_{ijk}, \quad (14.8)$$

where μ_{ijk} denotes the expected response based on the factorial effects, ζ_{ij} represents the error associated with whole units, and η_{ijk} is the error asso-

ciated with the split units. The variance for Rep_i does not affect any of the tests, assuming all factorial effect contrasts are orthogonal to Replicates.

For Example 3.5 analyzed in Section 3.5.2, there was only a single replicate, with 16 whole units of size 2. The whole units correspond to an unreplicated 2^4 factorial, so effect sparsity was assumed and Lenth's method used to estimate the standard error for the 15 whole-unit contrasts. Under the model (14.8), the standard error for each of these contrasts is

$$[(\sigma_\eta^2 + 2\sigma_\zeta^2)/32]^{1/2},$$

which is estimated by $PSE = 2.475$ (see Table 3.15). Similarly, we calculated $PSE = 0.216$ from the 16 factorial effect estimates that are orthogonal to whole units, to estimate $\sigma_\eta/(32)^{1/2}$, the standard error of the split-unit contrasts.

Mixed models can also be written to describe any of the designs with multiple blocking factors described in Section 3.6. For unreplicated designs, the analysis is typically based on Lenth's method, with PSEs computed from each set of contrasts that share a common standard error. For replicated designs, we may use REML to estimate the variance components rather than Lenth's method. This is more common for full factorial designs with few whole-unit factors. All of the split-unit and split-split-unit designs analyzed in this book are unreplicated. Example 3.6 analyzed in Section 3.6.2 is a replicated 2^3 with replicates (squares) and additional restrictions due to animals and periods. A mixed-model analysis is shown in Figure 14.4. The only difference between this and the ANOVA presented earlier in Table 3.18 is that we obtain an estimate for each variance component. As would be expected, the animal variance component is by far the largest. The three-factor interaction is confounded with Square. If we were to include this term in the model, the variance component for Square would be nonestimable. The fact that the Square variance estimate in Figure 14.4 is negative indicates the lack of evidence for both the three-factor interaction and Square-to-Square variation. The apparent difference between Squares indicated in Table 3.18 is because that test used an inappropriately small denominator mean square (i.e., one that did not include the animal or period variance components).

Finally, mixed models can be used to analyze data where Treatment*Block interactions are allowed, as was mentioned in Section 3.2. For randomized block designs, with Blocks and every Block*Factor interaction declared random, the mean square for each main effect is compared with its interaction with blocks. To do this effectively, one must have four or more blocks, since otherwise each denominator mean square will have too few degrees of freedom.

This section has just scratched the surface regarding the estimation of variance components and testing of effects for mixed models. Goos, Langhans, and Vandebroek (2006) provides a concise overview. For book-length coverage, see Searle, Casella and McCulloch (2006) and Littell et al. (2006).

Response Milk, kg/d**Summary of Fit**

Mean of Response	23.6397
Observations	32

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	23.6397	1.4503	16.30	0.0390
Grain[Coarse]	-0.5447	0.1813	-3.00	0.0110
Forage[Long]	0.0434	0.1813	0.24	0.8147
F/C[High F]	-0.5353	0.1813	-2.95	0.0121
Grain[Coarse]*Forage[Long]	-0.1459	0.1813	-0.81	0.4364
Grain[Coarse]*F/C[High F]	0.0453	0.1813	0.25	0.8068
Forage[Long]*F/C[High F]	-0.1278	0.1813	-0.71	0.4942

REML Variance Component Estimates

Random Effect	Var Component	Std Error	95% Lower	95% Upper
Square	-6.879	8.281	-23.11	9.35
Animal[Square]	39.332	22.860	-5.47	84.14
Period[Square]	4.749	2.896	-0.93	10.42
Residual	1.051	0.429	0.54	2.87
Total	38.253			

-2 LogLikelihood = 148.397

Fixed Effect Tests

Source	DF	DFDen	F Ratio	Prob > F
Grain	1	12	9.029	0.0110
Forage	1	12	0.057	0.8147
F/C	1	12	8.721	0.0121
Grain*Forage	1	12	0.648	0.4364
Grain*F/C	1	12	0.062	0.8068
Forage*F/C	1	12	0.497	0.4942

Fig. 14.4. Mixed-model analysis of Example 3.6

14.6 Highly Multivariate Response Data

When there are several response variables, typically we analyze each separately. For example, the Bermejo-Marrera et al. (2001) experiments for measuring trace elements in seafood were analyzed individually (see Example 6.6). However, sometimes a combined analysis can be done in addition to or in place of separate analyses.

Langsrud (2001) presented data from a screening experiment involving five ingredient and processing factors for baguettes:

- A: Flour quality
- B: Garlic content
- C: Mixing time
- D: Ascorbic acid
- E: Type of process

For a reason not described, the resolution IV 2^{5-1} design with $I = \mathbf{BCDE}$ was utilized. Fifteen sensory attributes, S2–S16, were evaluated by a panel of judges:

- S2: Glossiness
- S3: Crackles on the crust
- S4: Porosity
- S5: Increase in crust area due to cuts
- S6: Elasticity
- S7: Odor intensity
- S8: Fresh odor
- S9: Garlic odor
- S10: Flavor intensity
- S11: Fresh flavor
- S12: Salt flavor
- S13: Garlic flavor
- S14: Firmness
- S15: Moistness
- S16: Crispness of the crust

(In Langsrud 2001, S1 and S4 were identical by mistake, and so we simply omit S1.) Table 14.13 contains the treatment combinations and a summary of the panel responses to be described later. Table 14.14 shows the panel's mean value for each attribute/treatment combination.

Table 14.13. Langsrud (2001) 2^{5-1} factorial for baguettes, with four principal components

t.c.	A	B	C	D	E	PC1	PC2	PC3	PC4
1	-1	-1	-1	-1	-1	-1.11	-1.70	2.35	-1.82
2	-1	-1	-1	1	1	-3.39	-2.12	-0.44	0.57
3	-1	-1	1	-1	1	-2.31	-1.48	0.63	0.85
4	-1	-1	1	1	-1	-0.79	-0.13	1.91	1.64
5	-1	1	-1	-1	1	2.10	-3.24	-2.12	-1.95
6	-1	1	-1	1	-1	3.74	0.08	1.56	1.40
7	-1	1	1	-1	-1	2.96	-0.05	1.58	-0.46
8	-1	1	1	1	1	1.97	-1.69	-0.12	0.10
9	1	-1	-1	-1	-1	-0.91	1.18	0.80	-0.10
10	1	-1	-1	1	1	-2.53	0.32	0.11	-0.30
11	1	-1	1	-1	1	-3.52	0.87	-1.66	0.45
12	1	-1	1	1	-1	-0.86	2.78	2.01	-0.79
13	1	1	-1	-1	1	0.94	0.08	-2.77	2.64
14	1	1	-1	1	-1	1.62	2.82	-1.11	-0.18
15	1	1	1	-1	-1	-0.31	2.01	-2.23	-2.20
16	1	1	1	1	1	2.39	0.29	-0.48	0.17

Table 14.14. Sensory responses for Langsrud (2001) baguette experiment

t.c.	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16
1	1.27	3.22	2.63	2.91	6.57	4.89	6.36	1.18	5.15	5.76	3.49	1.13	5.02	4.44	3.68
2	1.24	3.42	1.59	6.02	5.94	4.79	5.72	1.15	4.82	5.60	3.09	1.19	5.51	4.06	4.67
3	1.41	3.67	1.59	6.32	6.25	5.07	6.08	1.02	4.99	5.91	3.17	1.03	5.25	4.13	5.45
4	1.59	4.87	2.66	5.85	6.54	5.38	6.30	1.05	5.30	6.11	3.34	1.02	5.32	4.60	5.39
5	1.23	3.05	1.39	2.55	5.66	7.62	6.08	7.64	7.44	5.98	3.10	7.76	4.65	4.26	4.37
6	1.86	4.71	3.33	4.31	6.69	7.31	6.46	6.87	6.92	6.72	3.46	6.71	5.17	5.00	5.52
7	1.87	4.46	2.96	3.96	6.57	6.81	6.37	6.20	6.35	6.30	3.52	5.81	4.47	4.65	5.27
8	1.21	2.91	2.24	5.87	6.10	7.13	6.63	7.03	6.97	6.16	3.35	6.82	5.16	4.74	5.55
9	1.26	4.94	4.17	4.17	6.60	5.30	6.31	1.16	5.73	5.83	3.18	1.40	4.91	4.53	7.32
10	1.20	4.11	2.78	5.78	6.44	5.09	6.04	1.12	5.30	5.18	3.33	1.36	5.22	4.22	6.34
11	1.22	4.61	3.94	6.83	6.19	5.34	5.54	1.10	5.21	5.25	2.89	1.16	5.15	3.98	6.81
12	1.35	3.78	5.44	6.51	7.07	5.45	6.10	1.17	5.63	5.63	3.64	1.45	4.95	4.67	7.15
13	1.48	5.68	2.87	5.56	5.48	7.12	5.99	6.99	6.85	5.99	3.11	6.42	5.53	4.39	7.59
14	1.46	5.53	3.79	6.64	6.74	7.26	5.72	6.83	7.04	5.50	3.59	7.04	4.79	4.51	7.13
15	1.10	4.38	4.82	3.45	6.61	6.89	5.42	5.80	6.57	5.26	3.44	5.78	5.06	4.05	7.16
16	1.42	4.37	2.97	5.29	6.47	7.64	6.30	7.63	7.15	5.75	3.43	7.04	5.15	4.78	5.93

One may use Lenth's method to analyze the 15 separate responses in Table 14.14; the results are summarized in Figure 14.5, where Lenth t statistics exceeding the 0.05 critical value of 2.156 are labeled. No effects are statistically significant for S2, S3, S5, S8, and S15. Factor **B** has the sole dominant effect for S7, S9, S10 and S13, the intensity and garlic measures for odor and flavor. The high level for flour (**A** = 1) increases both porosity and crispiness of the crust. The low level for Process (**E** = -1) enhances porosity, elasticity, and salt flavor. Furthermore, when **A** = 1, high mixing time (**C** = 1) enhances porosity. Factor **D** appears only for salt flavor and firmness. Which of each aliased pair (**BE** = **CD**, **DE** = **BC**, and **ABD** = **ACE**) is considered more likely might well be resolved by the experts.

The analysis summarized in Figure 14.5 involves 240 tests and hence is prone to making numerous Type I errors. Using a more stringent criterion would not likely change the conclusions here. However, it is possible to reduce the number of tests by analyzing only a few linear combinations of the original response variables. We now illustrate that for the baguette data.

Principal components is a multivariate statistical technique based on an eigenanalysis of the correlation matrix of the response data matrix. For details, see Härdle and Simar (2007, Ch. 9) or some other book on applied multivariate statistics. Table 14.13 reports the first four principal components, which together account for 85% of the variation in the sensory variables. By computing the correlations between the sensory responses S2–S16 and the principal components PC1–PC4, you can see which responses are most represented in each component; see Table 14.15. Note that PC1 is most correlated with S7, S9, S10, and S13, the four responses affected so strongly by **B**. PC2 is dom-

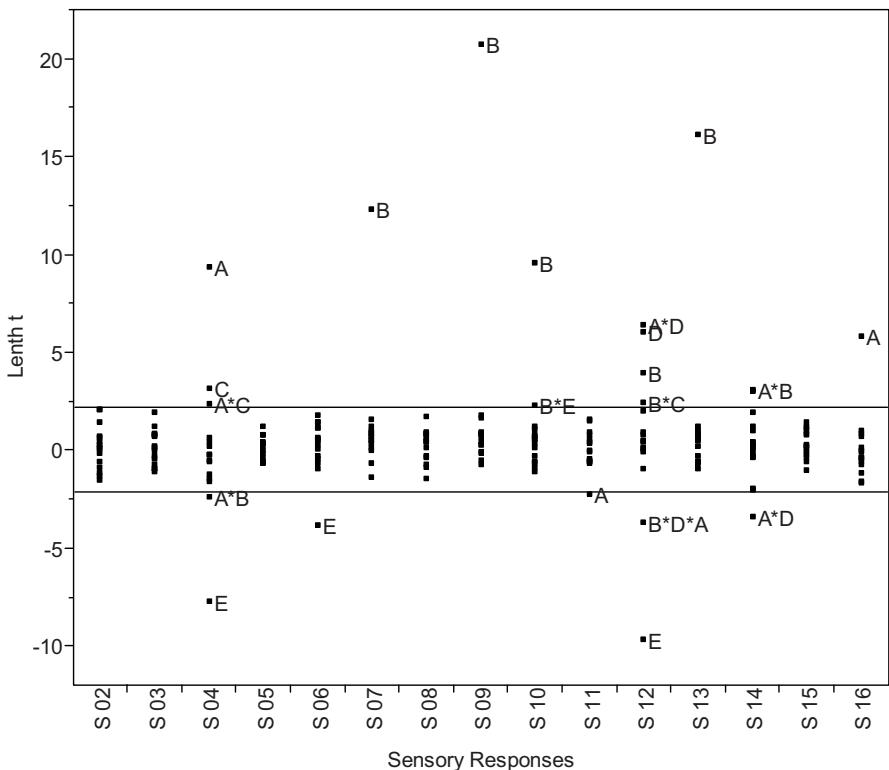


Fig. 14.5. Lenth t statistics for 15 sensory responses

inated by S4 and S16. Figure 14.6 shows the Lenth t statistics when we take PC1–PC4 as the responses. Given the correlations in Table 14.15, it is no surprise that **B** is a dominant effect for PC1, whereas factors **A** and **E**, and to a lesser extent **C**, affect PC2. For data such as in Table 14.14 where a few factors account for most of the variation, we expect to see their effects concentrated in a few principal components. As is the case here, that simplifies the analysis.

This example illustrates the type of methods that are useful for analyzing two-level designs with many responses. Langsrud (2001) presented other methods that utilize principal components in combination with a step-down test procedure. The article discussed a second example based on a 2^3 factorial with centerpoint replicates for mayonnaise production. The response for each run is based on reflectance spectroscopy, with absorbance values at 350 equally spaced wavelengths. Here one would forego the individual analyses and only analyze a limited number of functions of the 350 responses. One difference between the baguette and mayonnaise examples is that the mayonnaise responses are ordered (according to wavelength). For such cases, other methods

for functional response and repeated measures data become relevant; see the concluding paragraph of Section 2.8.4.

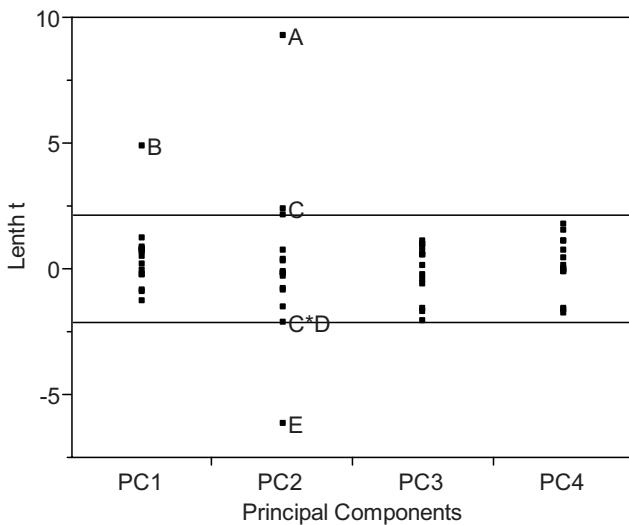


Fig. 14.6. Lenth t statistics for first four principal components

Table 14.15. Correlations between sensory responses and the first four principal components for baguette data

	PC1	PC2	PC3	PC4
S02	0.578	0.078	0.431	0.473
S03	0.156	0.655	-0.204	0.514
S04	-0.011	0.910	0.065	-0.195
S05	-0.391	0.349	-0.035	0.569
S06	0.051	0.658	0.648	-0.285
S07	0.883	0.053	-0.445	-0.025
S08	0.503	-0.349	0.633	0.170
S09	0.877	-0.038	-0.453	-0.027
S10	0.879	0.064	-0.434	-0.088
S11	0.675	-0.364	0.395	0.384
S12	0.474	0.449	0.509	-0.310
S13	0.873	-0.019	-0.462	-0.066
S14	-0.449	-0.143	-0.156	0.635
S15	0.746	0.146	0.523	0.219
S16	-0.063	0.821	-0.379	0.237

14.7 Four Analysis Blunders to Avoid

Analysis of two-level designs is generally quite straightforward, because the effect estimates are uncorrelated. The primary analysis difficulty relates to the choice of an estimator for the error variance σ^2 , which is necessary for calculating standard errors for the factorial effect estimates. Statistical software does not make this foolproof. Here are four errors to avoid regarding the computation of standard errors for the estimates.

1. Make sure that replication captures all the sources of run-to-run variation

Replication is not always necessary. However, if replication is utilized to provide a pure error estimate of the error variance, make sure that you have true replication, not pseudo-replication. True replication requires that two observations at the same treatment combination be performed just as independently of one another as are observations at different treatment combinations. If not, then the pure error mean square will likely underestimate σ^2 , because the “pseudo-replicates” share some of the same error in common.

Lai, Pan, and Tzeng (2003) reported conducting two lovastatin production experiments. Each involved eight distinct treatment combinations, and the authors claimed that each was performed as a completely randomized 24-run design. Our analysis in Section 6.1 assumed that the measurements were in fact from true replicates. However, Figure 6.2 exhibits little variation among “replicate” observations. Such data should at least raise the question as to how the replicates were performed.

Yuan, Murch, and Goss (2003) conducted a 2^{7-4} screening experiment to identify which factors influence thermal transmittance measurements. Smaller values of the “U-factor” response indicate less heat loss. The order of the treatment combinations was randomized. Then, for each, data were collected from four consecutive runs, each lasting 4 hours. There is no problem with this design if one analyzes the eight mean values (or eight $\ln(s)$ values) using Lenth’s method. However, if the repeat runs are used to estimate the run-to-run error variance, this is a bad design—one that will underestimate the true error variance if there is any positive autocorrelation in the errors. It would have been better to have done just 16 or 24 runs with complete randomization of order than to perform these 32 runs grouped by treatment combination.

2. Avoid basing your estimation of the error variance on only 1 or 2 degrees of freedom

This warning applies whether the few degrees of freedom are based on pure error or omitted terms. For instance, Example 7.3 is a 16-run fractional factorial with $n_0 = 2$ centerpoint replicates. If one fits a saturated model, most software will use the mean square error with 1 df to perform tests. The large t

percentiles in Appendix A for 1 df indicates how imprecise are standard errors based on 1 df. If effect sparsity is reasonable, it is preferable to incorporate 1 or 2 df for pure error into Lenth's method, either as described by Edwards and Mee (2008) or by treating the pure error as "null effect" contrasts as JMP's screening platform does. Su and Lua (2006) analyzed their 2^{4-1} by dropping the smallest interaction estimate from the model and using this 1 df for error. Such an analysis violates both this and the next warning.

3. Remember that the mean square error for a reduced model can seriously underestimate the error variance

Consider first the unfortunate case where no effects are present. Thus, the true model is $y = \beta_0 + \epsilon$, with $\text{Var}(\epsilon) = \sigma^2$ to be estimated from an experiment without replication. To illustrate the potential for bias in the mean square error of a reduced model, consider the case where the largest estimates are always considered significant. We compute the mean square error for the reduced model that retains no more than 13% of the estimates. The expected value of the MSE/σ^2 is

No. of Effects			
<i>N</i>	in Model	df for Error	$E(\text{MSE}/\sigma^2)$
8	1	6	0.623
16	1	14	0.752
	2	13	0.598
32	1	30	0.845
	2	29	0.739
	4	27	0.586

This illustrates the dramatic bias to the MSE that results from just a few Type I errors. If one declares the largest two effects statistically significant, out of 15, the MSE is on average less than 60% of σ^2 .

Suppose instead that we use stepwise regression to select a model, sequentially adding terms until we obtain a p -value greater than α . Simulation results are as follows for $\alpha = .05$ and a true model with no active effects:

Mean No. of			
<i>N</i>	EER	Effects Added	$E(\text{MSE}/\sigma^2)$
8	0.35	0.35	0.624
16	0.66	0.71	0.746
32	0.91	1.60	0.784

The EER is the probability that the (simple regression) t statistic for the largest estimate will exceed $t_{.05/2,N-2}$ in absolute value, since adding one or more terms constitutes a Type I error. While the average number of Type I errors is not excessive, the mean square error still seriously underestimates σ^2 .

To base the error variance on only a small number of the smallest estimates greatly exacerbates the problem. Su and Lua (2006) used an MSE based on just the least significant estimate; such an estimate is both biased and imprecise.

For this reason, when we do not have sufficient replication to estimate σ^2 , Section 2.4 proposed using either Lenth's method, provided effect sparsity is reasonable, or using the MSE from a model assumed prior to data analysis.

The simulation results reported above are for the case where all β_i 's are zero. When some β_i 's are not zero, Type I errors (including in the model effects with $\beta_i = 0$) and Type II errors (leaving in error effects for which $\beta_i \neq 0$) can tend to cancel one another. The resulting MSE will be biased, but the direction of the bias depends on the magnitude of the true β_i 's. Conceptually, we have a mixture of distributions for the b_i 's. For an orthogonal design, the b_i 's are independent normal random variables, each with its own mean β_i and common variance σ^2/N . Without replication or some prior knowledge regarding σ^2 , one looks for a few b_i 's that stand out as much larger than the rest. The inference procedures presented in Section 2.4 assume that certain β_i 's, or a preponderance of β_i 's, are zero. If we use these procedures to identify significant effects, but most β_i 's are not zero, we are, in effect, concluding that their true coefficients are both larger than the majority of other β_i 's, and larger than their standard error $\sigma/N^{1/2}$. For deterministic computer models, $\sigma = 0$, and so model selection simply becomes a choice between parsimony and accuracy.

4. Recognize randomization restrictions when conducting tests for effects

Randomization restrictions such as blocking and split-unit structures generally make a design easier to conduct. However, they make the analysis more complex and require more understanding by the data analyst to obtain the correct results. Effects are often estimated the same way for these designs as for completely randomized designs, but the standard errors for the estimates depend on the unit structure for the design.

Conclusion

Plotting both the data and the effect estimates is always recommended. Generally, an effective plot and an intuitive understanding of the analysis method will keep one from mistaken claims of significance. A well-planned experiment should be relatively easy to analyze, especially given the aid of statistical software. It is the author's hope that you will be more successful in your work because of the widespread utility of factorial experimentation.

Part IV

Appendices and Tables

A

Upper Percentiles of t Distributions, t_α

$P(t_{\text{df}} > t_\alpha) = \alpha$ and $P(|t_{\text{df}}| > t_\alpha) = 2\alpha$

df	$\alpha = .20$.10	.075	.05	.025	.01	.005	.0025	.001	
1	1.376	3.078	4.165	6.314	12.706	31.821	63.657	127.321	318.309	
2	1.061	1.886	2.282	2.920	4.303	6.965	9.925	14.089	22.327	
3	0.978	1.638	1.924	2.353	3.182	4.541	5.841	7.453	10.215	
4	0.941	1.533	1.778	2.132	2.776	3.747	4.604	5.598	7.173	
5	0.920	1.476	1.699	2.015	2.571	3.365	4.032	4.773	5.893	
6	0.906	1.440	1.650	1.943	2.447	3.143	3.707	4.317	5.208	
7	0.896	1.415	1.617	1.895	2.365	2.998	3.499	4.029	4.785	
8	0.889	1.397	1.592	1.860	2.306	2.896	3.355	3.833	4.501	
9	0.883	1.383	1.574	1.833	2.262	2.821	3.250	3.690	4.297	
10	0.879	1.372	1.559	1.812	2.228	2.764	3.169	3.581	4.144	
11	0.876	1.363	1.548	1.796	2.201	2.718	3.106	3.497	4.025	
12	0.873	1.356	1.538	1.782	2.179	2.681	3.055	3.428	3.930	
13	0.870	1.350	1.530	1.771	2.160	2.650	3.012	3.372	3.852	
14	0.868	1.345	1.523	1.761	2.145	2.624	2.977	3.326	3.787	
15	0.866	1.341	1.517	1.753	2.131	2.602	2.947	3.286	3.733	
16	0.865	1.337	1.512	1.746	2.120	2.583	2.921	3.252	3.686	
17	0.863	1.333	1.508	1.740	2.110	2.567	2.898	3.222	3.646	
18	0.862	1.330	1.504	1.734	2.101	2.552	2.878	3.197	3.610	
19	0.861	1.328	1.500	1.729	2.093	2.539	2.861	3.174	3.579	
20	0.860	1.325	1.497	1.725	2.086	2.528	2.845	3.153	3.552	
21	0.859	1.323	1.494	1.721	2.080	2.518	2.831	3.135	3.527	
22	0.858	1.321	1.492	1.717	2.074	2.508	2.819	3.119	3.505	
23	0.858	1.319	1.489	1.714	2.069	2.500	2.807	3.104	3.485	
24	0.857	1.318	1.487	1.711	2.064	2.492	2.797	3.091	3.467	
25	0.856	1.316	1.485	1.708	2.060	2.485	2.787	3.078	3.450	
30	0.854	1.310	1.477	1.697	2.042	2.457	2.750	3.030	3.385	
35	0.852	1.306	1.472	1.690	2.030	2.438	2.724	2.996	3.340	
40	0.851	1.303	1.468	1.684	2.021	2.423	2.704	2.971	3.307	
45	0.850	1.301	1.465	1.679	2.014	2.412	2.690	2.952	3.281	
50	0.849	1.299	1.462	1.676	2.009	2.403	2.678	2.937	3.261	
60	0.848	1.296	1.458	1.671	2.000	2.390	2.660	2.915	3.232	
80	0.846	1.292	1.453	1.664	1.990	2.374	2.639	2.887	3.195	
100	0.845	1.290	1.451	1.660	1.984	2.364	2.626	2.871	3.174	
∞	0.842	1.282	1.440	1.645	1.960	2.326	2.576	2.807	3.090	
	$2\alpha =$.40	.20	.15	.10	.05	.02	.01	.005	.002

B**Upper Percentiles of F Distributions, F_α**

$F_{.10}$ critical values, where $P(F_{df_1, df_2} > F_{.10}) = 0.10$

df ₂	df ₁										
	1	2	3	4	5	6	8	10	15	30	60
1	39.86	49.50	53.59	55.83	57.24	58.20	59.44	60.19	61.22	62.26	62.79
2	8.53	9.00	9.16	9.24	9.29	9.33	9.37	9.39	9.42	9.46	9.47
3	5.54	5.46	5.39	5.34	5.31	5.28	5.25	5.23	5.20	5.17	5.15
4	4.54	4.32	4.19	4.11	4.05	4.01	3.95	3.92	3.87	3.82	3.79
5	4.06	3.78	3.62	3.52	3.45	3.40	3.34	3.30	3.24	3.17	3.14
6	3.78	3.46	3.29	3.18	3.11	3.05	2.98	2.94	2.87	2.80	2.76
7	3.59	3.26	3.07	2.96	2.88	2.83	2.75	2.70	2.63	2.56	2.51
8	3.46	3.11	2.92	2.81	2.73	2.67	2.59	2.54	2.46	2.38	2.34
9	3.36	3.01	2.81	2.69	2.61	2.55	2.47	2.42	2.34	2.25	2.21
10	3.29	2.92	2.73	2.61	2.52	2.46	2.38	2.32	2.24	2.16	2.11
11	3.23	2.86	2.66	2.54	2.45	2.39	2.30	2.25	2.17	2.08	2.03
12	3.18	2.81	2.61	2.48	2.39	2.33	2.24	2.19	2.10	2.01	1.96
13	3.14	2.76	2.56	2.43	2.35	2.28	2.20	2.14	2.05	1.96	1.90
14	3.10	2.73	2.52	2.39	2.31	2.24	2.15	2.10	2.01	1.91	1.86
15	3.07	2.70	2.49	2.36	2.27	2.21	2.12	2.06	1.97	1.87	1.82
16	3.05	2.67	2.46	2.33	2.24	2.18	2.09	2.03	1.94	1.84	1.78
17	3.03	2.64	2.44	2.31	2.22	2.15	2.06	2.00	1.91	1.81	1.75
18	3.01	2.62	2.42	2.29	2.20	2.13	2.04	1.98	1.89	1.78	1.72
19	2.99	2.61	2.40	2.27	2.18	2.11	2.02	1.96	1.86	1.76	1.70
20	2.97	2.59	2.38	2.25	2.16	2.09	2.00	1.94	1.84	1.74	1.68
21	2.96	2.57	2.36	2.23	2.14	2.08	1.98	1.92	1.83	1.72	1.66
22	2.95	2.56	2.35	2.22	2.13	2.06	1.97	1.90	1.81	1.70	1.64
23	2.94	2.55	2.34	2.21	2.11	2.05	1.95	1.89	1.80	1.69	1.62
24	2.93	2.54	2.33	2.19	2.10	2.04	1.94	1.88	1.78	1.67	1.61
25	2.92	2.53	2.32	2.18	2.09	2.02	1.93	1.87	1.77	1.66	1.59
30	2.88	2.49	2.28	2.14	2.05	1.98	1.88	1.82	1.72	1.61	1.54
35	2.85	2.46	2.25	2.11	2.02	1.95	1.85	1.79	1.69	1.57	1.50
40	2.84	2.44	2.23	2.09	2.00	1.93	1.83	1.76	1.66	1.54	1.47
50	2.81	2.41	2.20	2.06	1.97	1.90	1.80	1.73	1.63	1.50	1.42
60	2.79	2.39	2.18	2.04	1.95	1.87	1.77	1.71	1.60	1.48	1.40
70	2.78	2.38	2.16	2.03	1.93	1.86	1.76	1.69	1.59	1.46	1.37
80	2.77	2.37	2.15	2.02	1.92	1.85	1.75	1.68	1.57	1.44	1.36
100	2.76	2.36	2.14	2.00	1.91	1.83	1.73	1.66	1.56	1.42	1.34
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.67	1.60	1.49	1.34	1.24

$F_{.05}$ critical values, where $P(F_{df_1, df_2} > F_{.05}) = 0.05$

df ₂	df ₁									
	1	2	3	4	5	6	8	10	15	30
1	161.45	199.50	215.71	224.58	230.16	233.99	238.88	241.88	245.95	250.10
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.40	19.43	19.46
3	10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.79	8.70	8.62
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.96	5.86	5.75
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.74	4.62	4.50
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.06	3.94	3.81
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.64	3.51	3.38
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.35	3.22	3.08
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.14	3.01	2.86
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.98	2.85	2.70
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.85	2.72	2.57
12	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.75	2.62	2.47
13	4.67	3.81	3.41	3.18	3.03	2.92	2.77	2.67	2.53	2.38
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.60	2.46	2.31
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.54	2.40	2.25
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.49	2.35	2.19
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.45	2.31	2.15
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.41	2.27	2.11
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.38	2.23	2.07
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.35	2.20	2.04
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.32	2.18	2.01
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.30	2.15	1.98
23	4.28	3.42	3.03	2.80	2.64	2.53	2.37	2.27	2.13	1.96
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.25	2.11	1.94
25	4.24	3.39	2.99	2.76	2.60	2.49	2.34	2.24	2.09	1.92
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.16	2.01	1.84
35	4.12	3.27	2.87	2.64	2.49	2.37	2.22	2.11	1.96	1.79
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.08	1.92	1.74
45	4.06	3.20	2.81	2.58	2.42	2.31	2.15	2.05	1.89	1.71
50	4.03	3.18	2.79	2.56	2.40	2.29	2.13	2.03	1.87	1.69
60	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.99	1.84	1.65
70	3.98	3.13	2.74	2.50	2.35	2.23	2.07	1.97	1.81	1.62
80	3.96	3.11	2.72	2.49	2.33	2.21	2.06	1.95	1.79	1.60
100	3.94	3.09	2.70	2.46	2.31	2.19	2.03	1.93	1.77	1.57
∞	3.84	3.00	2.60	2.37	2.21	2.10	1.94	1.83	1.67	1.46

$F_{.01}$ critical values, where $P(F_{df_1, df_2} > F_{.01}) = 0.01$

df ₂	df ₁									
	1	2	3	4	5	6	8	10	15	30
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.40	99.43	99.47
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.23	26.87	26.50
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.55	14.20	13.84
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	10.05	9.72	9.38
6	13.75	10.92	9.78	9.15	8.75	8.47	8.10	7.87	7.56	7.23
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.62	6.31	5.99
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.81	5.52	5.20
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.26	4.96	4.65
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.85	4.56	4.25
11	9.65	7.21	6.22	5.67	5.32	5.07	4.74	4.54	4.25	3.94
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.30	4.01	3.70
13	9.07	6.70	5.74	5.21	4.86	4.62	4.30	4.10	3.82	3.51
14	8.86	6.51	5.56	5.04	4.69	4.46	4.14	3.94	3.66	3.35
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.80	3.52	3.21
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.69	3.41	3.10
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.59	3.31	3.00
18	8.29	6.01	5.09	4.58	4.25	4.01	3.71	3.51	3.23	2.92
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.43	3.15	2.84
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.37	3.09	2.78
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.31	3.03	2.72
22	7.95	5.72	4.82	4.31	3.99	3.76	3.45	3.26	2.98	2.67
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.21	2.93	2.62
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.17	2.89	2.58
25	7.77	5.57	4.68	4.18	3.85	3.63	3.32	3.13	2.85	2.54
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.98	2.70	2.39
35	7.42	5.27	4.40	3.91	3.59	3.37	3.07	2.88	2.60	2.28
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.80	2.52	2.20
45	7.23	5.11	4.25	3.77	3.45	3.23	2.94	2.74	2.46	2.14
50	7.17	5.06	4.20	3.72	3.41	3.19	2.89	2.70	2.42	2.10
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.63	2.35	2.03
70	7.01	4.92	4.07	3.60	3.29	3.07	2.78	2.59	2.31	1.98
80	6.96	4.88	4.04	3.56	3.26	3.04	2.74	2.55	2.27	1.94
100	6.90	4.82	3.98	3.51	3.21	2.99	2.69	2.50	2.22	1.89
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.51	2.32	2.04	1.70

C

Upper Percentiles for Lenth t Statistics, c_α^{IER} and c_α^{EER}

Individual error rate: $P(|b_s|/PSE > c_\alpha^{\text{IER}}) = \alpha$

m	α										
	.20	.15	.10	.08	.06	.05	.04	.03	.02	.01	
7	1.202	1.406	1.710	1.888	2.132	2.297	2.536	3.086	3.752	5.065	
11	1.240	1.433	1.709	1.866	2.074	2.211	2.383	2.719	3.212	4.073	
15	1.256	1.441	1.701	1.846	2.034	2.156	2.308	2.521	2.955	3.625	
19	1.264	1.444	1.694	1.831	2.008	2.121	2.261	2.443	2.808	3.380	
23	1.268	1.446	1.688	1.820	1.989	2.096	2.227	2.397	2.711	3.224	
27	1.271	1.446	1.684	1.812	1.975	2.078	2.203	2.365	2.643	3.120	
31	1.274	1.446	1.680	1.805	1.964	2.064	2.185	2.341	2.593	3.044	
35	1.275	1.446	1.677	1.800	1.955	2.052	2.170	2.321	2.551	2.986	
39	1.276	1.446	1.674	1.796	1.949	2.044	2.160	2.307	2.520	2.942	
43	1.277	1.446	1.672	1.792	1.943	2.037	2.151	2.295	2.496	2.906	
47	1.278	1.446	1.671	1.789	1.938	2.031	2.143	2.284	2.481	2.876	
63	1.280	1.445	1.665	1.781	1.925	2.014	2.121	2.256	2.442	2.797	
127	1.282	1.444	1.657	1.767	1.905	1.989	2.089	2.215	2.385	2.685	
255	1.283	1.443	1.652	1.761	1.894	1.976	2.073	2.194	2.357	2.632	
511	1.283	1.442	1.650	1.757	1.889	1.970	2.065	2.184	2.344	2.606	
∞	1.284	1.442	1.648	1.754	1.884	1.963	2.057	2.174	2.330	2.580	

Note: These critical values are for two-sided tests.

For $m > 20$ contrasts, the .05 critical value is well approximated by $1.9633 + 3.33/m - 6.66/m^2$. The following MATLAB® code may be used to estimate p -values for Lenth t statistics via simulation for orthogonal model matrices.

```
% Simulation for computing Lenth's t statistic p-values
tobs=[ ]; pvalue=[ ];
m = input('What is the number of contrasts, m? ');
tobs(1) = input('Input the m observed t statistics, one at a time ');
j = 1;
while j < m
j=j+1;
tobs(j) = input('Next observed t? ');
end
ns = input('How many simulated samples? (>=40,000 recommended)');
c=zeros(1,m);
for i=1:ns
Z=normrnd(0,1,m,1); Z=sort(abs(Z));
Upper=1.5*median(Z)*2.5;
test=m;
while Z(test,1)>Upper
test=test-1;
end
PSE=1.5*median(Z(1:test));
t=Z/PSE;
for i=1:m
j=m;
while j>0 & t(j)> tobs(i)
j=j-1;
end
c(i) = c(i) + m - j;
end
end
for i=1:m
pvalue(i)=c(i)/(m*ns);
end
[tobs', pvalue']
```

Experimentwise error rate: $P(\max\{|b_1|, |b_2|, \dots, |b_m|\}/\text{PSE} > c_\alpha^{\text{EER}}) = \alpha$

m	α									
	.50	.40	.30	.25	.20	.15	.10	.05	.03	.01
7	1.608	1.803	2.055	2.217	2.420	3.046	3.688	4.867	6.023	9.714
11	1.828	2.015	2.248	2.393	2.738	3.104	3.563	4.452	5.251	7.448
15	1.967	2.146	2.363	2.519	2.843	3.121	3.505	4.233	4.846	6.423
19	2.068	2.239	2.445	2.685	2.894	3.134	3.477	4.105	4.612	5.857
23	2.146	2.311	2.546	2.751	2.928	3.147	3.459	4.019	4.457	5.498
27	2.210	2.370	2.643	2.793	2.956	3.161	3.452	3.962	4.354	5.264
31	2.263	2.419	2.694	2.827	2.981	3.176	3.448	3.919	4.276	5.091
35	2.309	2.461	2.731	2.856	3.003	3.189	3.447	3.890	4.219	4.961
39	2.350	2.509	2.762	2.882	3.024	3.203	3.449	3.866	4.175	4.867
43	2.385	2.580	2.788	2.904	3.042	3.215	3.452	3.850	4.142	4.786
47	2.418	2.618	2.812	2.926	3.060	3.227	3.455	3.835	4.114	4.726
63	2.554	2.715	2.890	2.994	3.116	3.267	3.470	3.805	4.045	4.560
127	2.799	2.928	3.074	3.161	3.261	3.383	3.546	3.808	3.992	4.373
255	3.015	3.127	3.254	3.328	3.413	3.517	3.656	3.876	4.030	4.345
511	3.218	3.317	3.430	3.495	3.572	3.664	3.787	3.982	4.118	4.397

Note: c_α^{EER} values for $m = 255$ (511) are based on 9 (1.8) million samples.

To simulate simultaneous p -values, replace the last 13 lines of the code for IER p -values with:

```
t=Z(m)/PSE;
for i=1:m
if t(j)> tobs(i)
c(i) = c(i) + 1;
end
end
end
for i=1:m
pvalue(i)=c(i)/ns;
end
[tobs', pvalue']
```

D

Computing Upper Percentiles for Maximum Studentized Residual

The largest Studentized residual is useful for determining whether an experiment contains a single outlier. The null distribution of this statistic, assuming a correctly specified model and normally distributed errors with common variance, is easily simulated. The Studentized residuals follow a multivariate t distribution with a correlation matrix that depends on \mathbf{X} .

The model matrix \mathbf{X} below and the maximum Studentized residual $rmax = 2.0997$ pertain to the example in Section 2.6.4. Output from this program is $p90 = 2.415$, $p95 = 2.531$, $Prob = 0.395$; that is, for this situation, there is a 39.5% chance of getting one or more Studentized residuals larger than the value 2.0997, which we observed. To be considered unusual, the largest Studentized residual would need to exceed 2.415, the estimated 90th percentile, or 2.531, the 95th percentile. For your application, simple replace \mathbf{X} and $rmax$.

```
%Distribution of largest |Studentized residual|, given X
iter=100000
rmax=2.0997
X =[ 1 1 -1 -1 -1 1;
1 1 1 -1 1 1;
1 -1 -1 -1 -1 1;
1 1 -1 1 -1 1;
1 -1 1 1 -1 -1;
1 1 1 1 1 1;
1 -1 -1 1 -1 1;
1 1 1 1 -1 -1;
1 1 -1 1 1 -1;
1 1 1 -1 -1 -1;
1 -1 1 1 1 1;
1 -1 1 -1 -1 -1;
1 1 -1 -1 1 -1;
1 -1 1 -1 1 1;
1 -1 -1 1 1 -1;
```

```
1 -1 -1 -1 1 -1 ];
[N,r]=size(X); mu=zeros(N,1); var=ones(N,1);
L=eye(N)-X*X'/N;
count=0;
for i=1:iter
Z=normrnd(mu,var,N,1); R=L*Z; t(i)=max(abs(R))/sqrt(Z'*R/N);
if t(i)>rmax
count=count+1;
end
end
t=sort(t);
%The 90th and 95th percentiles
p90=t(.90*iter)
p95=t(.95*iter)
% Prob[max |studentized residual| > rmax]
Prob=count/iter
```

E

Orthogonal Blocking for Full 2^k Factorial Designs

The following table presents generators for partitioning a 2^k factorial design into 2^b blocks, confounding $2^b - 1$ interactions optimally according to the minimum aberration criterion of Sun, Wu, and Chen (1997). For each design, we underline b interactions that may be used to generate the blocks. If the estimability is “ $e.f$ ”, then one can estimate all effects of order “ e ” and a fraction “ $.f$ ” of order $e + 1$.

No. Factors	No. Blocks	Block Size	Order of Estimability	Confounded Interactions
3	2	4	2	<u>ABC</u>
	4	2	1	<u>AB</u> , <u>AC</u> , BC
4	2	8	3	<u>ABCD</u>
	4	4	1.83	<u>ABC</u> , <u>ABD</u> , CD
	8	2	1	<u>AB</u> , <u>AC</u> , BC, <u>AD</u> , BD, CD, ABCD

Example: 2^4 in four blocks of size 4. Using **ABC** and **ABD** to generate the blocks, we also confound with blocks **ABC(ABD) = CD**. Since all main effects and five of six two-factor interactions are clear of confounding with blocks, the order of estimability is 1.83. Using the formula $0.5(5+\mathbf{ABC}+2\mathbf{ABD})$ to create the block number, the partitioning of the design is

Block 1 ABC = ABD = -1	Block 2 ABC = -ABD = 1	Block 3 -ABC = ABD = 1	Block 4 ABC = ABD = 1
----	--+-	-+-+	-++-
+---	+--+	+--+	+---
-+++	-+-+	-++-	-+--
++--	+++-	++-+	++++

No. Factors	No. Blocks	Block Size	Order of Estimability	Confounded Interactions
5	2	16	4	<u>ABCDE</u>
	4	8	2.8	<u>ABC</u> , <u>ADE</u> , BCDE
	8	4	1.8	<u>ABC</u> , <u>ADE</u> , BCDE, <u>BD</u> , ACD, ABE, CE
	16	2	1	All 2- and 4-factor interactions
6	2	32	5	<u>ABCDEF</u>
	4	16	3.8	<u>ABCE</u> , <u>ABDF</u> , CDEF
	8	8	2.8	<u>ABCE</u> , <u>ABDF</u> , CDEF, <u>ACD</u> , BDE, BCF, AEF
	16	4	1.8	<u>AB</u> , <u>CD</u> , ABCD, <u>ACE</u> , BCE, ADE, BDE, <u>ACE</u> , BCF, ADF, BDF, EF, ABEF, CDEF, ABCDEF
	32	2	1	All 2-, 4- and 6-factor interactions
7	2	64	6	<u>ABCDEFG</u>
	4	32	3.97	<u>ABCDF</u> , <u>ABCEG</u> , DEFG
	8	16	3.8	<u>ABCE</u> , <u>ABDF</u> , CDEF, <u>ACDG</u> , BDEG, BCFG, AEFG
	16	8	2.8	<u>ABC</u> , <u>ADE</u> , BCDE, <u>BDE</u> , ACDF, ABEF, CEF, <u>ABDG</u> , CDG, BEG, ACEG, AFG, BCFG, DEFG, ABCDEFG
	32	4	1.76	<u>AB</u> , <u>AC</u> , BC, <u>DE</u> , ABDE, ACDE, BCDE, <u>ADF</u> , BDF, CDF, AEF, BEF, CEF, <u>ADG</u> , BDG, CDG, AEG, BEG, CEG, FG, ABFG, ACFG, BCFG, DEFG, plus 7 higher-order interactions
	64	2	1	All 2-, 4- and 6-factor interactions
	128	1		
8	2	128	7	<u>ABCDEFGHI</u>
	4	64	4.96	<u>ABCDG</u> , <u>ABEFH</u> , CDEFGH
	8	32	3.96	<u>CDEF</u> , <u>ABDEG</u> , ABCFG, <u>ABCEH</u> , ABDFH, CDGH, EFGH
	16	16	3.8	<u>BCDE</u> , <u>ACDF</u> , ABEF, <u>ABDG</u> , ACEG, BCFG, DEFG, <u>ABCH</u> , ADEH, BDFH, CEFH, CDGH, BEGH, AFGH, ABCDEFGH
	32	8	1.96	<u>ABCD</u> , <u>ADE</u> , BCE, <u>ACE</u> , BDF, ABEF, CDEF, ACEG, ADFG, BCFG, BDEG, ABG, CDG, EFG, <u>CEGH</u> , DFGH, AH, BGH, CFH, DEH, BCDH, BEFH, plus 9 higher-order interactions
	64	4	1.75	<u>AB</u> , <u>AC</u> , BC, <u>DE</u> , <u>DF</u> , EF, <u>GH</u> , ADG, ADH, AEG, AEH, AFG, AFH, BDG, BDH, BEG, BEH, BFG, BFH, CDG, CDH, CEG, CEH, CFG, CFH, plus 38 higher-order interactions
	128	2	1	All 2-, 4-, 6- and 8-factor interactions

F

Column Labels of Generators for Regular Fractional Factorial Designs

Many tables of designs use numbers rather than strings of letters to identify interaction columns as generators. Table F is provided to assist with the use of these identifiers. Columns 1, 2, 4, 8, 16, 32, and 64 are basic columns and are labeled with the letters **A**, **B**, **C**, **D**, **E**, **F**, and **G**, respectively. Other columns are interactions of these. To identify the interaction corresponding to a given number, find the number in the body of Table F and then combine the letters in the column and row heading for that number. For example, interaction column 15 (in Table F's column **ABC** and row **D**) equals **ABCD**, whereas interaction column 96 (in Table F's blank column and row **FG**) equals **FG**.

Table F. Column number labels for interaction columns

	A	B	AB	C	AC	BC	ABC
	1	2	3	4	5	6	7
D	8	9	10	11	12	13	14
E	16	17	18	19	20	21	22
DE	24	25	26	27	28	29	30
F	32	33	34	35	36	37	38
DF	40	41	42	43	44	45	46
EF	48	49	50	51	52	53	54
DEF	56	57	58	59	60	61	62
G	64	65	66	67	68	69	70
DG	72	73	74	75	76	77	78
EG	80	81	82	83	84	85	86
DEG	88	89	90	91	92	93	94
FG	96	97	98	99	100	101	102
DFG	104	105	106	107	108	109	110
EFG	112	113	114	115	116	117	118
DEFG	120	121	122	123	124	125	126
							127

Column numbers above 127 require additional basic columns. For column numbers between 128 and 257, subtract 128 from the interaction column number, use Table F to find the interaction, and then append the letter **H**. For instance,

- For interaction column 200, we compute $200 - 128 = 72$. From Table F, we find that interaction column 72 corresponds to **DG**. Appending the letter **H**, we have that column 200 is **DGH**.
- For interaction column 250, we compute $250 - 128 = 122$, and find this number in the **B** column and **DEFG** row of Table F. So interaction column 250 is **BDEFGH**.

In general, powers of 2 higher than 2^8 denote additional basic columns. We use **J**, **K**, . . . , to denote these. Interactions are numbered according to the sum of the basic column numbers. Subtract successively smaller powers of 2 to decompose the interaction column number into basic columns. For instance, interaction column 300 is decomposed as follows:

$$300 - 256 = 44,$$

$$44 - 32 = 12,$$

$$12 - 8 = 4,$$

$$4 - 4 = 0.$$

Since columns 4, 8, 32, and 256 correspond to basic columns **C**, **D**, **F**, and **J**, column 300 is the **CDFJ** interaction.

G

Tables of Minimum Aberration Regular Fractional Factorial Designs

Tables G.1–G.9 present generators for all minimum aberration 2^{k-f} designs of sizes 8, 16, 32, 64, and 128. Each design is created by combining the $k-f$ basic columns with f generators. The generators are interactions of the basic columns, as described in Section 5.2.4. For convenience, Appendix F lists the interaction corresponding to each generator's number label. Examples are given following Table G.1 and G.2. The resolution and the first part of the word length pattern, A_3, A_4, \dots , are given for each design.

Table G.1. Minimum aberration designs of size 8

Res.	No. Factors	Design ID	Generator Columns	A_3	A_4	A_5
IV	4	4-1.1	7	0	1	
III	5	5-2.1	6, 7	2	1	0
	6	6-3.1	5, 6, 7	4	3	0
	7	7-4.1	3, 5, 6, 7	7	7	0

Example

The minimum aberration 2^{5-2} design may be obtained using $k-f=5-2=3$ basic columns and interaction columns numbered 6 and 7 as generators for the additional factors. From Table F, we know that columns 6 and 7 correspond to **BC** and **ABC**, respectively. So this design is obtained as a full factorial in **A**, **B**, **C**, plus **D** = **BC** and **E** = **ABC**. The defining relation is

$$\mathbf{I} = \mathbf{BCD} = \mathbf{ABCE} = \mathbf{ADE},$$

so the word length pattern is $(A_3, A_4, A_5) = (2, 1, 0)$.

Table G.2. Minimum aberration designs of size 16

No.	Res.	Factors	Design ID	Generator Columns	A_3	A_4	A_5
V		5	5-1.1	15	0	0	1
IV		6	6-2.1	7, 11	0	3	0
		7	7-3.1	7, 11, 13	0	7	0
		8	8-4.1	7, 11, 13, 14	0	14	0
III		9	9-5.1	7, 11-14	4	14	8
		10	10-6.1	7, 11-15	8	18	16
		11	11-7.1	7, 10-15	12	26	28
		12	12-8.1	7, 9-15	16	39	48
		13	13-9.1	6, 7, 9-15	22	55	72
		14	14-10.1	5-7, 9-15	28	77	112
		15	15-11.1	3, 5-7, 9-15	35	105	168

Example

The minimum aberration 2^{9-5} design is obtained using four basic columns, plus interaction columns numbered 7 and 11–14 as generators for the additional factors. From Table F, we know that columns 7, 11, 12, 13, and 14 correspond to **A****B****C**, **A****B****D**, **C****D**, **A****C****D**, and **B****C****D**, respectively. So this design is obtained as a full factorial in **A**, **B**, **C**, **D**, plus **E** = **A****B****C**, **F** = **A****B****D**, **G** = **C****D**, **H** = **A****C****D**, and **J** = **B****C****D**. The full defining relation will contain $2^5 = 32$ terms. From just the first three generators, the defining relation is

$$\mathbf{I} = \mathbf{ABCE} = \mathbf{ABDF} = \mathbf{CDEF} = \mathbf{CDG} = \mathbf{ABDEG} = \mathbf{ABCFG} = \mathbf{EFG}. \quad (\text{G.1})$$

Augment this portion by adding the product of (G.1) with **ACDH**:

$$\begin{aligned} &= \mathbf{ACDH} = \mathbf{BDEH} = \mathbf{BCFH} = \mathbf{AEFH} \\ &= \mathbf{AGH} = \mathbf{BCEGH} = \mathbf{BDFGH} = \mathbf{ACDEFGH}. \end{aligned} \quad (\text{G.2})$$

Finally, we complete the defining relation by multiplying (G.1) and (G.2) by **BCDJ**:

$$\begin{aligned} &= \mathbf{BCDJ} = \mathbf{ADEJ} = \mathbf{ACFJ} = \mathbf{BEFJ} = \mathbf{BGJ} = \mathbf{ACEGJ} \\ &= \mathbf{ADFGJ} = \mathbf{BCDEFGJ} = \mathbf{ABHJ} = \mathbf{CEHJ} = \mathbf{DFHJ} \\ &= \mathbf{ABCDEFHJ} = \mathbf{ABCDGHJ} = \mathbf{DEGHJ} = \mathbf{CFGHJ} = \mathbf{ABEFGHJ}. \end{aligned} \quad (\text{G.3})$$

The defining relation is the combination of (G.1), (G.2), and (G.3), and the full word length pattern is $(A_3, \dots, A_9) = (4, 14, 8, 0, 4, 1, 0)$.

Table G.3. Minimum aberration designs of size 32

No.	Design		Generator Columns	A_3	A_4	A_5
Res.	Factors	ID				
VI						
	6	6-1.1	31		0	0
IV						
	7	7-2.1	7, 27		0	1
	8	8-3.1	15, 19, 21		0	3
	9	9-4.1	15, 19, 21, 25		0	6
	10	10-5.1	15, 19, 21, 25, 30		0	10
	11	11-6.1	7, 11, 13, 21, 25, 31		0	25
	12	12-7.1	7, 11, 13, 14, 21, 25, 31		0	38
	13	13-8.1	7, 11, 13, 14, 19, 21, 25, 31		0	55
	14	14-9.1	7, 11, 13, 14, 19, 21, 22, 25, 31		0	77
	15	15-10.1	7, 11, 13, 14, 19, 21, 22, 25, 26, 31		0	105
	16	16-11.1	7, 11, 13, 14, 19, 21, 22, 25, 26, 28, 31		0	140
III						
	17	17-12.1	Design 16-11.1, plus 3		8	140
	18	18-13.1	Design 17-12.1, plus 5		16	148
	19	19-14.1	Design 18-13.1, plus 9		24	164
	20	20-15.1	Design 19-14.1, plus 17		32	188
	21	21-16.1	Design 20-15.1, plus 30		40	220
	22	22-17.1	Design 23-18.1, minus 27		48	263
	23	23-18.1	Design 24-19.1, minus 28		56	315
	24	24-19.1	Design 25-20.1, minus 6		64	378
	25	25-20.1	Design 26-21.1, minus 10		76	442
	26	26-21.1	Design 27-22.1, minus 18		88	518
	27	27-22.1	Design 28-23.1, minus 30		100	606
	28	28-23.1	Design 29-24.1, minus 12		112	707
	29	29-24.1	Design 30-25.1, minus 20		126	819
	30	30-25.1	Design 31-26.1, minus 24		140	945
	31	31-26.1	3, 5-7, 9-15, 17-31		155	1085
						5208

Table G.4a. Minimum aberration designs of size 64, with 7–32 factors

Res.	No. Factors	Design ID	Generator Columns	A_3	A_4	A_5
VII						
	7	7-1.1	63	0	0	0
V						
	8	8-2.1	31, 39	0	0	2
IV						
	9	9-3.1	31, 39, 41	0	1	4
	10	10-4.1	31, 39, 41, 51	0	2	8
	11	11-5.1	31, 39, 41, 42, 51	0	4	14
	12	12-6.1	31, 39, 41, 42, 51, 60	0	6	24
	13	13-7.1	21, 22, 31, 39, 41, 42, 51	0	14	28
	14	14-8.1	11, 13, 21, 31, 39, 41, 51, 52	0	22	40
	15	15-9.1	11, 13, 21, 31, 39, 41, 51, 52, 58	0	30	60
	16	16-10.1	Same as design 15-9.1, plus 22	0	43	81
	17	17-11.1	Same as design 16-10.1, plus 25	0	59	108
	18	18-12.1	Same as design 17-11.1, plus 28	0	78	144
	19	19-13.1	Same as design 18-12.1, plus 46	0	100	192
	20	20-14.1	Same as design 19-13.1, plus 61	0	125	256
	21	21-15.1	7, 11, 13, 14, 19, 21, 25, 31, 35, 37, 44, 52, 55, 61, 62	0	204	0
	22	22-16.1	Same as design 21-15.1, plus 49	0	250	0
	23	23-17.1	Same as design 22-16.1, plus 22	0	304	0
	24	24-18.1	Same as design 23-17.1, plus 41	0	365	0
	25	25-19.1	Same as design 24-18.1, plus 38	0	435	0
	26	26-20.1	Same as design 25-19.1, plus 26	0	515	0
	27	27-21.1	Same as design 26-20.1, plus 28	0	605	0
	28	28-22.1	Same as design 27-21.1, plus 42	0	706	0
	29	29-23.1	Same as design 28-22.1, plus 47	0	819	0
	30	30-24.1	Same as design 29-23.1, plus 50	0	945	0
	31	31-25.1	Same as design 30-24.1, plus 56	0	1085	0
	32	32-26.1	Same as design 31-25.1, plus 59	0	1240	0

Table G.4b. Minimum aberration designs of size 64, with 33–63 factors

Res.	No. Factors	Design ID	Generator Columns	A_3	A_4
III					
	33	33-27.1	Same as design 32-26.1, plus 63	16	1240
	34	34-28.1	Same as design 33-27.1, plus 60	32	1256
	35	35-29.1	Same as design 34-28.1, plus 43	48	1288
	36	36-30.1	Same as design 35-29.1, plus 12	64	1336
	37	37-31.1	33–63	80	1400
	38	38-32.1	31, 33–63	96	1480
	39	39-33.1	7, 27, 33–63	112	1577
	40	40-34.1	7, 11, 29, 33–63	128	1691
	41	41-35.1	7, 11, 19, 29, 33–63	144	1822
	42	42-36.1	7, 11, 19, 29, 30, 33–63	160	1970
	43	43-37.1	7, 11, 13, 19, 21, 25, 33–63	176	2145
	44	44-38.1	Same as design 43-37.1, plus 14	192	2334
	45	45-39.1	Same as design 44-38.1, plus 22	208	2543
	46	46-40.1	Same as design 45-39.1, plus 26	224	2773
	47	47-41.1	Same as design 46-40.1, plus 28	240	3025
	48	48-42.1	Same as design 47-41.1, plus 31	256	3300
	49	49-43.1	Same as design 48-42.1, plus 27	280	3556
	50	50-44.1	Same as design 49-43.1, plus 20	304	3836
	51	51-45.1	Same as design 50-44.1, plus 12	328	4140
	52	52-46.1	Same as design 51-45.1, plus 30	352	4468
	53	53-47.1	Same as design 52-46.1, plus 29	376	4820
	54	54-48.1	7, 11, 17–31, 33–63	400	5199
	55	55-49.1	Same as design 54-48.1, plus 13	424	5603
	56	56-50.1	Same as design 55-49.1, plus 14	448	6034
	57	57-51.1	Same as design 56-50.1, plus 12	476	6482
	58	58-52.1	Same as design 57-51.1, plus 15	504	6958
	59	59-53.1	Same as design 58-52.1, plus 10	532	7462
	60	60-54.1	Same as design 59-53.1, plus 9	560	7995
	61	61-55.1	Same as design 60-54.1, plus 6	590	8555
	62	62-56.1	Same as design 61-55.1, plus 5	620	9145
	63	63-57.1	Same as design 62-56.1, plus 3 I.e., 3, 5–7, 9–15, 17–31, 33–63	651	9765

Table G.5a. Minimum aberration designs of size 128, with 8–33 factors

Res.	No. Factors	Design ID	Generator Columns	A_4	A_5
VIII					
	8	8-1.1	127	0	0
VI					
	9	9-2.1	31, 103	0	0
V					
	10	10-3.1	31, 43, 103	0	3
	11	11-4.1	31, 43, 85, 103	0	6
IV					
	12	12-5.1	31, 43, 85, 103, 121	1	8
	13	13-6.1	31, 43, 44, 85, 86, 103	2	16
	14	14-7.1	31, 43, 46, 61, 85, 103, 114	3	24
	15	15-8.1	Same as design 14-7.1, plus 67	7	32
	16	16-9.1	31, 43, 44, 53, 85, 86, 88, 103, 110	10	48
	17	17-10.1	31, 43, 46, 61, 67, 78, 85, 103, 114, 116	15	60
	18	18-11.1	Same as design 17-10.1, plus 121	20	80
	19	19-12.1	31, 43, 46, 55, 58, 61, 67, 78, 85, 86, 103, 114	27	120
	20	20-13.1	Same as design 19-12.1, plus 91	36	152
	21	21-14.1	25, 31, 43, 44, 54, 56, 78, 82, 85, 88, 103, 104, 123, 125	51	200
	22	22-15.1	28, 31, 43, 44, 53, 58, 78, 83, 85, 86, 88, 97, 103, 104, 114	65	248
	23	23-16.1	25, 31, 43, 44, 49, 54, 56, 78, 82, 85, 88, 103, 104, 112, 123, 125	83	316
	24	24-17.1	19, 26, 28, 31, 43, 44, 53, 57, 67, 85, 86, 88, 98, 100, 103, 105, 110	102	384
	25	25-18.1	31, 38, 43, 44, 53, 58, 79, 83, 85, 86, 88, 97, 103, 104, 110, 114, 123, 124	124	482
	26	26-19.1	7, 11, 19, 29, 35, 46, 53, 57, 70, 73, 76, 82, 87, 94, 100, 109, 118, 120, 123	152	568
	27	27-20.1	Same as design 26-19.1, plus 97	180	690
	28	28-21.1	Same as design 27-20.1, plus 60	210	840
	29	29-22.1	Same as design 28-21.1, plus 69	266	945
	30	30-23.1	23, 25, 26, 39, 43, 45, 46, 51, 53, 56, 63, 71, 73, 74, 76, 81, 84, 88, 99, 101, 102, 104, 112	335	972
	31	31-24.1	7, 11, 19, 21, 22, 25, 26, 35, 45, 46, 49, 60, 67, 77, 78, 81, 95, 101, 105, 108, 116, 120, 123, 126	391	1134
	32	32-25.1	Same as design 30-23.1, plus 28, 82	452	1322
	33	33-26.1	Same as design 32-25.1, plus 54	518	1543

Table G.5b. Minimum aberration designs of size 128, with 34–58 factors

Res.	No. Factors	Design ID	Generator Columns	A_4	A_5
IV	34	34-27.1	Same as design 33-26.1, plus 95	589	1800
	35	35-28.1	Same as design 34-27.1, plus 111	665	2100
	36	36-29.1	Same as design 35-28.1, plus 15	756	2401
	37	37-30.1	Same as design 36-29.1, plus 119	854	2744
	38	38-31.1	Same as design 37-30.1, plus 123	959	3136
	39	39-32.1	Same as design 38-31.1, plus 125	1071	3584
	40	40-33.1	Same as design 39-32.1, plus 126	1190	4096
	41	41-34.1	7, 13, 19, 22, 25, 26, 31, 37, 38, 41, 42, 47, 49, 50, 52, 55, 56, 59, 73, 74, 76, 82, 84, 88, 93, 97, 100, 103, 104, 107, 109, 110, 112, 115	1648	0
	42	42-35.1	Same as design 41-34.1, plus 118	1822	0
	43	43-36.1	Same as design 42-35.1, plus 124	2009	0
	44	44-37.1	11, 13, 25, 26, 28, 35, 37, 38, 41, 42, 44, 50, 52, 55, 56, 59, 61, 62, 69, 70, 73, 74, 76, 79, 81, 87, 91, 97, 98, 100, 107, 110, 117, 118, 121, 122, 124	2214	0
	45	45-38.1	Same as design 44-37.1, plus 31	2430	0
	46	46-39.1	Same as design 45-38.1, plus 115	2665	0
	47	47-40.1	Same as design 46-39.1, plus 103	2915	0
	48	48-41.1	Same as design 47-40.1, plus 19	3180	0
	49	49-42.1	Same as design 48-41.1, plus 127	3466	0
	50	50-43.1	Same as design 49-42.1, plus 112	3770	0
	51	51-44.1	7, 11, 13, 22, 25, 26, 28, 31, 37, 38, 41, 42, 44, 47, 49, 50, 52, 55, 56, 61, 62, 69, 70, 73, 74, 76, 79, 81, 82, 84, 87, 88, 93, 94, 97, 98, 100, 103, 107, 109, 110, 115, 117, 118	4091	0
	52	52-45.1	Same as design 50-43.1, plus 82, 93	4433	0
	53	53-46.1	Same as design 52-45.1, plus 109	4797	0
	54	54-47.1	Same as design 53-46.1, plus 104	5182	0
	55	55-48.1	Same as design 54-47.1, plus 88	5589	0
	56	56-49.1	19, 21, 22, 25, 26, 28, 35, 37, 38, 41, 42, 44, 49, 50, 52, 55, 56, 59, 61, 62, 67, 69, 70, 73, 74, 76, 81, 82, 84, 87, 88, 91, 93, 94, 97, 98, 100, 103, 104, 107, 109, 110, 115, 117, 118, 121, 122, 124, 127	6020	0
	57	57-50.1	Same as design 55-48.1, plus 7, 21	6475	0
	58	58-51.1	Same as design 57-50.1, plus 14	6955	0

Table G.5c. Minimum aberration designs of size 128, with 59–91 factors

No.	Design						
Res.	Factors	ID	Generator	Columns		A_3	A_4
IV							
	59	59-52.1	Same as design 58.51.1, plus 22			0	7461
	60	60-53.1	Same as design 59-52.1, plus 47			0	7994
	61	61-54.1	Same as design 60-53.1, plus 49			0	8555
	62	62-55.1	Same as design 61-54.1, plus 67			0	9145
	63	63-56.1	Same as design 62-55.1, plus 84			0	9765
	64	64-57.1	Same as design 63-56.1, plus 94			0	10416
III							
	65	65-58.1	Same as design 64-57.1, plus 3			32	10416
	66	66-59.1	Same as design 65-58.1, plus 5			64	10448
	67	67-60.1	Same as design 66-59.1, plus 9			96	10512
	68	68-61.1	Same as design 67-60.1, plus 17			128	10608
	69	69-62.1	Same as design 68-61.1, plus 33			160	10736
	70	70-63.1	Same as design 69-62.1, plus 65			192	10896
	71	71-64.1	63, 65–127			224	11088
	72	72-65.1	45, 51, 65–127			256	11312
	73	73-66.1	Same as design 72-65.1, plus 7			288	11569
	74	74-67.1	Same as design 73-66.1, plus 29			320	11858
	75	75-68.1	Same as design 74-67.1, plus 11			352	12180
	76	76-69.1	Same as design 75-68.1, plus 62			384	12534
	77	77-70.1	7, 11, 21, 25, 38, 58, 60, 65–127			416	12926
	78	78-71.1	7, 11, 19, 30, 37, 41, 49, 60, 65–127			448	13350
	79	79-72.1	Same as design 78-71.1, plus 63			480	13806
	80	80-73.1	7, 11, 13, 19, 21, 35, 37, 57, 58, 60, 65–127			512	14299
	81	81-74.1	Same as design 80-73.1, plus 14			544	14827
	82	82-75.1	Same as design 81-74.1, plus 22			576	15390
	83	83-76.1	Same as design 82-75.1, plus 38			608	15988
	84	84-77.1	Same as design 83-76.1, plus 63			640	16621
	85	85-78.1	7, 11, 13, 14, 19, 21, 22, 25, 35, 41, 42, 49, 52, 56, 62, 64–127			672	17340
	86	86-79.1	Same as design 85-78.1, plus 37			704	18058
	87	87-80.1	Same as design 86-79.1, plus 26			736	18816
	88	88-81.1	Same as design 87-80.1, plus 38			768	19613
	89	89-82.1	Same as design 88-81.1, plus 28			800	20451
	90	90-83.1	7, 11, 13, 14, 19, 21, 22, 25, 26, 28, 35, 37, 38, 41, 42, 44, 49, 50, 52, 56, 65–127			832	21331
	91	91-84.1	Same as design 90-83.1, plus 31			864	22253

Table G.5d. Minimum aberration designs of size 128, with 92–127 factors

Res.	No. Factors	Design ID	Generator Columns	A_3	A_4
III					
	92	92-85.1	Same as design 91-84.1, plus 47	896	23218
	93	93-86.1	Same as design 92-85.1, plus 55	928	24227
	94	94-87.1	Same as design 93-86.1, plus 59	960	25281
	95	95-88.1	Same as design 94-87.1, plus 61	992	26381
	96	96-89.1	Same as design 95-88.1, plus 62	1024	27528
	97	97-90.1	Same as design 96-89.1, plus 63	1072	28552
	98	98-91.1	Same as design 97-90.1, plus 60	1120	29624
	99	99-92.1	Same as design 98-91.1, plus 43	1168	30744
	100	100-93.1	Same as design 99-92.1, plus 12	1216	31912
	101	101-94.1	33-63, 65-127	1264	33128
	102	102-95.1	31, 33-63, 65-127	1312	34392
	103	103-96.1	7, 27, 33-63, 65-127	1360	35705
	104	104-97.1	7, 11, 29, 33-63, 65-127	1408	37067
	105	105-98.1	7, 11, 19, 29, 33-63, 65-127	1456	38478
	106	106-99.1	7, 11, 19, 29, 30, 33-63, 65-127	1504	39938
	107	107-100.1	7, 11, 13, 19, 21, 25, 33-63, 65-127	1552	41457
	108	108-101.1	Same as design 107-100.1, plus 14	1600	43022
	109	109-102.1	Same as design 108-101.1, plus 22	1648	44639
	110	110-103.1	Same as design 109-102.1, plus 26	1696	46309
	111	111-104.1	Same as design 110-103.1, plus 28	1744	48033
	112	112-105.1	Same as design 111-104.1, plus 31	1792	49812
	113	113-106.1	Same as design 112-105.1, plus 27	1848	51604
	114	114-107.1	Same as design 113-106.1, plus 20	1904	53452
	115	115-108.1	Same as design 114-107.1, plus 12	1960	55356
	116	116-109.1	Same as design 115-108.1, plus 30	2016	57316
	117	117-110.1	Same as design 116-109.1, plus 29	2072	59332
	118	118-111.1	7, 11, 17-31, 33-63, 65-127	2128	61407
	119	119-112.1	Same as design 118-111.1, plus 13	2184	63539
	120	120-113.1	Same as design 119-112.1, plus 14	2240	65730
	121	121-114.1	Same as design 120-113.1, plus 12	2300	67970
	122	122-115.1	Same as design 121-114.1, plus 15	2360	70270
	123	123-116.1	Same as design 122-115.1, plus 10	2420	72630
	124	124-117.1	Same as design 123-116.1, plus 9	2480	75051
	125	125-118.1	Same as design 124-117.1, plus 6	2542	77531
	126	126-119.1	Same as design 125-118.1, plus 5	2604	80073
	127	127-120.1	Same as design 126-119.1, plus 3	2667	82677

H

Minimum Aberration Blocking Schemes for Fractional Factorial Designs

Below are the optimal 8-, 16-, 32-, 64-, and 128-run designs based on the W_1 criteria proposed by Cheng and Wu (2002). In most cases, the fractional factorial is the same design listed in Tables G. For the few exceptions, the generators for the required fractional factorial designs are listed.

Table H.1. Blocking for 16-run fractional factorial designs

No. Factors	Design ID	No. Blocks	Block Generators	$A_{2,1}$	$A_{3,1}$
4	4-1.1	2	3	2	0
	4-1.1	4	3, 5	6	0
5	5-2.1	2	3	2	2
6	6-3.1	2	3	3	4

Table H.2. Blocking for 16-run fractional factorial designs

No. Factors	Design ID	No. Blocks	Block Generators	$A_{2.1}$	$A_{3.1}$
5	5-1.1	2	3	1	1
	5-1.1	4	3, 5	3	3
	5-1.2*	8	9, 10, 12	10	0
6	6-2.1	2	13	0	4
	6-2.1	4	13, 14	3	8
	6-2.1	8	5, 6, 9	15	0
7	7-3.1	2	14	0	7
	7-3.1	4	3, 5	9	0
	7-3.1	8	3, 5, 9	21	0
8	8-4.1	2	3	4	0
	8-4.1	4	3, 5	12	0
	8-4.1	8	3, 5, 9	28	0
9	9-5.1	2	3	4	4
	9-5.1	4	3, 5	12	12
10	10-6.1	2	5	4	8
	10-6.1	4	3, 5	13	24
11	11-7.1	2	9	4	13
	11-7.1	4	3, 5	15	36
12	12-8.1	2	3	6	16
	12-8.1	4	3, 5	18	48
13	13-9.1	2	3	6	22
14	14-10.1	2	3	7	28

*Generator for 16-run fractional factorial not in Appendix G.

- Design 5-1.2: column 7 (i.e., $\mathbf{E} = \mathbf{ABC}$)

Table H.3a. Blocking for 32-run fractional factorial designs, 6–13 factors

No. Factors	Design ID	No. Blocks	Block Generators	$A_{2.1}$	$A_{3.1}$
6	6-1.1	2	7	0	2
	6-1.1	4	3, 13	1	4
	6-1.1	8	3, 12, 21	3	8
	6-1.1	16	3, 5, 9, 17	15	0
7	7-2.1	2	21	0	2
	7-2.1	4	13, 19	1	6
	7-2.1	8	5, 11, 19	5	12
	7-2.2*	16	3, 5, 9, 17	21	0
8	8-3.1	2	25	0	3
	8-3.1	4	7, 25	1	10
	8-3.1	8	7, 9, 17	8	16
	8-3.2*	16	3, 5, 9, 17	28	0
9	9-4.1	2	30	0	4
	9-4.1	4	3, 29	4	8
	9-4.1	8	3, 5, 24	12	16
	9-4.3*	16	3, 5, 9, 17	36	0
10	10-5.1	2	3	2	4
	10-5.1	4	3, 5	6	12
	10-5.1	8	3, 5, 17	17	24
	10-5.2*	16	3, 5, 9, 17	45	0
11	11-6.1	2	14	0	13
	11-6.1	4	5, 19	4	26
	11-6.1	8	3, 5, 17	25	0
	11-6.1	16	3, 5, 9, 17	55	0
12	12-7.1	2	19	0	17
	12-7.1	4	5, 19	5	34
	12-7.1	8	3, 5, 17	30	0
	12-7.1	16	3, 5, 9, 17	66	0
13	13-8.1	2	22	0	22
	13-8.1	4	6, 26	6	44
	13-8.1	8	3, 5, 17	36	0
	13-8.1	16	3, 5, 9, 17	78	0

Table H.3b. Blocking for 32-run fractional factorial designs, 14–22 factors

No. Factors	Design ID	No. Blocks	Block		
			Generators	$A_{2.1}$	$A_{3.1}$
14	14-9.1	2	26	0	28
	14-9.1	4	6, 26	7	56
	14-9.1	8	3, 9, 17	42	0
	14-9.1	16	3, 5, 9, 17	91	0
15	15-10.1	2	28	0	35
	15-10.1	4	3, 5	21	0
	15-10.1	8	3, 5, 9	49	0
	15-10.1	16	3, 5, 9, 17	105	0
16	16-11.1	2	3	8	0
	16-11.1	4	3, 5	24	0
	16-11.1	8	3, 5, 9	56	0
	16-11.1	16	3, 5, 9, 17	120	0
17	17-12.1	2	5	8	8
	17-12.1	4	5, 9	24	24
	17-12.1	8	5, 9, 17	56	56
18	18-13.1	2	9	8	16
	18-13.1	4	9, 17	24	48
	18-13.1	8	6, 9, 17	57	112
19	19-14.1	2	17	8	24
	19-14.1	4	15, 17	24	73
	19-14.1	8	6, 10, 17	59	168
20	20-15.1	2	30	8	32
	20-15.1	4	6, 27	25	98
	20-15.1	8	6, 10, 18	62	224
21	21-16.1	2	6	9	41
	21-16.1	4	6, 10	27	123
	21-16.2*	8	6, 10, 18	66	280
22	22-17.1	2	27	8	52
	22-17.1	4	6, 27	27	152
	22-17.1	8	6, 10, 18	71	336

Table H.3c. Blocking for 32-run fractional factorial designs, 23–30 factors

No. Factors	Design ID	No. Blocks	Block		
			Generators	$A_{2.1}$	$A_{3.1}$
23	23-18.1	2	29	8	63
	23-18.1	4	6, 10	33	168
	23-18.1	8	6, 10, 18	77	392
24	24-19.1	2	6	12	64
	24-19.1	4	6, 10	36	192
	24-19.1	8	6, 10, 18	84	448
25	25-20.1	2	10	12	76
	25-20.1	4	10, 18	36	228
26	26-21.1	2	18	12	88
	26-21.1	4	12, 18	37	264
27	27-22.1	2	30	12	101
	27-22.1	4	12, 20	39	300
28	28-23.1	2	12	14	112
	28-23.1	4	12, 20	42	336
29	29-24.1	2	20	14	126
30	30-25.1	2	24	15	140

***Generators for 32-run fractional factorials not in Appendix G.** For cases in Tables H.3a and H3.b requiring a fraction other than the minimum aberration fraction ($k-f.1$), use the following columns as generators:

- Design 7-2.2: 7, 31 (i.e., $\mathbf{F} = \mathbf{ABC}$, $\mathbf{G} = \mathbf{ABCDE}$)
- Design 8-3.2: add 11 to Design 7-2.2 (i.e., $\mathbf{H} = \mathbf{ABD}$)
- Design 9-4.3: add 21 to Design 8-3.2 (i.e., $\mathbf{J} = \mathbf{ACE}$)
- Design 10-5.2: add 25 to Design 9-4.3 (i.e., $\mathbf{K} = \mathbf{ADE}$)
- Design 21-16.2: 3, 5, 7, 9, 11, 13-15, 17, 19, 21, 22, 25, 26, 28, 31

Table H.4a. Blocking for 64-run fractional factorial designs, 7–13 factors

No. Factors	Design ID	No. Blocks	Block Generators	$A_{2.1}$	$A_{3.1}$
7	7-1.1	2	7	0	1
	7-1.1	4	7, 25	0	3
	7-1.1	8	7, 25, 42	0	7
	7-1.2*	16	3, 12, 21, 33	5	12
	7-1.2*	32	3, 5, 9, 17, 33	21	0
8	8-2.1	2	41	0	1
	8-2.1	4	11, 49	0	4
	8-2.1	8	7, 9, 50	2	8
	8-2.1	16	3, 12, 21, 33	7	18
	8-2.2*	32	3, 5, 9, 17, 33	28	0
9	9-3.1	2	51	0	1
	9-3.1	4	19, 46	0	6
	9-3.1	8	6, 11, 49	2	14
	9-3.1	16	3, 12, 21, 33	9	27
	9-3.4*	32	3, 5, 9, 17, 33	36	0
10	10-4.1	2	42	0	2
	10-4.1	4	11, 53	0	8
	10-4.1	8	5, 11, 48	4	16
	10-4.1	16	3, 12, 21, 33	12	36
	10-4.7*	32	3, 5, 9, 17, 33	45	0
11	11-5.1	2	60	0	2
	11-5.1	4	11, 21	1	12
	11-5.1	8	5, 11, 18	5	24
	11-5.1	16	5, 11, 18, 35	15	48
	11-5.14*	32	3, 5, 9, 17, 33	55	0
12	12-6.1	2	11	0	8
	12-6.1	4	5, 11	2	16
	12-6.1	8	5, 11, 18	6	32
	12-6.1	16	5, 11, 18, 35	18	64
	12-6.17*	32	3, 5, 9, 17, 33	66	0
13	13-7.1	2	52	0	8
	13-7.1	4	12, 52	2	18
	13-7.1	8	5, 10, 19	7	42
	13-7.1	16	3, 5, 9, 49	23	75
	13-7.20*	32	3, 5, 9, 17, 33	78	0

Table H.4b. Blocking for 64-run fractional factorial designs, 14–20 factors

No. Factors	Design ID	No. Blocks	Block Generators	$A_{2.1}$	$A_{3.1}$
14	14-8.1	2	58	0	8
	14-8.1	4	19, 46	2	25
	14-8.1	8	5, 25, 35	9	52
	14-8.1	16	5, 10, 19, 33	26	100
	14-8.18*	32	3, 5, 9, 17, 33	91	0
15	15-9.1	2	22	0	13
	15-9.1	4	5, 25	3	32
	15-9.2*	8	5, 10, 19	11	66
	15-9.1	16	5, 10, 19, 33	30	125
	15-9.14*	32	3, 5, 9, 17, 33	105	0
16	16-10.1	2	25	0	16
	16-10.1	4	5, 25	3	41
	16-10.1	8	5, 25, 33	12	81
	16-10.1	16	5, 9, 17, 33	45	108
	16-10.10*	32	3, 5, 9, 17, 33	120	0
17	17-11.1	2	28	0	19
	17-11.1	4	19, 46	3	50
	17-11.1	8	15, 19, 33	13	99
	17-11.1	16	7, 9, 19, 33	48	144
	17-11.9*	32	3, 5, 9, 17, 33	136	0
18	18-12.1	2	46	0	22
	18-12.1	4	19, 46	3	60
	18-12.1	8	3, 5, 40	21	85
	18-12.1	16	7, 9, 19, 33	57	160
	18-12.4*	32	3, 5, 9, 17, 33	153	0
19	19-13.1	2	61	0	25
	19-13.1	4	3, 61	8	49
	19-13.1	8	3, 5, 56	24	97
	19-13.1	16	3, 9, 20, 33	67	176
	19-13.2*	32	3, 5, 9, 17, 33	171	0
20	20-14.1	2	3	4	16
	20-14.1	4	3, 5	12	48
	20-14.1	8	3, 5, 9	34	96
	20-14.1	16	3, 5, 9, 17	78	192
	20-14.2*	32	3, 5, 9, 17, 33	190	0

Table H.4c. Blocking for 64-run fractional factorial designs, 21–27 factors

No. Factors	Design ID	No. Blocks	Block Generators	$A_{2.1}$	$A_{3.1}$
21	21-15.1	2	49	0	46
	21-15.1	4	26, 33	6	94
	21-15.1	8	6, 26, 41	21	189
	21-15.1	16	3, 5, 17, 40	100	0
	21-15.1	32	3, 5, 9, 17, 33	210	0
22	22-16.1	2	22	0	54
	22-16.1	4	5, 42	7	108
	22-16.1	8	3, 17, 41	24	216
	22-16.1	16	3, 5, 24, 33	110	0
	22-16.1	32	3, 5, 9, 17, 33	231	0
23	23-17.1	2	41	0	61
	23-17.1	4	5, 42	8	123
	23-17.1	8	3, 17, 41	27	246
	23-17.1	16	3, 5, 9, 33	121	0
	23-17.1	32	3, 5, 9, 17, 33	253	0
24	24-18.1	2	38	0	70
	24-18.1	4	3, 56	9	141
	24-18.1	8	9, 20, 38	30	280
	24-18.1	16	3, 5, 17, 40	132	0
	24-18.1	32	3, 5, 9, 17, 33	276	0
25	25-19.1	2	26	0	80
	25-19.1	4	9, 50	10	160
	25-19.1	8	3, 5, 57	66	0
	25-19.1	16	3, 5, 9, 48	144	0
	25-19.1	32	3, 5, 9, 17, 33	300	0
26	26-20.1	2	28	0	90
	26-20.1	4	3, 56	11	180
	26-20.1	8	3, 5, 57	72	0
	26-20.1	16	3, 5, 9, 48	156	0
	26-20.1	32	3, 5, 9, 17, 33	325	0
27	27-21.1	2	42	0	101
	27-21.1	4	3, 56	12	202
	27-21.1	8	3, 12, 33	78	0
	27-21.1	16	3, 5, 9, 33	169	0
	27-21.1	32	3, 5, 9, 17, 33	351	0

Table H.4d. Blocking for 64-run fractional factorial designs, 28–32 factors

No. Factors	Design ID	No. Blocks	Block Generators	$A_{2.1}$	$A_{3.1}$
28	28-22.1	2	47	0	113
	28-22.1	4	3, 56	13	226
	28-22.1	8	5, 27, 33	84	0
	28-22.1	16	3, 5, 24, 33	182	0
	28-22.1	32	3, 5, 9, 17, 33	378	0
29	29-23.1	2	50	0	126
	29-23.1	4	3, 56	14	252
	29-23.1	8	5, 17, 33	91	0
	29-23.1	16	3, 5, 17, 33	196	0
	29-23.1	32	3, 5, 9, 17, 33	406	0
30	30-24.1	2	56	0	140
	30-24.1	4	3, 56	15	280
	30-24.1	8	5, 9, 17	98	0
	30-24.1	16	5, 9, 17, 33	210	0
	30-24.1	32	3, 5, 9, 17, 33	435	0
31	31-25.1	2	59	0	155
	31-25.1	4	3, 5	45	0
	31-25.1	8	3, 5, 9	105	0
	31-25.1	16	3, 5, 9, 17	225	0
	31-25.1	32	3, 5, 9, 17, 33	465	0
32	32-26.1	2	3	16	0
	32-26.1	4	3, 5	48	0
	32-26.1	8	3, 5, 9	112	0
	32-26.1	16	3, 5, 9, 17	240	0
	32-26.1	32	3, 5, 9, 17, 33	496	0

***Generators for 64-run fractional factorials not in Appendix G.** For cases in Tables H.4a and H.4b requiring a fraction other than the minimum aberration fraction ($k-f.1$), use the following columns as generators:

- Design 7-1.2: 31 (i.e., $\mathbf{G} = \mathbf{ABCDE}$)
- Design 8-2.2: Add 35 to Design 7-1.2 (i.e., $\mathbf{H} = \mathbf{ABF}$)
- Design 9-3.4: Add 13 to Design 8-2.2 (i.e., $\mathbf{J} = \mathbf{ACD}$)
- Design 10-4.7: Add 52 to Design 9-3.4 (i.e., $\mathbf{K} = \mathbf{CEF}$)
- Design 11-5.14: Add 14 to Design 10-4.7 (i.e., $\mathbf{L} = \mathbf{BCD}$)
- Design 12-6.17: Add 55 to Design 11-5.14 (i.e., $\mathbf{M} = \mathbf{ABCEF}$)
- Design 13-7.20: Add 21 to Design 12-6.17 (i.e., $\mathbf{N} = \mathbf{ACE}$)

- Design 14-8.18: Add 37, 61 to Design 12-6.17
- Design 15-9.2: 11, 13, 21, 31, 39, 41, 46, 52
- Design 15-9.14: Add 11 to Design 14-8.18
- Design 16-10.10: Add 19 to Design 15-9.14
- Design 17-11.9: Add 21 to Design 16-10.10
- Design 18-12.4: Add 44 to Design 17-11.9
- Design 19-13.2: Add 7 to Design 18-12.4
- Design 20-14.2: Add 62 to Design 19-13.2

Table H.5a. Blocking for 128-run fractional factorial designs, 8–13 factors

No. Factors	Design ID	No. Blocks	Block Generators	$A_{2.1}$	$A_{3.1}$
8	8-1.1	2	15	0	0
	8-1.1	4	15, 51	0	0
	8-1.1	8	15, 51, 85	0	0
	8-1.2*	16	7, 25, 42, 65	1	10
	8-1.2*	32	3, 5, 24, 40, 73	7	18
	8-1.1	64	3, 5, 9, 17, 33, 65	28	0
9	9-2.1	2	43	0	0
	9-2.1	4	43, 85	0	0
	9-2.1	8	7, 41, 82	0	6
	9-2.1	16	3, 13, 37, 84	2	14
	9-2.1	32	7, 9, 18, 33, 66	9	27
	9-2.1	64	3, 5, 9, 17, 33, 65	36	0
10	10-3.1	2	85	0	0
	10-3.1	4	44, 81	0	2
	10-3.1	8	7, 49, 74	0	9
	10-3.1	16	3, 12, 53, 69	3	19
	10-3.1	32	3, 9, 17, 36, 69	12	36
	10-3.3*	64	3, 5, 9, 17, 33, 65	45	0
11	11-4.1	2	121	0	1
	11-4.1	4	14, 115	0	4
	11-4.1	8	7, 49, 91	0	12
	11-4.1	16	6, 11, 49, 67	4	25
	11-4.1	32	3, 9, 20, 36, 69	15	48
	11-4.1e*	64	3, 5, 9, 17, 33, 65	55	0
12	12-5.1	2	13	0	2
	12-5.1	4	13, 49	0	6
	12-5.1	8	7, 49, 91	0	16
	12-5.2*	16	3, 13, 52, 69	5	34
	12-5.1	32	3, 9, 20, 36, 69	18	64
	12-5.1e*	64	3, 5, 9, 17, 33, 65	66	0
13	13-6.1	2	88	0	2
	13-6.1	4	25, 105	0	8
	13-6.2*	8	13, 55, 67	0	22
	13-6.2*	16	3, 13, 52, 69	6	44
	13-6.1	32	5, 11, 18, 35, 67	22	80
	13-6.1e*	64	3, 5, 9, 17, 33, 65	78	0

Table H.5b. Blocking for 128-run fractional factorial designs, 14–19 factors

No. Factors	Design ID	No. Blocks	Block Generators	$A_{2,1}$	$A_{3,1}$
14	14-7.1	2	67	0	4
	14-7.1	4	13, 67	0	12
	14-7.1	8	13, 55, 67	0	28
	14-7.1	16	13, 21, 34, 67	7	56
	14-7.1	32	3, 5, 17, 33, 72	26	100
	14-7.1e*	64	3, 5, 9, 17, 33, 65	91	0
15	15-8.1	2	78	0	4
	15-8.1	4	13, 116	0	12
	15-8.1	8	13, 55, 86	0	35
	15-8.1	16	13, 21, 34, 70	9	68
	15-8.1	32	3, 5, 17, 33, 73	30	125
	15-8.1e*	64	3, 5, 9, 17, 33, 65	105	0
16	16-9.1	2	19	0	6
	16-9.1	4	19, 97	0	18
	16-9.1	8	9, 19, 97	2	40
	16-9.1	16	3, 13, 20, 101	12	80
	16-9.1	32	3, 5, 9, 49, 65	36	144
	16-9.1e*	64	3, 5, 9, 17, 33, 65	120	0
17	17-10.1	2	121	0	5
	17-10.2	4	58, 86	0	20
	17-10.1	8	7, 25, 97	3	49
	17-10.3*	16	3, 13, 20, 101	13	99
	17-10.1	32	3, 5, 17, 33, 73	40	180
	17-10.1e*	64	3, 5, 9, 17, 33, 65	136	0
18	18-11.1	2	7	0	11
	18-11.2*	4	24, 100	1	24
	18-11.2*	8	13, 21, 100	5	56
	18-11.2*	16	6, 9, 19, 66	16	116
	18-11.2*	32	3, 5, 17, 33, 73	46	210
	18-11.1e*	64	3, 5, 9, 17, 33, 65	153	0
19	19-12.1	2	91	0	9
	19-12.1	4	13, 100	2	30
	19-12.1	8	3, 37, 89	7	66
	19-12.1	16	5, 10, 19, 99	19	134
	19-12.1	32	3, 5, 17, 33, 73	51	252
	19-12.1e*	64	3, 5, 9, 17, 33, 65	171	0

Table H.5c. Blocking for 128-run fractional factorial designs, 20–25 factors

No. Factors	Design ID	No. Blocks	Block Generators	$A_{2.1}$	$A_{3.1}$
20	20-13.1	2	77	0	16
	20-13.1	4	3, 105	3	36
	20-13.1	8	3, 12, 101	9	76
	20-13.1	16	6, 9, 19, 35	22	160
	20-13.1	32	3, 12, 21, 37, 68	62	256
	20-13.1e*	64	3, 5, 9, 17, 33, 65	190	0
21	21-14.1	2	112	0	14
	21-14.1	4	17, 97	2	42
	21-14.2*	8	9, 19, 97	6	96
	21-14.2*	16	6, 9, 18, 98	22	186
	21-14.1	32	3, 12, 20, 37, 68	64	336
	21-14.1e*	64	3, 5, 9, 17, 33, 65	210	0
22	22-15.1	2	124	0	20
	22-15.1	4	13, 113	2	48
	22-15.1	8	13, 23, 102	8	100
	22-15.1	16	7, 10, 51, 66	24	216
	22-15.1	32	6, 9, 17, 33, 67	71	384
	22-15.1e*	64	3, 5, 9, 17, 33, 65	231	0
23	23-16.1	2	73	0	24
	23-16.2*	4	34, 89	3	49
	23-16.2*	8	9, 48, 66	11	110
	23-16.2*	16	6, 9, 18, 98	29	238
	23-16.1	32	5, 10, 18, 34, 67	77	448
	23-16.1e*	64	3, 5, 9, 17, 33, 65	253	0
24	24-17.1	2	62	0	24
	24-17.1	4	15, 112	2	64
	24-17.1	8	7, 58, 74	10	136
	24-17.1	16	5, 11, 35, 81	30	280
	24-17.1	32	3, 5, 9, 49, 65	84	512
	24-17.1e*	64	3, 5, 9, 17, 33, 65	276	0
25	25-18.1	2	3	2	19
	25-18.1	4	3, 13	6	64
	25-18.1	8	3, 9, 17	16	140
	25-18.1	16	3, 5, 9, 17	38	294
	25-18.2*	32	7, 9, 18, 33, 66	92	576
	25-18.1e*	64	3, 5, 9, 17, 33, 65	300	0

***Generators for 128-run fractional factorials not in Appendix G.**

For cases in Tables H.5a to H.5c requiring a fraction other than the minimum aberration fraction ($k-f.1$), use the following columns as generators:

- Design 8-1.2: 63
- Design 10-3.3: 31, 41, 103
- Design 11-4.8: 31, 41, 82, 103
- Design 12-5.2: 31, 43, 44, 85, 103
- Design 12-5.1e: 31, 41, 82, 103, 124
- Design 13-6.2: 31, 43, 46, 61, 85, 103
- Design 13-6.1e: 31, 41, 44, 82, 93, 103
- Design 14-7.1e: Add 7 to Design 13-6.1e
- Design 15-8.1e: 7, 31, 41, 52, 82, 94, 103, 124
- Design 16-9.1e: Add 109 to Design 15-8.1e
- Design 17-10.2: 31, 43, 46, 55, 61, 67, 78, 85, 103, 114
- Design 17-10.3: 31, 38, 43, 44, 53, 58, 85, 86, 88, 103
- Design 17-10.1e: Add 11 to Design 16-9.1e
- Design 18-11.2: 31, 43, 46, 55, 58, 61, 67, 78, 85, 103, 114
- Design 18-11.1e: Add 52 to Design 17-10.1e
- Design 19-12.1e: Add 122 to Design 18-11.1e
- Design 20-13.1e: 7, 19, 25, 31, 41, 44, 61, 82, 87, 93, 97, 103, 107
- Design 21-14.2: 31, 38, 43, 44, 53, 58, 79, 83, 85, 86, 88, 103, 110, 124
- Design 21-14.1e: Add 117 to Design 20-13.1e
- Design 22-15.1e: Add 11 to Design 21-14.1e
- Design 23-16.2: 31, 38, 43, 44, 53, 58, 79, 83, 85, 86, 88, 97, 103, 104, 110, 124
- Design 23-16.1e: Add 37 to Design 22-15.1e
- Design 24-17.1e: Add 70 to Design 23-16.1e
- Design 25-18.2: 31, 43, 44, 53, 54, 56, 79, 83, 85, 86, 88, 97, 98, 103, 104, 112, 123, 124
- Design 25-18.1e: 7, 11, 19, 22, 31, 41, 44, 49, 50, 74, 82, 87, 93, 97, 103, 107, 110, 117

I

Alias Matrix Derivation

Let \mathbf{X}_f denote the model matrix for the fitted model corresponding to a design \mathbf{D} , and let \mathbf{X}_o denote the matrix of columns for omitted terms that are in fact active. The true model is

$$E(y) = \mathbf{X}_f \beta_f + \mathbf{X}_o \beta_o, \quad (\text{I.1})$$

but one fits a simpler model using least squares to obtain

$$\mathbf{b}_f = (\mathbf{X}'_f \mathbf{X}_f)^{-1} \mathbf{X}_f \mathbf{Y}. \quad (\text{I.2})$$

Taking the expectation of (I.2) and substituting using (I.1), one obtains

$$E(\mathbf{b}_f) = (\mathbf{X}'_f \mathbf{X}_f)^{-1} \mathbf{X}_f [\mathbf{X}_f \beta_f + \mathbf{X}_o \beta_o] \\ = \beta_f + \mathbf{A} \beta_o,$$

where

$$\mathbf{A} = (\mathbf{X}'_f \mathbf{X}_f)^{-1} \mathbf{X}'_f \mathbf{X}_o.$$

The matrix \mathbf{A} is known as the alias matrix.

If $\mathbf{X}'_f \mathbf{X}_f$ is a diagonal matrix and \mathbf{X}_o consists of omitted interactions, then the elements of \mathbf{A} represent correlations between the columns of \mathbf{X}_f and the columns of \mathbf{X}_o . For regular fractional factorial designs, these correlations are ± 1 for aliased effects and 0 otherwise. For instance, for the 2^{5-2} fraction in Table 5.3, suppose we fit a first-order model and we consider aliasing from omitted two-factor interactions. Then \mathbf{X}_f has six columns, \mathbf{X}_o has 10 columns, and the alias matrix is

The first row of zeros indicates that the two-factor interactions do not bias the intercept. Two nonzero entries in the second row indicate that two interactions are aliased with the first main effect. That four columns have only zeros indicates that they are not aliased with main effects.

For nonregular designs, some correlations take values other than -1 , 0 , and 1 . For instance, for the 12-run Plackett–Burman design (e.g., Table 6.19), any omitted interaction will bias 9 of the 11 main effect estimates in a saturated first-order model. The first 12 columns of the alias matrix for this case are

	AB	AC	AD	AE	AF	AG	AH	AJ	AK	AL	BC	BD	...
I	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	0	0	0	0	$-1/3$	$1/3$	
B	$0 -1/3$	$1/3$	$1/3$	$-1/3$	$-1/3$	$1/3$	$-1/3$	$-1/3$	$-1/3$	0	0		
C	$-1/3$	$0 -1/3$	$-1/3$	$1/3$	$-1/3$	$1/3$	$-1/3$	$-1/3$	$1/3$	$-0 -1/3$			
D	$1/3 -1/3$	$0 -1/3$	$-1/3$	$-1/3$	$-1/3$	$-1/3$	$-1/3$	$1/3$	$1/3 -1/3$	-0			
E	$1/3 -1/3$	$-1/3$	0	$1/3$	$1/3$	$-1/3$	$-1/3$	$-1/3$	$-1/3$	$1/3 -1/3$			
F	$-1/3$	$1/3 -1/3$	$1/3$	0	$-1/3$	$-1/3$	$-1/3$	$1/3$	$-1/3$	$1/3 -1/3$			
G	$-1/3$	$-1/3 -1/3$	$1/3$	$-1/3$	0	$-1/3$	$1/3$	$-1/3$	$1/3 -1/3$	$1/3$			
H	$1/3$	$1/3 -1/3$	$-1/3$	$-1/3$	$-1/3$	0	$1/3$	$-1/3$	$-1/3$	$-1/3 -1/3$			
J	$-1/3$	$-1/3 -1/3$	$-1/3$	$-1/3$	$1/3$	$1/3$	0	$1/3$	$-1/3$	$1/3 -1/3$			
K	$-1/3$	$-1/3$	$1/3 -1/3$	$1/3$	$-1/3$	$-1/3$	$1/3$	0	$-1/3$	$-1/3 -1/3$			
L	$-1/3$	$1/3$	$1/3 -1/3$	$-1/3$	$1/3$	$-1/3$	$-1/3$	$-1/3$	0	$-1/3 -1/3$			

Consider now the 12-run, seven-factor experiment of Bermego-Barrera et al. (2001), where we adopted a two-factor interaction model for **A**, **B**, and **G**, and omitted **C**–**F**; refer to Tables 6.22 and 6.24. If we have ignored an active main effect, what would be the consequence to our estimates in Table 6.24? Here \mathbf{X}_f has seven columns and \mathbf{X}_o has four columns. The corresponding alias matrix is

	C	D	E	F
I	0	0	0	0
A	$-1/8$	$1/8$	$-1/8$	$1/8$
B	$-1/8$	$-1/8$	$1/8$	$-1/8$
G	$-1/8$	$1/8$	$1/8$	$-1/8$
AB	$-3/8$	$3/8$	$3/8$	$-3/8$
AG	$-3/8$	$-3/8$	$3/8$	$-3/8$
BG	$-3/8$	$3/8$	$-3/8$	$3/8$

Including interactions in the model affects how the main effect estimates b_A , b_B , and b_G are calculated. As a result, these estimators are biased by the omission of any active main effects, although the bias is minimal with the coefficient in the alias matrix of $\pm 1/8$. The potential bias to the interaction estimators is greater.

J

Distinguishing Among Fractional Factorial Designs

Fractional factorial designs can be usefully distinguished via their word length patterns. Whereas for larger designs the word length pattern does not fully distinguish designs, it does so for all regular (resolution III and higher) fractions of size 4, 8, 16, and 32. Even more useful is the fact that for regular designs, there is a one-to-one correspondence between the word length pattern and any row or column of the *row coincidence matrix* $\mathbf{T} = \mathbf{D}'\mathbf{D}$, where \mathbf{D} is the $n \times k$ design matrix with ± 1 coding. If two treatment combinations are opposite, then their inner product is $-k$, whereas if they coincide perfectly, their product is $+k$; otherwise, they assume a value in the range $-k + 2, \dots, k - 2$.

When one row of \mathbf{D} is the treatment combination with all factors “+1,” then the row coincidence distribution is obtained by summing the k columns together. For instance, Hsieh and Goodwin (1986) performed a nine-factor, 16-run experiment with treatment combinations shown in Table J.1. (When this example was considered in Section 2.8.2, we ignored factors x_5, \dots, x_9 and treated it as a simple 2^4 .) The last column gives the row sums of the nine main effect columns. How one recognizes this as Chen, Sun, and Wu’s (1993) (CSW) design 9-5.2 will be explained below.

Table J.1 Hsieh and Goodwin's (1986) 2^{9-5} factorial

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	Sum
-1	-1	1	-1	-1	-1	1	-1	1	-3
-1	-1	-1	-1	-1	1	1	1	-1	-3
-1	1	1	1	-1	-1	-1	-1	-1	-3
-1	1	-1	1	-1	1	-1	1	1	1
-1	1	-1	-1	1	-1	1	1	1	1
-1	1	1	-1	1	1	1	-1	-1	1
-1	-1	-1	1	1	-1	-1	1	-1	-3
-1	-1	1	1	1	1	-1	-1	1	1
1	1	1	-1	-1	-1	-1	1	1	1
1	1	-1	-1	-1	1	-1	-1	-1	-3
1	-1	1	1	-1	-1	1	1	-1	1
1	-1	-1	1	-1	1	1	-1	1	1
1	-1	-1	-1	1	-1	-1	-1	1	-3
1	-1	1	-1	1	1	-1	1	-1	1
1	1	-1	1	1	-1	1	-1	-1	1
1	1	1	1	1	1	1	1	1	9

Define the N elements of any row or column of \mathbf{T} by t_1, \dots, t_N and let the r^{th} moment be defined as

$$M_r = [t_1^r + t_2^r + \dots + t_N^r]/N.$$

Butler (2003b) showed that, for regular fractions of resolution III or higher, $M_1 = 0$, $M_2 = k$, and subsequent moments could be used to compute A_3, \dots, A_6 as follows:

$$\begin{aligned} A_3 &= M_3/6 \\ A_4 &= [M_4 - (3k - 2)k]/24, \\ A_5 &= [M_5 - 10(k - 2)M_3]/120, \\ A_6 &= [M_6 - 5(3k - 8)M_4 + 2(15k^2 - 60k + 32)k]/720. \end{aligned}$$

For the design in Table J.1, the row coincidence moments are

$$\begin{aligned} M_2 &= [6(-3)^2 + 9(1)^2 + (9)^2]/16 = 9, \\ M_3 &= [6(-3)^3 + 9(1)^3 + (9)^3]/16 = 36, \\ M_4 &= [6(-3)^4 + 9(1)^4 + (9)^4]/16 = 441, \\ M_5 &= [6(-3)^5 + 9(1)^5 + (9)^5]/16 = 3600, \\ M_6 &= [6(-3)^6 + 9(1)^6 + (9)^6]/16 = 33489. \end{aligned}$$

Therefore, for Hsieh and Goodwin's 2^{9-5} design,

$$\begin{aligned} A_3 &= 36/6 = 6, \\ A_4 &= (441 - 225)/24 = 9, \\ A_5 &= (3600 - 2520)/120 = 9, \\ A_6 &= (33489 - 41895 + 12726)/720 = 6, \end{aligned}$$

which matches the word length pattern for CSW's design 9-5.2.

CSW found five nonisomorphic 2^{9-5} designs; their word length patterns are listed in Section 6.2 of this book. In Table J.2 we list the row coincidence distribution for these five designs, obtained by constructing each design based on the generators in CSW and then summing the nine columns together.

Table J.2 Row coincidence distribution for five 2^{9-5}_{III} designs

Design 9-5.1	Design 9-5.2	Design 9-5.3	Design 9-5.4	Design 9-5.5
-7	-3	-5	-3	-3
-1	-3	-3	-3	-3
-1	-3	-3	-3	-3
-1	-3	-1	-3	-1
-1	-3	-1	-1	-1
-1	-3	-1	-1	-1
-1	1	-1	-1	-1
-1	1	-1	-1	-1
-1	1	-1	-1	-1
1	1	1	-1	-1
1	1	1	1	-1
1	1	1	1	1
1	1	1	1	1
1	1	3	3	5
9	9	9	9	9
M_3	24	36	36	42
M_4	561	441	465	441
				465

Note that the minimum aberration ranking is equivalent to sorting on M_3 , then M_4 , etc. The best design is the one with the least skewness in the row coincidence distribution. Note also how the best design maximizes the minimum distance between points. For the best two designs, the largest coincidence between two treatment combinations is agreement on five factors and opposite on four, resulting in a row coincidence value of 1. The poorer designs have pairs of runs that agree on six or seven factors, producing a t_i value of 3 or 5, respectively.

Two 2^{k-f} designs have the same row coincidence distribution if and only if they have the same word length pattern. For regular designs with 32 or fewer runs, we can distinguish designs by simply computing one column of the \mathbf{T} matrix. However, for designs of size 64 and larger, there are nonisomorphic designs with identical word length patterns and, hence, identical t_1, \dots, t_n . Block (2003) found that by taking pairs of columns of \mathbf{T} , he could distinguish all 64-run designs and most 128-run designs. For the practitioner, usually one

begins with a set of generators, as in Appendix G, and then constructs the design. However, when encountering an unknown fractional factorial design, such as in Hsieh and Goodwin (1986), the row coincidence distribution is a simple tool for identifying the particular fractional factorial.

The generalized word length pattern for nonregular designs may be computed from the row coincidence matrix \mathbf{T} (Butler 2003b). The formula for B_3, \dots, B_6 are identical to those given for A_3, \dots, A_6 , except that the M_r moments must be computed using the entire \mathbf{T} matrix; that is, for nonregular designs, the columns of \mathbf{T} may have distributions that differ. For example, taking any eight columns from the 12-run orthogonal array in Table 6.16, the row coincidence matrix is

$$\mathbf{T} = \begin{bmatrix} 8 & 2 & 0 & 0 & -2 & -2 & -2 & -2 & -2 & 0 & -2 & 2 \\ 2 & 8 & -2 & -2 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & -4 \\ 0 & -2 & 8 & 0 & -2 & -2 & 2 & 2 & -2 & 0 & -2 & -2 \\ 0 & -2 & 0 & 8 & -2 & 2 & -2 & -2 & -2 & 0 & 2 & -2 \\ -2 & 0 & -2 & -2 & 8 & 0 & 0 & 0 & -4 & 2 & 0 & 0 \\ -2 & 0 & -2 & 2 & 0 & 8 & 0 & 0 & 0 & -2 & -4 & 0 \\ -2 & 0 & 2 & -2 & 0 & 0 & 8 & -4 & 0 & -2 & 0 & 0 \\ -2 & 0 & 2 & -2 & 0 & 0 & -4 & 8 & 0 & -2 & 0 & 0 \\ -2 & 0 & -2 & -2 & -4 & 0 & 0 & 0 & 8 & 2 & 0 & 0 \\ 0 & -2 & 0 & 0 & 2 & -2 & -2 & -2 & 2 & 8 & -2 & -2 \\ -2 & 0 & -2 & 2 & 0 & -4 & 0 & 0 & 0 & -2 & 8 & 0 \\ 2 & -4 & -2 & -2 & 0 & 0 & 0 & 0 & 0 & -2 & 0 & 8 \end{bmatrix}.$$

Note that some columns contain a -4 while others do not. Thus, the moments must be computed using the entire $N \times N$ matrix. Here, $M_3 = 37\bar{3}$, so $B_3 = 37\bar{3}/6 = 6\bar{2}$.

References

- Abraham, B., Chipman, H., and Vijayan, K. (1999). Some Risks in the Construction and Analysis of Supersaturated Designs. *Technometrics* **41** (2), 135–141.
- Abraham, B. and MacKay, J. (1993). Variation Reduction and Designed Experiments. *International Statistical Review* **61** (1), 121–129.
- Addelman, S. (1961). Irregular Fractions of 2^n Factorial Experiments. *Technometrics* **3** (4), 479–496.
- Addelman, S. (1962). Orthogonal Main Effect Plans for Asymmetrical Factorial Experiments. *Technometrics* **4** (1), 21–46.
- Addelman, S. (1964). Some Two-Level Factorial Plans with Split Plot Confounding. *Technometrics* **6** (3), 253–258.
- Addelman, S. (1969). Sequences of Two-Level Fractional Factorial Plans. *Technometrics* **11** (3), 477–509.
- Ahuja, S. K., Ferreira, G. M., and Moreira, A. R. (2004). Application of Plackett–Burman Design and Response Surface Methodology to Achieve Exponential Growth for Aggregated Shipworm Bacterium. *Biotechnology and Bioengineering* **85** (6), 666–675.
- Ai, M., Xu, X., and Wu, C. F. J. (2008). Optimal Blocking and Foldover Plans for Regular Two-Level Designs. Technical Report No. 26, Peking University, Institute of Mathematics, School of Mathematical Sciences.
- Ai, M. Y., Yang, G. J., and Zhang, R. C. (2006). Minimum Aberration Blocking of Regular Mixed Factorial Designs. *Journal of Statistical Planning and Inference* **136** (4), 1493–1511.
- Anbari, F. T. and Lucas, J. M. (2008). Designing and Running Super-Efficient Experiments: Optimum Blocking with One Hard-to-Change Factor. *Journal of Quality Technology* **40** (1), 31–45.
- Anderson, C. M. and Wu, C. F. J. (1995). Measuring Location Effects from Factorial Experiments with a Directional Response. *International Statistical Review* **63** (3), 345–363.

- Anderson, C. M. and Wu, C. F. J. (1996). Dispersion Measures and Analysis for Factorial Directional Data with Replicates. *Applied Statistics* **45** (1), 47–61.
- Anderson-Cook, C. M. (2001). Understanding the Influence of Several Factors on a Cylindrical Response. *Journal of Quality Technology* **33** (2), 167–180.
- Arber, S., Mckinlay, J., Adams, A., Marceau, L., Link, C., and O'Donnell, A. (2006). Patient Characteristics and Inequalities in Doctors' Diagnostic and Management Strategies Relating to Coronary Heart Disease: A Video-Simulation Experiment. *Social Science & Medicine* **62** (1), 103–115.
- Atkinson, A. C. and Donev, A. N. (1989). The Construction of Exact D-Optimum Experimental Designs with Application to Blocking Response Surface Designs. *Biometrika* **76** (3), 515–526.
- Atkinson, A. C. and Donev, A. N. (1992). *Optimum Experimental Designs*. Oxford University Press, New York.
- Atkinson, A. C., Donev A., and Tobias, R. D. (2007). *Optimum Experimental Designs, with SAS*. Oxford University Press, New York.
- Baardseth, P., Bjerke, F., Aaby, K., and Mielnik, M. (2005). A Screening Experiment to Identify Factors Causing Rancidity During Meat Loaf Production. *European Food Research and Technology* **221** (5), 653–661.
- Bafna, S. S. and Beall, A. M. (1997). A Design of Experiments Study on the Factors Affecting Variability in the Melt Index Measurement. *Journal of Applied Polymer Science* **65** (2), 277–288.
- Bandurek, G. (1999). Practical Application of Supersaturated Arrays. *Quality and Reliability Engineering International* **15** (2), 123–133.
- Barnett, J., Czitrom, V., John, P. W. M., and León, R. V. (1997). Using Fewer Wafers to Resolve Confounding in Screening Experiments. In: *Statistical Case Studies for Industrial Process Improvement*, Czitrom, V. and Spagon, P. D. (Eds.), SIAM, Philadelphia, pp. 235–250.
- Bartlett, M. S. and Kendall, D. G. (1946). The Statistical Analysis of Variance-Heterogeneity and the Logarithmic Transformation. *Journal of the Royal Statistical Society, Ser. B* **8** (1), 128–138.
- Barton, R. R. (1997). Pre-Experiment Planning for Designed Experiments: Graphical Methods. *Journal of Quality Technology* **29** (3), 307–316.
- Beattie, S. D., Fong, D. K. H., and Lin, D. K. J. (2002). A Two-Stage Bayesian Model Selection Strategy for Supersaturated Designs. *Technometrics* **44** (1), 55–63.
- Beckman, R. J. and Cook, R. D. (1983). Outliers... *Technometrics* **25** (2), 119–149.
- Belcher-Novosad, S. and Ingram, D. (2003). Identifying Minimum G Aberration Designs from Hadamard Matrices of Order 28. *Journal of the Arkansas Academy of Science* **57**, 202–205.
- Bell, G. H., Ledolter, J., and Swersey, A. J. (2006). Experimental Design on the Front Lines of Marketing: Testing New Ideas to Increase Direct Mail Sales. *International Journal of Research in Marketing* **23** (3), 309–319.

- Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate — A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* **57** (1), 289–300.
- Benjamini, Y. and Hochberg, Y. (2000). On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics. *Journal of Educational and Behavioral Statistics* **25** (1), 60–83.
- Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive Linear Step-Up Procedures that Control the False Discovery Rate. *Biometrika* **93** (3), 491–507.
- Bergerud, W. A. (1996). Displaying Factor Relationships in Experiments. *The American Statistician* **50** (3), 228–233.
- Bermejo-Barrera, P., Muniz-Naveiro, O., Moreda-Pineiro, A., and Bermejo-Barrera, A. (2001). The Multivariate Optimisation of Ultrasonic Bath-Induced Acid Leaching for the Determination of Trace Elements in Seafood Products by Atomic Absorption Spectrometry. *Analytica Chimica Acta* **439** (2), 211–227.
- Bingham, D. R., Schoen, E. D., and Sitter, R. R. (2004). Designing Fractional Factorial Split-Plot Experiments with Few Whole-Plot Factors. *Applied Statistics* **53** (2), 325–339. Corrigendum 2005, pp. 955–958.
- Bingham, D. R., Sitter, R., Kelly, E., Moore, L., and Olivas, J. D. (2008). Factorial Designs with Multiple Levels of Randomization. *Statistica Sinica* **18** (2), 493–513.
- Bingham, D. R. and Sitter, R. R. (1999). Minimum Aberration Two-Level Fractional Factorial Split-Plot Designs. *Technometrics* **41** (1), 62–70.
- Bingham, D. R. and Sitter, R. R. (2001). Design Issues in Fractional Factorial Split-Plot Experiments. *Journal of Quality Technology* **33** (1), 2–15.
- Bisgaard, S. (1994). Blocking Generators for Small 2^{k-p} Designs. *Journal of Quality Technology* **26** (4), 288–296.
- Bisgaard, S. (1997). Designing Experiments for Tolerancing Assembled Products. *Technometrics* **39** (2), 142–152.
- Bisgaard, S. (2000). The Design and Analysis of $2^{k-p} \times 2^{q-r}$ Split Plot Experiments. *Journal of Quality Technology* **32** (1), 39–56.
- Bisgaard, S. and Fuller, H. T. (1995). Sample-Size Estimates for 2^{k-p} Designs with Binary Responses. *Journal of Quality Technology* **27** (4), 344–354. Corrigendum 1995, p. 496.
- Bisgaard, S., Fuller, H. T., and Barrios, E. (1996). Two-Level Factorials Run as Split Plot Experiments. *Quality Engineering* **8** (4), 705–708.
- Block, R. M. (2003). Theory and Construction Methods for Large Regular Resolution IV Designs. Ph.D. thesis, University of Tennessee, Knoxville, available from <http://etd.utk.edu/2003/BlockRobert.pdf>.
- Block, R. M. and Mee, R. W. (2003). Second Order Saturated Resolution IV Designs. *Journal of Statistical Theory and Applications* **2** (2), 96–112.
- Block, R. M. and Mee, R. W. (2005). Resolution IV Designs with 128 Runs. *Journal of Quality Technology* **37** (4), 282–293. Corrigendum 2006, p. 196.

- Blom, G. (1958). *Statistical Estimates and Transformed Beta Variates*. Wiley, New York.
- Booth, K. H. V. and Cox, D. R. (1962). Some Systematic Supersaturated Designs. *Technometrics* **4** (4), 489–495.
- Bouler, J. M., Trecant, M., Delecrin, J., Royer, J., Passuti, N., and Daculsi, G. (1996). Macroporous Biphasic Calcium Phosphate Ceramics: Influence of Five Synthesis Parameters on Compressive Strength. *Journal of Biomedical Materials Research* **32** (4), 603–609.
- Box, G. E. P. and Behnken, D. W. (1960). Some New Three Level Designs for the Study of Quantitative Variables. *Technometrics* **2** (4), 455–475.
- Box, G. E. P. and Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society, Ser. B* **26** (2), 211–252.
- Box, G. E. P. and Draper, N. R. (2007). *Response Surfaces, Mixtures, and Ridge Analyses*, 2nd Edition. Wiley, Hoboken, NJ.
- Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. Wiley, New York.
- Box, G. E. P., Hunter, W. G., and Hunter, J. S. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*, 2nd Edition. Wiley, Hoboken, NJ.
- Box, G. E. P. and Jones, S. (1992). Split-Plot Designs for Robust Product Experimentation. *Journal of Applied Statistics* **19** (1), 3–26.
- Box, G. E. P. and Liu, P. Y. T. (1999). Statistics as a Catalyst to Learning by Scientific Method Part I — An Example. *Journal of Quality Technology* **31** (1), 1–15.
- Box, G. E. P. and Meyer, R. D. (1986). An Analysis for Unreplicated Fractional Factorials. *Technometrics* **28** (1), 11–18.
- Box, G. E. P. and Meyer, R. D. (1993). Finding the Active Factors in Fractionated Screening Experiments. *Journal of Quality Technology* **25** (2), 94–105.
- Box, G. E. P. and Wilson, K. B. (1951). On the Experimental Attainment of Optimum Conditions. *Journal of the Royal Statistical Society, Ser. B* **13** (1), 1–45.
- Brenneman, W. A. and Nair, V. N. (2001). Methods for Identifying Dispersion Effects in Unreplicated Factorial Experiments: A Critical Analysis and Proposed Strategies. *Technometrics* **43** (4), 388–405.
- Brien, C. J. and Bailey, R. A. (2006). Multiple Randomizations. *Journal of the Royal Statistical Society, Ser. B* **68** (4), 571–599.
- Broudiscou, L. P., Papon, Y., and Broudiscou, A. F. (2000). Effects of Dry Plant Extracts on Fermentation and Methanogenesis in Continuous Culture of Rumen Microbes. *Animal Feed Science and Technology* **87** (3-4), 263–277.
- Bullington, R. G., Lovin, S., Miller, D. M., and Woodall, W. H. (1993). Improvement of an Industrial Thermostat Using Designed Experiments. *Journal of Quality Technology* **25** (4), 262–270.

- Bulutoglu, D. A. and Cheng, C. S. (2003). Hidden Projection Properties of Some Nonregular Fractional Factorial Designs and Their Applications. *Annals of Statistics* **31** (3), 1012–1026.
- Butler, N. A. (2003a). Some Theory for Constructing Minimum Aberration Fractional Factorial Designs. *Biometrika* **90** (1), 233–238.
- Butler, N. A. (2003b). Minimum Aberration Construction Results for Nonregular Two-Level Fractional Factorial Designs. *Biometrika* **90** (4), 891–898.
- Butler, N. A. (2004). Construction of Two-Level Split-Lot Fractional Factorial Designs for Multistage Processes. *Technometrics* **46** (4), 445–451.
- Butler, N. A. (2005). Classification of Efficient Two-Level Fractional Factorial Designs of Resolution IV or More. *Journal of Statistical Planning and Inference* **131** (1), 145–159.
- Butler, N. A. (2006). Optimal Blocking of Two-Level Factorial Designs. *Biometrika* **93** (2), 289–302.
- Butler, N. A., Mead, R., Eskridge, K. M., and Gilmour, S. G. (2001). A General Method of Constructing $E(s^2)$ -Optimal Supersaturated Designs. *Journal of the Royal Statistical Society, Ser. B* **63** (3), 621–632.
- Carpinteiro, J., Quintana, J. B., Martinez, E., Rodriguez, I., Carro, A. M., Lorenzo, R. A., and Cela, R. (2004). Application of Strategic Sample Composition to the Screening of Anti-Inflammatory Drugs in Water Samples Using Solid-Phase Microextraction. *Analytica Chimica Acta* **524** (1–2), 63–71.
- Carroll, R. J. and Cline, D. B. H. (1988). An Asymptotic Theory for Weighted Least Squares with Weights Estimated by Replication. *Biometrika* **75** (1), 35–43.
- Chen, H. G. and Cheng, C. S. (1999). Theory of Optimal Blocking of 2^{n-m} Designs. *Annals of Statistics* **27** (6), 1948–1973.
- Chen, H. G. and Hedayat, A. S. (1996). 2^{n-l} Designs with Weak Minimum Aberration. *Annals of Statistics* **24** (6), 2536–2548.
- Chen, H. G. and Hedayat, A. S. (1998). 2^{n-m} Designs with Resolution III or IV Containing Clear Two-Factor Interactions. *Journal of Statistical Planning and Inference* **75** (1), 147–158.
- Chen, H. H. and Cheng, C. S. (2006). Doubling and Projection: A Method of Constructing Two-Level Designs of Resolution IV. *Annals of Statistics* **34** (1), 546–558.
- Chen, J. H. and Lin, D. K. J. (1998). On the Identifiability of a Supersaturated Design. *Journal of Statistical Planning and Inference* **72** (1–2), 99–107.
- Chen, J. H., Sun, D. X., and Wu, C. F. J. (1993). A Catalog of Two-Level and Three-Level Fractional Factorial Designs with Small Runs. *International Statistical Review* **61** (1), 131–145.
- Cheng, C. S. (1995). Some Projection Properties of Orthogonal Arrays. *Annals of Statistics* **23** (4), 1223–1233.
- Cheng, C. S. (1997). $E(s^2)$ -Optimal Supersaturated Designs. *Statistica Sinica* **7** (4), 929–939.

- Cheng, C. S. (1998). Some Hidden Projection Properties of Orthogonal Arrays with Strength Three. *Biometrika* **85** (2), 491–495.
- Cheng, C. S. and Jacroux, M. (1988). The Construction of Trend-Free Run Orders of Two-Level Factorial Designs. *Journal of the American Statistical Association* **83** (404), 1152–1158.
- Cheng, C. S. and Li, C. C. (1993). Constructing Orthogonal Fractional Factorial Designs When Some Factor-Level Combinations Are Debarred. *Technometrics* **35** (3), 277–283.
- Cheng, C. S., Mee, R. W., and Yee, O. (2008). Second Order Saturated Orthogonal Arrays of Strength Three. *Statistica Sinica* **18** (1), 105–119.
- Cheng, C. S. and Steinberg, D. M. (1991). Trend Robust Two-Level Factorial Designs. *Biometrika* **78** (2), 325–336.
- Cheng, C. S., Steinberg, D. M., and Sun, D. X. (1999). Minimum Aberration and Model Robustness for Two-Level Fractional Factorial Designs. *Journal of the Royal Statistical Society, Ser. B* **61** (1), 85–93.
- Cheng, C. S. and Tang, B. X. (2001). Upper Bounds on the Number of Columns in Supersaturated Designs. *Biometrika* **88** (4), 1169–1174.
- Cheng, S. W. and Wu, C. F. J. (2002). Choice of Optimal Blocking Schemes in Two-Level and Three-Level Designs. *Technometrics* **44** (3), 269–277.
- Cheng, S. W., Wu, C. F. J., and Wu, H. Q. (2003). Finding Defining Generators with Extreme Lengths. *Journal of Statistical Planning and Inference* **113** (1), 315–321.
- Chipman, H. A. and Hamada, M. S. (1996). Follow-Up Designs to Resolve Confounding in Multifactor Experiments. Discussion: Factor-Based or Effect-Based Modeling? Implications for Design. *Technometrics* **38** (4), 317–320.
- Choueiki, M. H., Mount-Campbell, C. A., and Ahalt, S. C. (1997). Building a ‘Quasi Optimal’ Neural Network to Solve the Short-Term Load Forecasting Problem. *IEEE Transactions on Power Systems* **12** (4), 1432–1439.
- Clark, J. B. and Dean, A. M. (2001). Equivalence of Fractional Factorial Designs. *Statistica Sinica* **11** (2), 537–547.
- Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*, 2nd Edition. Wiley, New York.
- Coleman, D. E. and Montgomery, D. C. (1993). A Systematic Approach to Planning for a Designed Industrial Experiment. *Technometrics* **35** (1), 1–12.
- Collett, D. (2002). *Modelling Binary Data*, 2nd Edition. Chapman & Hall/CRC, Boca Raton, FL.
- Cook, R. D. and Nachtsheim, C. J. (1989). Computer-Aided Blocking of Factorial and Response-Surface Designs. *Technometrics* **31** (3), 339–346.
- Cox, D. R. (1958). *Planning of Experiments*. John Wiley, New York.
- Cox, D. R. and Reid, N. (2000). *The Theory of the Design of Experiments*. Chapman & Hall/CRC, Boca Raton, FL.
- Crosier, R. B. (2000). Some New Two-Level Saturated Designs. *Journal of Quality Technology* **32** (2), 103–110.

- Czitrom, V., Mohammadi, P., Flemming, M., and Dyas, B. (1998). Robust Design Experiment to Reduce Variance Components. *Quality Engineering* **10** (4), 645–655.
- Daniel, C. (1959). Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments. *Technometrics* **1** (4), 311–341.
- Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*. Wiley, New York.
- Daniel, C. and Wilcoxon, F. (1966). Factorial 2^{p-q} Plans Robust Against Linear and Quadratic Trends. *Technometrics* **8** (2), 259–278.
- Davidian, M. and Haaland, P. D. (1990). Regression and Calibration with Nonconstant Error Variance. *Chemometrics and Intelligent Laboratory Systems* **9** (3), 231–248.
- Davies, O. L. (Ed.) (1954). *The Design and Analysis of Industrial Experiments*. Oliver and Boyd, London.
- Dean, A. M. and Voss, D. T. (1999). *Design and Analysis of Experiments*. Springer-Verlag, New York.
- Deng, L. Y., Lin, D. K. J., and Wang, J. N. (1996). Marginally Oversaturated Designs. *Communications in Statistics: Theory and Methods* **25** (11), 2557–2573.
- Deng, L. Y., Lin, D. K. J., and Wang, J. N. (1999). A Resolution Rank Criterion for Supersaturated Designs. *Statistica Sinica* **9** (2), 605–610.
- Deng, L. Y. and Tang, B. X. (1999). Generalized Resolution and Minimum Aberration Criteria for Plackett–Burman and Other Nonregular Factorial Designs. *Statistica Sinica* **9** (4), 1071–1082.
- Deng, L. Y. and Tang, B. X. (2002). Design Selection and Classification for Hadamard Matrices Using Generalized Minimum Aberration Criteria. *Technometrics* **44** (2), 173–184.
- Derringer, G. and Suich, R. (1980). Simultaneous Optimization of Several Response Variables. *Journal of Quality Technology* **12** (4), 214–219.
- Dickinson, A. W. (1974). Some Run Orders Requiring a Minimum Number of Factor Level Changes for 2^4 and 2^5 Main Effect Plans. *Technometrics* **16** (1), 31–37.
- Dong, F. (1993). On the Identification of Active Contrasts in Unreplicated Fractional Factorials. *Statistica Sinica* **3** (1), 209–217.
- Draper, N. R. and Lin, D. K. J. (1990). Capacity Considerations for Two-Level Fractional Factorial Designs. *Journal of Statistical Planning and Inference* **24** (1), 25–35.
- Draper, N. R. and Stoneman, D. M. (1964). Estimating Missing Values in Unreplicated Two-Level Factorial and Fractional Factorial Designs. *Biometrics* **20** (3), 443–458.
- Draper, N. R. and Stoneman, D. M. (1968). Factor Changes and Linear Trends in Eight-Run Two Level Factorial Designs. *Technometrics* **10** (2), 301–311.
- Durig, T. and Fassihi, A. R. (1993). Identification of Stabilizing and Destabilizing Effects of Excipient-Drug Interactions in Solid Dosage Form Design. *International Journal of Pharmaceutics* **97** (1–3), 161–170.

- Dykstra, O. (1959). Partial Duplication of Factorial Experiments. *Technometrics* **1** (1), 63–75.
- Edmondson, R. N. (1991). Agricultural Response Surface Experiments Based on Four-Level Factorial Designs. *Biometrics* **47** (4), 1435–1448.
- Edwards, D. J. (2008). Analysis and Design of Screening Experiments, Assuming Effect Sparsity. Ph.D. thesis, University of Tennessee, Knoxville.
- Edwards, D. J. and Mee, R. W. (2008). Empirically Determined p-Values for Lenth t-Statistics. *Journal of Quality Technology* **40** (4), 368–380.
- Edwards, D. J. and Mee, R. W. (2009). Supersaturated Designs: Are Our Results Significant? Technical Report 2009-01, University of Tennessee, Department of Statistics, Operations, and Management Science.
- Ellekjaer, M. R., Ilseng, M. A., and Naes, T. (1996). A Case Study of the Use of Experimental Design and Multivariate Analysis in Product Improvement. *Food Quality and Preference* **7** (1), 29–36.
- Engel, J. (1992). Modeling Variation in Industrial Experiments. *Applied Statistics* **41** (3), 579–593.
- Engel, J. (2008). Factorial Effects, Random Blocks, and Longitudinal Data: Two Simple Analysis Methods. *Journal of Quality Technology* **40** (1), 97–108.
- Evangelaras, H., Georgiou, S., and Koukouvinos, C. (2003). Inequivalent Projections of Hadamard Matrices of Orders 16 and 20. *Metrika* **57** (1), 29–35.
- Evangelaras, H., Georgiou, S., and Koukouvinos, C. (2004). Evaluation of Inequivalent Projections of Hadamard Matrices of Order 24. *Metrika* **59** (1), 51–73.
- Evangelaras, H., Kolaiti, E., and Koukouvinos, C. (2006). Some Orthogonal Arrays with 32 Runs and Their Projection Properties. *Metrika* **63** (3), 271–281.
- Evangelaras, H., Koukouvinos, C., and Stylianou, S. (2005). Evaluation of Some Non-Orthogonal Saturated Designs with Two Levels. *Statistics & Probability Letters* **74** (4), 322–329.
- Filliben, J. J. and Li, K. C. (1997). A Systematic Approach to the Analysis of Complex Interaction Patterns in Two-Level Factorial Designs. *Technometrics* **39** (3), 286–297.
- Fisher, R. A. (1926). The Arrangement of Field Experiments. *Journal of the Ministry of Agriculture of Great Britain* **33**, 503–513.
- Fisher, R. A. (1950). *The Design of Experiments*, 6th Edition. Oliver and Boyd, London.
- Franklin, M. F. (1984). Constructing Tables of Minimum Aberration p^{n-m} Designs. *Technometrics* **26** (3), 225–232.
- Freeman, M. F. and Tukey, J. W. (1950). Transformations Related to the Angular and the Square Root. *Annals of Mathematical Statistics* **21** (4), 607–611.
- Fujisawa, H. (2000). Variance Stabilizing Transformation and Studentization for Estimator of Correlation Coefficient. *Statistics & Probability Letters* **47** (3), 213–217.

- Ganju, J. and Lucas, J. M. (1997). Bias in Test Statistics when Restrictions in Randomization Are Caused by Factors. *Communications in Statistics: Theory and Methods* **26** (1), 47–63.
- Gilmour, S. G. (2004). Irregular Four-Level Response Surface Designs. *Journal of Applied Statistics* **31** (9), 1043–1048.
- Gilmour, S. G. and Mead, R. (1996). Fixing a Factor in the Sequential Design of Two-Level Fractional Factorial Experiments. *Journal of Applied Statistics* **23** (1), 21–29.
- Goos, P. (2002). *The Optimal Design of Blocked and Split-Plot Experiments*. Springer, New York.
- Goos, P., Langhans, I., and Vandebroek, M. (2006). Practical Inference from Industrial Split-Plot Designs. *Journal of Quality Technology* **38** (2), 162–179.
- Goos, P. and Vandebroek, M. (2004). Outperforming Completely Randomized Designs. *Journal of Quality Technology* **36** (1), 12–26.
- Gray, J. B. and Woodall, W. H. (1994). The Maximum Size of Standardized and Internally Studentized Residuals in Regression Analysis. *The American Statistician* **48** (2), 111–113.
- Green, P. E. and Devita, M. T. (1975). Interaction Model of Consumer Utility. *Journal of Consumer Research* **2** (2), 146–153.
- Grundy, P. M. and Healy, M. J. R. (1950). Restricted Randomization and Quasi-Latin Squares. *Journal of the Royal Statistical Society, Ser. B* **12** (2), 286–291.
- Gunter, B. H. (1993). A Systematic Approach to Planning for a Designed Industrial Experiment. Discussion. *Technometrics* **35** (1), 13–14.
- Haaland, P. D. (1998). Analyzing Unreplicated Factorial Experiments: A Review with Some New Proposals. Comment. *Statistica Sinica* **8** (1), 31–35.
- Haaland, P. D. and O'Connell, M. A. (1995). Inference for Effect-Saturated Fractional Factorials. *Technometrics* **37** (1), 82–93.
- Hahn, G. J. (1984). Experimental Design in the Complex World. *Technometrics* **26** (1), 19–31.
- Hamada, M. and Balakrishnan, N. (1998). Analyzing Unreplicated Factorial Experiments: A Review with Some New Proposals. *Statistica Sinica* **8** (1), 1–28.
- Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association* **72** (2), 320–338.
- Härdle, W. and Simar, L. (2007). *Applied Multivariate Statistical Analysis*, 2nd Edition. Springer-Verlag, Berlin.
- Hebble, T. L. and Mitchell, T. J. (1972). Repairing Response Surface Designs. *Technometrics* **14** (3), 767–779.
- Hedayat, A. S. and Pesotan, H. (1992). Two-Level Factorial Designs for Main Effects and Selected Two-Factor Interactions. *Statistica Sinica* **2** (2), 453–464.

- Hedayat, A. S. and Pesotan, H. (1997). Designs for Two-Level Factorial Experiments with Linear Models Containing Main Effects and Selected Two-Factor Interactions. *Journal of Statistical Planning and Inference* **64** (1), 109–124.
- Hedayat, A. S., Sloane, N. J. A., and Stufken, J. (1999). *Orthogonal Arrays: Theory and Applications*. Springer, New York.
- Hoàng, E. M., Liauw, C. M., Allen, N. S., Fontán, E., and Lafuente, P. (2004). Effect of Additive Interactions on the Thermo-Oxidative Stabilization of a Film Grade Metallocene LLDPE. *Journal of Vinyl and Additive Technology* **10** (3), 149–156.
- Holcomb, D. R., Montgomery, D. C., and Carlyle, W. M. (2003). Analysis of Supersaturated Designs. *Journal of Quality Technology* **35** (1), 13–27.
- Holms, A. G. (1998). Design of Experiments as Expansible Sequences of Orthogonal Blocks with Crossed-Classification Block Effects. *Technometrics* **40** (3), 244–253.
- Holms, A. G. and Sidik, S. M. (1971). Design of Experiments as Doubly-Telescoping Sequences of Blocks with Application to a Nuclear Reactor Experiment. *Technometrics* **13** (3), 559–574.
- Hsieh, P. I. and Goodwin, D. E. (1986). Sheet Molded Compound Process Improvement, In: *Fourth Symposium on Taguchi Methods*, American Supplier Institute, Dearborn, MI, pp. 13–21.
- Hu, C. C. and Bai, A. (2001). Composition Control of Electroplated Nickel-Phosphorus Deposits. *Surface & Coatings Technology* **137** (2–3), 181–187.
- Huang, P., Chen, D. C., and Voelkel, J. O. (1998). Minimum Aberration Two-Level Split-Plot Designs. *Technometrics* **40** (4), 314–326.
- Ingram, D. and Tang, B. X. (2005). Minimum G Aberration Design Construction and Design Tables for 24 Runs. *Journal of Quality Technology* **37** (2), 101–114.
- Irvine, G. M., Clark, N. B., and Recupero, C. (1996). Extended Delignification of Mature and Plantation Eucalypt Wood. 2. The Effects of Chip Impregnation Factors. *Appita Journal* **49** (5), 347–352.
- Jacroux, M. (2006). Optimal Sequencing of Blocked 2^{m-k} Fractional Factorial Designs. *Journal of Statistical Planning and Inference* **136** (4), 1473–1492.
- John, P. W. M. (1961). Three-Quarter Replicates of 2^4 and 2^5 Designs. *Biometrics* **17** (2), 319–321.
- John, P. W. M. (1962). Three-Quarter Replicates of 2^n Designs. *Biometrics* **18** (2), 172–184.
- John, P. W. M. (1969). Some Non-Orthogonal Fractions of 2^n Designs. *Journal of the Royal Statistical Society, Ser. B* **31** (2), 270–275.
- John, P. W. M. (1990). Time Trends and Factorial Experiments. *Technometrics* **32** (3), 275–282.
- John, P. W. M. (1998). *Statistical Design and Analysis of Experiments*. Classics in Applied Mathematics. SIAM, Philadelphia.

- Jones, B. and DuMouchel, W. (1996). Follow-Up Designs to Resolved Confounding in Multifactor Experiments. Discussion. *Technometrics* **38** (4), 323–326.
- Jones, B., Lin, D. K. J., and Nachtsheim, C. J. (2008). Bayesian D-Optimal Supersaturated Designs. *Journal of Statistical Planning and Inference* **138** (1), 86–92.
- Jones, K. F., Marrs, I., Young, J., and Townend, C. (1995). Investigation and Improvement of a Process for Vacuum-Formed Ceramic Fiber Composites, Using a Fractional 2^n Experimental Design. *Journal of Applied Statistics* **22** (4), 459–467.
- Joshi, A., Yi, J. J., Bell, R. H. Jr., Eeckhout, L., John, L., and Lilja, D. (2006). Evaluating the Efficacy of Statistical Simulation for Design Space Exploration. *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software*, pp. 70–79.
- Ju, H. L. and Lucas, J. M. (2002). L^k Factorial Experiments with Hard-To-Change and Easy-To-Change Factors. *Journal of Quality Technology* **34** (4), 411–421.
- Kalil, S. J., Maugeri, F., and Rodrigues, M. I. (2000). Response Surface Analysis and Simulation as a Tool for Bioprocess Design and Optimization. *Process Biochemistry* **35** (6), 539–550.
- Katsaounis, T. I. and Dean, A. M. (2008). A Survey and Evaluation of Methods for Determination of Combinatorial Equivalence of Factorial Designs. *Journal of Statistical Planning and Inference* **138** (1), 245–258.
- Ke, W. M. and Tang, B. X. (2003). Selecting 2^{m-p} Designs Using a Minimum Aberration Criterion When Some Two-Factor Interactions Are Important. *Technometrics* **45** (4), 352–360.
- Ke, W. M., Tang, B. X., and Wu, H. Q. (2005). Compromise Plans with Clear Two-Factor Interactions. *Statistica Sinica* **15** (3), 709–715.
- Kelly, H. W. and Voelkel, J. O. (2000). Asymptotic Power Problems in the Analysis of Supersaturated Designs. *Statistics & Probability Letters* **47** (4), 317–324.
- Kerr, K. F. (2006). Efficient 2^k Factorial Designs for Blocks of Size 2 with Microarray Applications. *Journal of Quality Technology* **38** (4), 309–318.
- Kiefer, J. (1960). Optimum Experimental Design V, with Applications to Systematic and Rotatable Designs. *Proceedings of the 4th Berkeley Symposium*, vol. 1, pp. 381–405.
- Kimel, M. T., Benjamini, Y., and Steinberg, D. M. (2008). The False Discovery Rate for Multiple Testing in Factorial Experiments. *Technometrics* **50** (1), 32–39.
- Koukouvinos, C. and Stylianou, S. (2005). A Method for Analyzing Supersaturated Designs. *Communications in Statistics: Simulation and Computation* **34** (4), 929–937.
- Kramschuster, A., Cavitt, R., Ermer, D., Chen, Z. B., and Turng, L. S. (2005). Quantitative Study of Shrinkage and Warpage Behavior for Microcellu-

- lar and Conventional Injection Molding. *Polymer Engineering and Science* **45** (10), 1408–1418.
- Kulahci, M. and Bisgaard, S. (2005). The Use of Plackett–Burman Designs to Construct Split-Plot Designs. *Technometrics* **47** (4), 495–501.
- Kulahci, M., Ramirez, J. G., and Tobias, R. (2006). Split-Plot Fractional Designs: Is Minimum Aberration Enough? *Journal of Quality Technology* **38** (1), 56–64.
- Lai, L. S. T., Pan, C. C., and Tzeng, B. K. (2003). The Influence of Medium Design on Lovastatin Production and Pellet Formation with a High-Producing Mutant of *Aspergillus Terreus* in Submerged Cultures. *Process Biochemistry* **38** (9), 1317–1326.
- Lamb, R. H., Boos, D. D., and Brownie, C. (1996). Testing for Effects on Variance in Experiments with Factorial Treatment Structure and Nested Errors. *Technometrics* **38** (2), 170–177.
- Langhans, I., Goos, P., and Vandebroek, M. (2005). Identifying Effects under a Split-Plot Design Structure. *Journal of Chemometrics* **19** (1), 5–15.
- Langsrud, O. (2001). Identifying Significant Effects in Fractional Factorial Multiresponse Experiments. *Technometrics* **43** (4), 415–424.
- Le Thanh, M., Voilley, A., and Luu, R. P. T. (1993). Influence de la Composition d'un Milieu de Culture Modele sur le Coefficient de Partage Vapeur-Liquide de Substances Aromatisantes. *Sciences Des Aliments* **13**, 699–710 (in French).
- Leitnaker, M. G. and Mee, R. W. (2001). Analytic Use of Two-Level Factorials in Incomplete Blocks to Examine the Stability of Factor Effects. *Quality Engineering* **14** (1), 49–58.
- Lenth, R. V. (1989). Quick and Easy Analysis of Unreplicated Factorials. *Technometrics* **31** (4), 469–473.
- Lewis, S. M. and Dean, A. M. (2001). Detection of Interactions in Experiments on Large Numbers of Factors. *Journal of the Royal Statistical Society, Ser. B* **63** (4), 633–659.
- Li, F. and Jacroux, M. (2007). Optimal Foldover Plans for Blocked 2^{m-k} Fractional Factorial Designs. *Journal of Statistical Planning and Inference* **137** (7), 2439–2452.
- Li, H. and Mee, R. W. (2002). Better Foldover Fractions for Resolution III 2^{k-p} Designs. *Technometrics* **44** (3), 278–283.
- Li, P. F., Liu, M. Q., and Zhang, R. C. (2007). 2^m4^1 Designs with Minimum Aberration or Weak Minimum Aberration. *Statistical Papers* **48** (2), 235–248.
- Li, R. Z. and Lin, D. K. J. (2002). Data Analysis in Supersaturated Designs. *Statistics & Probability Letters* **59** (2), 135–144.
- Li, W. (2008). Analyzing Supersaturated Designs. IMS/SPES Spring Research Conference on Statistics in Industry and Technology, Atlanta.
- Li, W. and Lin, D. K. J. (2003). Optimal Foldover Plans for Two-Level Fractional Factorial Designs. *Technometrics* **45** (2), 142–149.

- Li, W., Lin, D. K. J., and Ye, K. Q. (2003). Optimal Foldover Plans for Two-Level Nonregular Orthogonal Designs. *Technometrics* **45** (4), 347–351.
- Li, W. and Nachtsheim, C. J. (2000). Model-Robust Factorial Designs. *Technometrics* **42** (4), 345–352.
- Li, W. and Wu, C. F. J. (1997). Columnwise-Pairwise Algorithms with Applications to the Construction of Supersaturated Designs. *Technometrics* **39** (2), 171–179.
- Liao, C. T. and Chai, F. S. (2009). Design and Analysis of Two-Level Factorial Experiments with Partial Replication. *Technometrics* **51** (1), 66–74.
- Liao, C. T., Iyer, H. K., and Vecchia, D. F. (1996). Construction of Orthogonal Two-Level Designs of User-Specified Resolution Where $N \neq 2^k$. *Technometrics* **38** (4), 342–353.
- Lin, D. K. J. (1993). A New Class of Supersaturated Designs. *Technometrics* **35** (1), 28–31.
- Lin, D. K. J. (1995). Generating Systematic Supersaturated Designs. *Technometrics* **37** (2), 213–225.
- Lin, D. K. J. and Draper, N. R. (1992). Projection Properties of Plackett and Burman Designs. *Technometrics* **34** (4), 423–428.
- Littell, R. C., Milliken, G. M., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. (2006). *SAS for Mixed Models*, 2nd Edition. SAS Institute, Inc., Cary, NC.
- Liu, M. Q. and Zhang, R. C. (2000). Construction of $E(s^2)$ Optimal Supersaturated Designs Using Cyclic BIBDs. *Journal of Statistical Planning and Inference* **91** (1), 139–150.
- Liu, Y. F. and Dean, A. (2004). k -Circulant Supersaturated Designs. *Technometrics* **46** (1), 32–43.
- Loepky, J. L. and Sitter, R. R. (2002). Analyzing Unreplicated Blocked or Split-Plot Fractional Factorial Designs. *Journal of Quality Technology* **34** (3), 229–243.
- Loh, W. Y. (2006). Regression Tree Models for Designed Experiments. In: *Optimality: The Second Erich L. Lehmann Symposium, Lecture Notes—Monograph Series*, Rojo, J. (Ed.), Institute of Mathematical Statistics, pp. 210–228.
- Lopez-Allvarez, T. and Aguirre-Torres, V. (1997). Improving Field Performance by Sequential Experimentation: A Successful Case Study in the Chemical Industry. *Quality Engineering* **9** (3), 391–403.
- Lu, X. and Wu, X. (2004). A Strategy of Searching Active Factors in Supersaturated Screening Experiments. *Journal of Quality Technology* **36** (4), 392–399.
- Lucas, J. M. (1974). Optimum Composite Designs. *Technometrics* **16** (4), 561–567.
- Lynch, R. O. (1993). Minimum Detectable Effects for 2^{k-p} Experimental Plans. *Journal of Quality Technology* **25** (1), 12–17.
- Lynch, R. O. and Markle, R. J. (1997). Understanding the Nature of Variability in a Dry Etch Process. In: *Statistical Case Studies for Industrial Process*

- Improvement*, Czitrom, V. and Spagon, P. D. (Eds.), SIAM, Philadelphia, pp. 71–85.
- Ma, C. X. and Fang, K. T. (2001). A Note on Generalized Aberration in Factorial Designs. *Metrika* **53** (1), 85–93.
- Margolin, B. H. (1969). Results on Factorial Designs of Resolution IV for 2^n and 2^n3^m Series. *Technometrics* **11** (3), 431–444.
- Martin, C. and Cuellar, J. (2004). Synthesis of Poly(Styrene-Co-Divinylbenzene)-Stainless Steel Beads Through a Factorial Design of Experiments. *Industrial & Engineering Chemistry Research* **43** (9), 2093–2103.
- Martinez, E., Cela, R., Carro, A. M., Cobas, J. C., and Garcia, B. (2002). Chemometrically Guided Sample Composition for Fast Screening of Trace Metals in Water Samples. *Journal of Analytical Atomic Spectrometry* **17** (10), 1373–1380.
- Martinez, E., Landin, P., Carro, A. M., Llompart, M. P., and Cela, R. (2002). Strategically Designed Sample Composition for Fastest Screening of Polychlorinated Biphenyl Congeners in Water Samples. *Journal of Environmental Monitoring* **4** (4), 490–497.
- McLeod, R. G. and Brewster, J. F. (2004). The Design of Blocked Fractional Factorial Split-Plot Experiments. *Technometrics* **46** (2), 135–146.
- Mee, R. W. (2001). Noncentral Composite Designs. *Technometrics* **43** (1), 34–43.
- Mee, R. W. (2004). Efficient Two-Level Designs for Estimating All Main Effects and Two-Factor Interactions. *Journal of Quality Technology* **36** (4), 400–412. Corrigendum 2005, p. 90.
- Mee, R. W. (2007). Optimal Three-Level Designs for Response Surfaces in Spherical Experimental Regions. *Journal of Quality Technology* **39** (4), 340–354.
- Mee, R. W. and Bates, R. L. (1998). Split-Lot Designs: Experiments for Multistage Batch Processes. *Technometrics* **40** (2), 127–140.
- Mee, R. W. and Peralta, M. (2000). Semifolding 2^{k-p} Designs. *Technometrics* **42** (2), 122–134.
- Mee, R. W. and Romanova, A. (2009). Constructing Two-Level Trend-Robust Designs. Technical Report 2009-02, University of Tennessee, Department of Statistics, Operations, and Management Science.
- Mee, R. W. and Xiao, J. (2008a). Optimal Foldovers and Semifolding for Minimum Aberration Even Fractional Factorial Designs. *Journal of Quality Technology* **40** (4), 448–460.
- Mee, R. W. and Xiao, J. (2008b). Steepest Ascent for Multiple-Response Applications. *Technometrics* **50** (4), 371–382.
- Meyer, R. D., Steinberg, D. M., and Box, G. (1996). Follow-Up Designs to Resolve Confounding in Multifactor Experiments. *Technometrics* **38** (4), 303–313.
- Miller, A. (1997). Strip-Plot Configurations of Fractional Factorials. *Technometrics* **39** (2), 153–161.

- Miller, A. (2005). The Analysis of Unreplicated Factorial Experiments Using All Possible Comparisons. *Technometrics* **47** (1), 51–63.
- Miller, A. and Sitter, R. R. (2001). Using the Folded-Over 12-Run Plackett–Burman Design to Consider Interactions. *Technometrics* **43** (1), 44–55.
- Miller, A. and Sitter, R. R. (2004). Choosing Columns from the 12-Run Plackett–Burman Design. *Statistics & Probability Letters* **67** (2), 193–201.
- Miller, A. and Sitter, R. R. (2005). Using Folded-Over Nonorthogonal Designs. *Technometrics* **47** (4), 502–513.
- Miller, J. J. (1978). Inverse of Freeman–Tukey Double Arcsine Transformation. *The American Statistician* **32** (4), 138.
- Mitchell, T. J. (1974). Algorithm for Construction of D-Optimal Experimental Designs. *Technometrics* **16** (2), 203–210.
- Moen, R. D., Nolan, T. W., and Provost, L. P. (1998). *Improving Quality Through Planned Experimentation*, 2nd Edition. McGraw Hill, New York.
- Montgomery, D. C. and Peck, E. A. (1992). *Introduction to Linear Regression*, 2nd Edition. Wiley, New York.
- Montgomery, D. C. and Runger, G. C. (1996). Foldovers of 2^{k-p} Resolution IV Experimental Designs. *Journal of Quality Technology* **28** (4), 446–450.
- Mukerjee, R. and Wu, C. F. J. (1995). On the Existence of Saturated and Nearly Saturated Asymmetrical Orthogonal Arrays. *Annals of Statistics* **23** (6), 2102–2115.
- Mukerjee, R. and Wu, C. F. J. (2001). Minimum Aberration Designs for Mixed Factorials in Terms of Complementary Sets. *Statistica Sinica* **11** (1), 225–239.
- Mukerjee, R. and Wu, C. F. J. (2006). *A Modern Theory of Factorial Designs*. Springer, New York.
- Nair, V. N., Taam, W., and Ye, K. Q. (2002). Analysis of Functional Responses from Robust Design Studies. *Journal of Quality Technology* **34** (4), 355–370.
- Nelder, J. A. and Lane, P. W. (1995). The Computer Analysis of Factorial Experiments: In Memoriam—Frank Yates. *The American Statistician* **49** (4), 382–385.
- Nguyen, N. K. (1996). An Algorithmic Approach to Constructing Supersaturated Designs. *Technometrics* **38** (1), 69–73.
- Nguyen, N. K. (2001). Cutting Experimental Designs into Blocks. *Australian & New Zealand Journal of Statistics* **43** (3), 367–374.
- Nguyen, N. K. and Dey, A. (1989). Computer-Aided Construction of D-Optimal 2^m Fractional Factorial Designs of Resolution V. *Australian Journal of Statistics* **31** (1), 111–117.
- Nguyen, N. K. and Miller, A. J. (1997). 2^m Fractional Factorial Designs of Resolution V with High A-Efficiency, $7 \leq m \leq 10$. *Journal of Statistical Planning and Inference* **59** (2), 379–384.
- O'Brien, R. G. (1979). A General ANOVA Method for Robust Tests of Additive Models for Variances. *Journal of the American Statistical Association* **74** (368), 877–880.

- O'Brien, R. G. (1981). A Simple Test for Variance Effects in Experimental Designs. *Psychological Bulletin* **89** (3), 570–574.
- Olguin, J. and Fearn, T. (1997). A New Look at Half-Normal Plots for Assessing the Significance of Contrasts for Unreplicated Factorials. *Applied Statistics* **46** (4), 449–462.
- Paniagua-Quiñones, C. and Box, G. E. P. (2008). Use of Strip-Strip-Block Design for Multi-Stage Processes to Reduce Cost of Experimentation. *Quality Engineering* **20** (1), 46–52.
- Paniagua-Quiñones, C. and Box, G. E. P. (2009). A Post-Fractionated Strip-Strip-Block Design for Multi-Stage Processes. *Quality Engineering* **21** (2), 156–167.
- Pensado, L., Blanco, E., Casais, M. C., Mejuto, M. C., Martinez, E., Carro, A. M., and Cela, R. (2004). Strategic Sample Composition in the Screening of Polycyclic Aromatic Hydrocarbons in Drinking Water Samples Using Liquid Chromatography with Fluorimetric Detection. *Journal of Chromatography A* **1056** (1–2), 121–130.
- Perutka, J. and Martell, A. E. (2001). Toward Understanding of the Synergistic Oxidation of Adamantane and Hydrogen Sulfide by Molecular Oxygen and with a Dinuclear Iron(II) Macrocyclic Complex as a Catalyst. *Analytica Chimica Acta* **435** (2), 385–391.
- Petersson, P., Lundell, N., and Markides, K. E. (1992). Chiral Separations in Supercritical Fluid Chromatography: A Multivariate Optimization Method. *Journal of Chromatography* **623** (1), 129–137.
- Plackett, R. L. and Burman, J. P. (1946). The Design of Optimum Multifactorial Experiments. *Biometrika* **33** (4), 305–325.
- Poorna, V. and Kulkarni, P. R. (1995). A Study of Inulinase Production in *Aspergillus Niger* Using Fractional Factorial Design. *Bioresource Technology* **54** (3), 315–320.
- Qu, X. G. (2006). A Maximum Estimability Criterion for Design Classification and Selection. *Journal of Statistical Planning and Inference* **136** (8), 2756–2769.
- Qu, X. G. (2007). Statistical Properties of Rechtschaffner Designs. *Journal of Statistical Planning and Inference* **137** (7), 2156–2164.
- Quenouille, M. H. and John, J. A. (1971). Paired Comparison Designs for 2^n Factorials. *Applied Statistics* **20** (1), 16–24.
- Quintana, J. B., Martinez, E., Carro, A. M., Lorenzo, R. A., and Cela, R. (2003). Screening of Polychlorinated Biphenyls in Water Samples by Strategic Sample Composition: Solid Phase Extraction and Gas Chromatography Tandem Mass Spectrometry. Comparison of Different Strategies for Sample Composition. *International Journal of Environmental Analytical Chemistry* **83** (4), 269–284.
- Rahni, N., Ramdani, N., Candau, Y., and Dalicieux, P. (1997). Application of Group Screening to Dynamic Building Energy Simulation Models. *Journal of Statistical Computation and Simulation* **57** (1–4), 285–304.

- Rechtschaffner, R. L. (1967). Saturated Fractions of 2^n and 3^n Factorial Designs. *Technometrics* **9** (4), 569–575.
- Rocke, D. M. (1993). On the Beta-Transformation Family. *Technometrics* **35** (1), 72–81.
- Rooda, J. E. and van der Schilden, N. (1982). Simulation of Maritime Transport and Distribution by Sea-Going Barges: An Application of Multiple Regression Analysis and Factor Screening. *Bulk Solids Handling* **2** (4), 813–824.
- Rosenbaum, P. R. (1994). Dispersion Effects from Fractional Factorials in Taguchi's Method of Quality Design. *Journal of the Royal Statistical Society Series B-Methodological* **56** (4), 641–652.
- Rosenbaum, P. R. (1996). Some Useful Compound Dispersion Experiments in Quality Design. *Technometrics* **38** (4), 354–364.
- Ross, S. M. (1998). *A First Course in Probability*, 5th Edition. Prentice-Hall, Upper Saddle River, NJ.
- Ryan, K. J. and Bulutoglu, D. A. (2007). $E(s^2)$ -Optimal Supersaturated Designs with Good Minimax Properties. *Journal of Statistical Planning and Inference* **137** (7), 2250–2262.
- Sanders, D. and Coleman, J. (2003). Recognition and Importance of Restrictions on Randomization in Industrial Experimentation. *Quality Engineering* **15** (4), 533–543.
- Sanders, D., Leitnaker, M. G., and McLean, R. A. (2001). Randomized Complete Block Designs in Industrial Studies. *Quality Engineering* **14** (1), 1–8.
- Satterthwaite, F. E. (1959). Random Balance Experimentation. *Technometrics* **1** (2), 111–137.
- Schoen, E. D. (1997). Cheesemaking with GENSTAT: A Case Study in Design of Industrial Experiments. *Genstat Newsletter* **33**, 20–29.
- Schoen, E. D. (1999). Designing Fractional Two-Level Experiments with Nested Error Structures. *Journal of Applied Statistics* **26** (4), 495–508.
- Schoen, E. D. and Kaul, E. A. A. (2000). Three Robust Scale Estimators to Judge Unreplicated Experiments. *Journal of Quality Technology* **32** (3), 276–283.
- Searle, S. R., Casella, G., and McCulloch, C. E. (2006). *Variance Components*. Wiley, New York.
- Seberry, J., Wysocki, B. J., and Wysocki, T. A. (2005). Some Applications of Hadamard Matrices. *Metrika* **62** (2–3), 221–239.
- Sheesley, J. (1985). Use of Factorial Designs in the Development of Lighting Products. In: *Experiments in Industry: Design, Analysis, and Interpretation of Results*, Snee, R. D., Hare, L. B., and Trout, J. R. (Eds.), American Society for Quality Control, Milwaukee, WI, pp. 47–57.
- Shen, Q. and Faraway, J. (2004). An F Test for Linear Models with Functional Responses. *Statistica Sinica* **14** (4), 1239–1257.
- Shen, Q. and Xu, H. (2007). Diagnostics for Linear Models with Functional Responses. *Technometrics* **49** (1), 26–33.

- Silknitter, K. O., Wisnowski, J. W., and Montgomery, D. C. (1999). The Analysis of Covariance: A Useful Technique for Analysing Quality Improvement Experiments. *Quality and Reliability Engineering International* **15** (4), 303–316.
- Sitter, R. R., Chen, J. H., and Feder, M. (1997). Fractional Resolution and Minimum Aberration in Blocked 2^{n-k} Designs. *Technometrics* **39** (4), 382–390.
- Snee, R. D. (1985). Experimenting with a Large Number of Variables. In: *Experiments in Industry: Design, Analysis, and Interpretation of Results*, Snee, R. D., Hare, L. B., and Trout, J. R. (Eds.), American Society for Quality Control, Milwaukee, WI, pp. 25–35.
- Starzec, P. and Andersson, J. (2002). Application of Two-Level Factorial Design to Sensitivity Analysis of Keyblock Statistics from Fracture Geometry. *International Journal of Rock Mechanics and Mining Sciences* **39** (2), 243–255.
- Steinberg, D. M. and Bursztyn, D. (1994). Dispersion Effects in Robust Design Experiments with Noise Factors. *Journal of Quality Technology* **26** (1), 12–20.
- Su, J. and Lua, A. C. (2006). Influence of Carbonisation Parameters on the Transport Properties of Carbon Membranes by Statistical Analysis. *Journal of Membrane Science* **278** (1–2), 335–343.
- Sun, D. X. and Wu, C. F. J. (1993). Statistical Properties of Hadamard Matrices of Order 16. In: *Quality Through Engineering Design*, Kuo, W. (Ed.), Elsevier, Amsterdam, pp. 169–179.
- Sun, D. X., Wu, C. F. J., and Chen, Y. Y. (1997). Optimal Blocking Schemes for 2^n and 2^{n-p} Designs. *Technometrics* **39** (3), 298–307.
- Tack, L. and Vandebroek, M. (2002). An Adjustment Algorithm for Optimal Run Orders in Design of Experiments. *Computational Statistics & Data Analysis* **40** (3), 559–577.
- Taguchi, G. (1987). *System of Experimental Design*, 2nd Edition. UNIPUB, New York.
- Tang, B. X. (2006). Orthogonal Arrays Robust To Nonnegligible Two-Factor Interactions. *Biometrika* **93** (1), 137–146.
- Tang, B. X. and Deng, L. Y. (1999). Minimum G_2 -Aberration for Nonregular Fractional Factorial Designs. *Annals of Statistics* **27** (6), 1914–1926.
- Tang, B. X., Ma, F. S., Ingram, D., and Wang, H. (2002). Bounds on the Maximum Number of Clear Two-Factor Interactions for 2^{m-p} Designs of Resolution III and IV. *Canadian Journal of Statistics* **30** (1), 127–136.
- Tang, B. X. and Wu, C. F. J. (1996). Characterization of Minimum Aberration 2^{n-k} Designs in Terms of Their Complementary Designs. *Annals of Statistics* **24** (6), 2549–2559.
- Tang, B. X. and Wu, C. F. J. (1997). A Method for Constructing Super-saturated Designs and Its $E\bar{s}^2$ Optimality. *Canadian Journal of Statistics* **25** (2), 191–201.

- Tobias, R. (1996). Saturated Second-Order Two-Level Designs: An Empirical Approach. Technical report, SAS Institute, available from <http://support.sas.com/rnd/app/papers/minres5.pdf>.
- Trinca, L. A. and Gilmour, S. G. (2000). An Algorithm for Arranging Response Surface Designs in Small Blocks. *Computational Statistics & Data Analysis* **33** (1), 25–43.
- Tyssedal, J. and Kulahci, M. (2005). Analysis of Split-Plot Designs with Mirror Image Pairs as Sub-Plots. *Quality and Reliability Engineering International* **21** (5), 539–551.
- Vander Heyden, Y., Kuttatharmmakkul, S., Smeyers-Verbeke, J., and Massart, D. L. (2000). Supersaturated Designs for Robustness Testing. *Analytical Chemistry* **72** (13), 2869–2874.
- Venter, J. H. and Steel, S. J. (1998). Identifying Active Contrasts by Stepwise Testing. *Technometrics* **40** (4), 304–313.
- Vindevogel, J. and Sandra, P. (1991). Resolution Optimization in Micellar Electrokinetic Chromatography: Use of Plackett–Burman Statistical Design for the Analysis of Testosterone Esters. *Analytical Chemistry* **63** (15), 1530–1536.
- Vine, A. E., Lewis, S. M., Dean, A. M., and Brunson, D. (2008). A Critical Assessment of Two-Stage Group Screening Through Industrial Experimentation. *Technometrics* **50** (1), 15–25.
- Vivacqua, C. A. and Bisgaard, S. (2009). Post-Fractionated Strip-Block Designs. *Technometrics* **51** (1), 47–55.
- Voss, D. T. and Wang, W. Z. (2006). On Adaptive Testing in Orthogonal Saturated Designs. *Statistica Sinica* **16** (1), 227–234.
- Wakeling, I. N., Hasted, A., and Buck, D. (2001). Cyclic Presentation Order Designs for Consumer Research. *Food Quality and Preference* **12** (1), 39–46.
- Walker, E. and Wright, S. P. (2002). Comparing Curves Using Additive Models. *Journal of Quality Technology* **34** (1), 118–129.
- Wang, J. X., Dipasquale, A. J., Bray, A. M., Maeji, N. J., and Geysen, H. M. (1993). Study of Stereo-Requirements of Substance-P Binding to NK1 Receptors Using Analogs with Systematic D-Amino-Acid Replacements. *Bioorganic & Medicinal Chemistry Letters* **3** (3), 451–456.
- Wang, P. C. and Jan, H. W. (1995). Designing Two-Level Factorial Experiments Using Orthogonal Arrays When the Run Order Is Important. *Statistician* **44** (3), 379–388.
- Watson, G. S. (1961). Study of Group Screening Method. *Technometrics* **3** (3), 371–388.
- Webb, D. F., Lucas, J. M., and Borkowski, J. J. (2004). Factorial Experiments when Factor Levels Are Not Necessarily Reset. *Journal of Quality Technology* **36** (1), 1–11.
- Webb, S. (1968). Non-Orthogonal Designs of Even Resolution. *Technometrics* **10** (2), 291–299.

- Westfall, P. H., Young, S. S., and Lin, D. K. J. (1998). Forward Selection Error Control in the Analysis of Supersaturated Designs. *Statistica Sinica* **8** (1), 101–117.
- Wheeler, D. J. and Lyday, R. W. (1989). *Evaluating the Measurement Process*, 2nd Edition. SPC Press, Knoxville, TN.
- Williams, E. J. (1949). Experimental Designs Balanced for the Estimation of Residual Effects of Treatments. *Australian Journal of Scientific Research Series A: Physical Sciences* **2** (2), 149–168.
- Wu, C. F. J. (1989). Construction of 2^m4^n Designs via a Grouping Scheme. *Annals of Statistics* **17** (4), 1880–1885.
- Wu, C. F. J. (1993). Construction of Supersaturated Designs Through Partially Aliased Interactions. *Biometrika* **80** (3), 661–669.
- Wu, C. F. J. and Chen, Y. (1991). A Graph-Aided Method for Planning Two-Level Experiments When Certain Interactions Are Important. Technical Report IIQP 91-07, University of Waterloo, Department of Statistics and Actuarial Science.
- Wu, C. F. J. and Chen, Y. Y. (1992). A Graph-Aided Method for Planning Two-Level Experiments When Certain Interactions Are Important. *Technometrics* **34** (2), 162–175.
- Wu, C. F. J. and Hamada, M. (2000). *Experiments: Planning, Analysis and Parameter Design Optimization*. Wiley, New York.
- Wu, C. F. J. and Zhang, R. C. (1993). Minimum Aberration Designs with Two-Level and Four-Level Factors. *Biometrika* **80** (1), 203–209.
- Wu, H. Q., Mee, R. W., and Tang, B. X. (2008). Fractional Factorial Designs with Admissible Sets of Clear Two-Factor Interactions. Technical Report Preprint 2008-10, Iowa State University, Department of Statistics and Statistical Laboratory.
- Wu, H. Q. and Wu, C. F. J. (2002). Clear Two-Factor Interactions and Minimum Aberration. *Annals of Statistics* **30** (5), 1496–1511.
- Wu, S. S., Mee, R. W., and Ford, J. (2009). Step-up Test with a Single Cutoff for Orthogonal Saturated Designs. Technical Report 2009-03, University of Tennessee, Department of Statistics, Operations and Management Science.
- Wu, S. S. and Wang, W. Z. (2008). A Note on Step-Up Test in Orthogonal Saturated Designs. *Journal of Statistical Planning and Inference* **138** (10), 3149–3156.
- Wu, W., Shaw, P., Ruan, J., Spence, F. J., and Wildsmith, S. E. (2005). Experimental Designs for Optimisation of the Image Analysis Process for cDNA Microarrays. *Chemometrics and Intelligent Laboratory Systems* **76** (2), 175–184.
- Xu, H. (2002). An Algorithm for Constructing Orthogonal and Nearly Orthogonal Arrays with Mixed Levels and Small Runs. *Technometrics* **44** (4), 356–368.
- Xu, H. (2005). Some Nonregular Designs from the Nordstrom–Robinson Code and Their Statistical Properties. *Biometrika* **92** (2), 385–397.
- Xu, H. (2008). Personal communication.

- Xu, H. (2009). Algorithmic Construction of Efficient Fractional Factorial Designs with Large Run Sizes. UCLA Statistics Electronic Publications, preprint 520 revised.
- Xu, H. and Cheng, C. S. (2008). A Complementary Design Theory for Doubling. *Annals of Statistics* **36** (1), 445–457.
- Xu, H. and Deng, L. Y. (2005). Moment Aberration Projection for Nonregular Fractional Factorial Designs. *Technometrics* **47** (2), 121–131.
- Xu, H. and Lau, S. (2006). Minimum Aberration Blocking Schemes for Two- and Three-Level Fractional Factorial Designs. *Journal of Statistical Planning and Inference* **136** (11), 4088–4118.
- Yamada, S. and Lin, D. K. J. (1997). Supersaturated Design Including an Orthogonal Base. *Canadian Journal of Statistics* **25** (2), 203–213.
- Yang, G. J. and Liu, M. Q. (2006). A Note on the Lower Bounds on Maximum Number of Clear Two-Factor Interactions for 2^{m-p}_{IV} Designs. *Communications in Statistics: Theory and Methods* **35** (5), 849–860.
- Yang, J. F., Zhang, R. C., and Liu, M. Q. (2007). Construction of Fractional Factorial Split-Plot Designs with Weak Minimum Aberration. *Statistics & Probability Letters* **77** (15), 1567–1573.
- Yang, W. Z., Beauchemin, K. A., and Rode, L. M. (2001). Effects of Grain Processing, Forage to Concentrate Ratio, and Forage Particle Size on Rumen pH and Digestion by Dairy Cows. *Journal of Dairy Science* **84** (12), 2203–2216.
- Yang, X., Shen, Q., Xu, H., and Shoptaw, S. (2007). Functional Regression Analysis Using an F Test for Longitudinal Data with Large Numbers of Repeated Measures. *Statistics in Medicine* **26** (7), 1552–1566.
- Yang, Y. Y. J. and Draper, N. R. (2003). Two-Level Factorial and Fractional Factorial Designs in Blocks of Size Two. *Journal of Quality Technology* **35** (3), 294–305.
- Yates, F. (1935). Complex Experiments. *Journal of the Royal Statistical Society* **2** (2 Suppl.), 181–247.
- Yates, F. (1937). *The Design and Analysis of Factorial Experiments*. Imperial Bureau of Soil Science, Harpenden, England.
- Yates, P. and Mee, R. W. (2000). Fractional Factorial Designs That Restrict the Number of Treatment Combinations for Factor Subsets. *Quality and Reliability Engineering International* **16** (5), 343–354.
- Ye, K. Q. and Hamada, M. (2000). Critical Values of the Lenth Method for Unreplicated Factorial Designs. *Journal of Quality Technology* **32** (1), 57–66.
- Ye, K. Q., Hamada, M., and Wu, C. F. J. (2001). A Step-Down Lenth Method for Analyzing Unreplicated Factorial Designs. *Journal of Quality Technology* **33** (2), 140–152.
- Yi, J. J., Lilja, D. J., and Hawkins, D. M. (2005). Improving Computer Architecture Simulation Methodology by Adding Statistical Rigor. *IEEE Transactions on Computers* **54** (11), 1360–1373.

- Youden, W. J., Kempthorne, O., Tukey, J. W., Box, G. E. P., and Hunter, J. S. (1959). Discussion of the Papers of Messrs. Satterthwaite and Budne. *Technometrics* **1** (2), 157–193.
- Young, J. C., Abraham, B., and Whitney, J. B. (1991). Design Implementation in a Foundry: A Case Study. *Quality Engineering* **3** (2), 167–180.
- Young, S. S. and Hawkins, D. M. (1995). Analysis of a 2^9 Full Factorial Chemical Library. *Journal of Medicinal Chemistry* **38** (14), 2784–2788.
- Yuan, S. Z., Murch, L., and Goss, W. P. (2003). Ruggedness Experiments for a Calibrated Hot Box Measuring Fenestration Systems Thermal Transmittance. *Journal of Testing and Evaluation* **31** (1), 44–51.
- Zhang, R. C. and Shao, Q. (2001). Minimum Aberration $(S^2)S^{n-k}$ Designs. *Statistica Sinica* **11** (1), 213–223.
- Zhao, Q. Y., Zhang, W., Jin, M. F., Yu, X. J., and Deng, M. C. (2005). Formulation of a Basal Medium for Primary Cell Culture of the Marine Sponge Hymeniacidon Perleve. *Biotechnology Progress* **21** (3), 1008–1012.
- Zhou, J. (2001). A Robust Criterion for Experimental Designs for Serially Correlated Observations. *Technometrics* **43** (4), 462–467.

Abbreviations and Symbols

Symbol	Definition	First-Use Section
α	Probability of a Type I error for an hypothesis test	2.2
α	Prior probability in Box–Meyer’s Bayesian procedure	2.5.2
α	Parameter determining the spacing for axial points	12.2
β	Probability of a Type II error = 1 – Power	13.1
β	Vector of true regression coefficients	1.3
β_i	Individual true regression coefficient	1.2
ϵ	Random error for assumed model	1.2
ϵ	Vector of random errors for assumed model	1.3
λ	Exponent in Box–Cox transformation	2.7
	Noncentrality parameter in the calculation of power	13.1
ρ	Correlation	6.3.1
Σ	Summation	1.2
σ^2	True error variance, $\text{Var}(\epsilon)$	1.3
A_j	Number of words of length j in defining relation of a regular fractional factorial	5.2.5
$A_j(\dots)$	Component of the confounding frequency vector for a nonregular design	6.3
a_j	Number of alias sets of two-factor interactions of size j	7.2.1
alp	Alias length pattern, (a_1, a_2, \dots, a_L)	7.2.1
ANOVA	Analysis of variance	1.3
A-opt	Optimal design criterion based on $\text{trace}(\mathbf{X}'\mathbf{X})^{-1}$	6.4
APC	All possible comparisons	14.2.1

Symbol	Definition	First-Use Section
B	Matrix of estimated second-order coefficients	12.1
b	Vector of least squares estimates	1.3
b	Vector of estimated first-order coefficients	12.1
<i>b</i>	Used to define the number of blocks, 2^b	3.3
b_0	Intercept for a fitted model	1.3
b_j	Estimated coefficient for factor j	1.3
B_j	Component of the generalized word length pattern	6.3
$b_{i \cdot i}$	Estimated coefficient for x_i^2	12.1
$b_{i \cdot j}$	Estimated coefficient for $x_i * x_j$ interaction	1.3
<i>c</i>	Sample count, either Binomial or Poisson	2.8
cfv	Confounding frequency vector	6.3
D	$N \times k$ design matrix	6.3
det	Determinant	6.4
df	Degrees of freedom	2.2
D-opt	Optimal design criterion based on $\det(\mathbf{X}'\mathbf{X})^{-1}$	6.4
$E(\cdot)$	Expected value	1.2
<i>e</i>	For designs with blocking, the maximum order for which all effects are estimable	3.3
e_i	Residual, $y_i - \hat{y}_i$	2.6
EER	Experimentwise (Type I) error rate	2.4.2
<i>f</i>	Defines the fraction; i.e., 2^{k-f} is a $(1/2)^f$ fraction	5.2.4
F_{ν_1, ν_2}	F random variable with df (ν_1, ν_2)	1.3
FDR	False discovery rate	14.2.2
FRD	Factor relationship diagram	4.3.3
FT	Freeman–Tukey transformations for Binomial or Poisson	2.8
gwlp	Generalized word length pattern, (B_3, B_4, \dots, B_k)	6.3
H _N	Hadamard matrix	6.3.1
h_{ii}	Diagonal element of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$	2.6.4
I	Identity column of +1's	5.2.1
I _{ν}	Identity matrix	1.3
IER	Individual test (Type I) error rate	2.4.2
$J_r(s)$	J-characteristic, sum of r -factor interaction column	6.3
<i>K</i>	Variance ratio in Box–Meyer’s Bayesian procedure	2.5.2
<i>k</i>	Number of factors	1.2

Symbol	Definition	First-Use Section
lof	Lack-of-fit	2.2
M	Degrees of freedom for two-factor interactions	7.2.1
MS	Mean square = SS/df	2.2
MSE	Mean square error	2.2
N	Number of factorial runs in design	2.2
n	Number of replications at each observed t.c.	2.2
n_0	Number of centerpoint replicates	2.3.1
p	True proportion for Binomial count data	2.8.1
\hat{p}	Sample proportion for Binomial count data	2.8.1
P_i	Proportion used for quantiles in normal effects plot	2.5
pe	“Pure error,” estimate for σ based entirely on replication	1.3
PSE	Lenth’s pseudo-standard-error	2.4.1
Q_i	Proportion used for quantiles in half-normal effects plot	2.5
R^2	Coefficient of determination	2.4
r	Number of parameters in a reduced model	2.2
r	Sample correlation coefficient	2.8.4
red	Abbreviation for “Reduced”	2.2
S_N	Sylvester-type Hadamard matrix	6.2
s^2	Sample variance, computed from within-run sampling	2.8.3
s_0	Preliminary estimate in Lenth’s procedure	2.5
sat	Abbreviation for “Saturated”	1.3
SOS	Second-order saturated	7.2.2
SS	Sum of squares	1.3
T	Row coincidence matrix	6.3.2
t.c.	Treatment combination	1.1
Var	Variance	1.3
VIF	Variance inflation factor	6.4
wlp	Word length pattern, (A_3, A_4, \dots, A_k)	5.2.1
\mathbf{X}	A model matrix	1.3
x_j	j^{th} coded factor	1.3
	Upper case boldface letters also denote coded factors	

Symbol	Definition	First-Use Section
\mathbf{Y}	Vector of values of y_i	1.3
$\hat{\mathbf{Y}}$	Vector of predicted values	2.2
y_i	i^{th} value of the response variable	1.3
\hat{y}	Predicted response for a fitted model	1.3
Z_{P_i}	Standard normal quantile used in normal plot of effects	2.5.1
Z_{Q_i}	Standard normal quantile used in half-normal plot	2.5.1

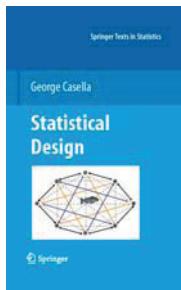
Index

- A-optimality, 227
Adaptive standard error (ASE), 440
Additivity of blocks and treatments, 79
Alias, 153
Alias length pattern, 268, 272, 358
Alias matrix, 196, 229, 511
All possible comparisons (APC), 446
Alternative hypothesis, 18
Analysis of variance, 16, 29, 34, 37, 92, 94, 96, 119, 133, 182, 300, 309, 381
Analysis strategy for full factorials, 27–29
Arcsin transformation, 58
Assumptions for fractional factorials, 154
Autocorrelation, 50
Avoiding cancellation of aliased estimates, 336, 386
Axial point for:
 asymmetric composite design, 406–407
 central composite design, 399–403
Basic factors, 156–160
Bayesian analysis, 47
Bayesian design augmentation, 342
Binomial distribution, 57, 418
Block size, 85
Block*Factor interactions, 79
Blocking, 25, 77–79, 373, 426, 430–431, 457–460, 468
Blocking for 2^k , 79–97, 104–109
Blocking for fractional factorial, 343–349, 355–358, 389–394
Box–Behnken design, 407–409
Box–Cox transformation, 55
Centerpoint runs, 31–35, 48, 86, 88, 135, 140, 252, 384, 394, 399, 401, 402, 407, 420, 438
Central composite design, 399
Check for curvature, 34
Coefficient of determination, 34
Confidence interval for mean response, 320
Confirmation run, 319
Confounding frequency vector (cfv), 199, 200, 226, 277
Confounding interactions with blocks, 80
Contrast, 41
Crossover design, 105
D-optimal design augmentation, 338–342
D-optimality, 86, 227–232, 244, 280, 296–298, 318, 354, 393, 401, 403, 406
Debarred combinations, 435
Defining contrast subgroup, 152
Defining relation, 152, 157, 177, 191, 267, 270, 271, 289, 294, 329, 330, 332, 333, 343, 365, 435
Degrees of freedom (df), 16
Directional response data, 66
Doubling, 272
Durbin–Watson test, 50

- Effect heredity assumption, 342
 Effect simplicity, 154
 Effect sparsity, 36, 154, 166
 Eigenanalysis, 409
 Estimability of blocking, 80, 483
 Estimation capacity, 268
 Even design, 269–271, 277
 Even/odd design, 271
 Experimentwise error rate (EER), 44, 124, 443–448, 467, 479
 False discovery rate, 447
 Family of fractional factorial designs, 152
 Fixed effects, 457
 Foldover, 186, 261, 269, 277–281, 318, 328–332, 337, 354, 371–373, 375–381, 435
 Freeman–Tukey transformation
 for Binomial, 58, 217
 for Poisson, 62
 Functional response data, 66
 Generalized interaction, 80
 Generalized least squares, 454–456
 Generalized word length pattern (gwlp), 200
 Hadamard design, 198–226
 Hadamard matrix, 190, 198, 208–210, 269, 277, 279
 Half-normal plot of effects, 45, 54, 72, 89, 148, 168, 185, 204, 315, 362, 367
 Hat matrix, 51
 Hierarchical model, 12, 18, 41
 Individual error rate (IER), 42, 444
 Interaction, 11
 Interaction plot, 19–21
 Interblock information, 86, 95–97, 349
 Isomorphic fractions, 159, 193
 Lack-of-fit, 30
 Latin square, 105
 Least squares, 13
 Lenth, 39
 Lenth t statistics, 42–44
 PSE, 39, 42
 Levels, 4, 25, 420–422
 Lifetime data, 66, 202
 Log transformation, 64, 66
 Mean squares, 15
 Minimum aberration, 160, 193, 267–269, 271–274, 282, 284, 337, 346, 350, 353, 425, 427, 483, 487–495
 Minimum G_2 -aberration, 200
 Minimum G -aberration, 199
 Missing data, 70–74, 342
 Mixed model, 457
 Model matrix, 14, 29, 197, 227, 245, 289, 296, 338, 481, 511
 Model simplicity, 149
 Model-dependent estimator for σ , 35, 36
 Nested unit structure, 110
 Noise factor, 353
 Noncentral composite design, 403
 Nonlinearity, 34
 Nonregular design, 194
 Normal plot of effects, 45, 54, 102, 189, 349
 Normality assumption, 50
 Null hypothesis, 17
 Observed significance level, 17
 Order of estimability, 84, 483
 Orthogonal array, 195
 Outliers, 51, 360
 Pareto optimal, 325
 Parsimony, 12, 265
 Partial aliasing, 196, 197, 210, 229, 264, 267, 354
 Partial confounding, 86
 Partial replication, 31, 67
 Performance measure modeling, 360
 Permutation test, 242
 Plackett–Burman design, *see* Hadamard design
 Planning strategy, 23–26
 Poisson distribution, 61, 418
 Power of a test, 415
 Predicted value, 29, 48
 Prediction interval for y , 320
 Principal block, 83

- Product array design, 353, 358
- Profile data, 66
- Profiler graph, 22
- Projection, 150, 192, 196
- Pure error, 15
- Pure quadratic curvature, 34
- Quasi-Latin square, 107
- Random effects, 349, 457
- Randomization, 50
- Rechtschaffner design, 295–296, 311–316
- Reduced model, 29, 30
- REML, 118, 458
- Replication, 4, 29, 31–35, 50, 67–69, 76, 114, 115, 178, 438, 466
- Pseudo, 466
- Replication vs. repeated measurements, 249, 466
- Residuals, 48–52
- Resolution, 152
 - III, 155, 158, 160, 173, 182, 193
 - IV, 160, 270, 271, 282
 - V, 160, 283, 285
- Response modeling, 360
- Response surface methodology (RSM), 397
- Robust parameter design, 277, 353, 423
- Root mean square error (RMSE), 36
- Row coincidence matrix, 200, 513
- Sample size determination, 415–420
- Second-order polynomial model, 399
- Second-order saturated, 269, 272
- Semifolding, 332–338, 342
- Sparsity, *see* Effect sparsity
- Split-unit designs, 97–104, 109–114, 140, 350–355, 358, 359, 361–369, 434, 459
- Square-root transformation, 61
- Standard error, 18
- Steepest ascent, 321–328, 338, 398
- Strength, 196
- Strength-2 OA, 195–226
- Strength-3 OA, 261–267, 277–279
- Strength-4 OA, 285–288
- Strip-block design, 110, 354, 370
- Studentized residual, 51
- Supersaturated design, 226, 231–243
- Three-quarter fraction, 288–293, 303–307, 372
- Trend-robust run order, 429
- True replication, 35, 466
- Two-factor interaction model, 10
- Two-factor interactions
 - clear, 268
 - df for, 268
- Unbalanced data, 67
- Unit structure, 110
- Variance inflation factor (VIF), 228, 229, 394
- Weak heredity, 264, 342
- Weak minimum aberration, 194, 273, 347
- Word length pattern (wlp), 159, 192, 267, 270, 272, 344, 487

Statistical Design

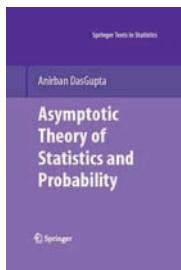


George Casella

The goal of this book is to describe the principles that drive good design, paying attention to both the theoretical background and the problems arising from real experimental situations. Designs are motivated through actual experiments, ranging from the timeless agricultural randomized complete block, to microarray experiments, which naturally lead to split plot designs and balanced incomplete blocks.

2008. XXII, 307 p. (Springer Texts in Statistics) Hardcover
ISBN 978-0-387-75964-7

Asymptotic Theory of Statistics and Probability

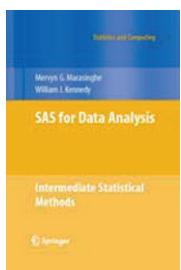


Anirban DasGupta

This book is an encyclopedic treatment of classic as well as contemporary large sample theory, dealing with both statistical problems and probabilistic issues and tools. It is written in an extremely lucid style, with an emphasis on the conceptual discussion of the importance of a problem and the impact and relevance of the theorems. It can be used as a graduate text, as a versatile research reference, as a source for independent reading on a wide assembly of topics, and as a window to learning the latest developments in contemporary topics.

2008. XXVIII, 724 p. (Springer Texts in Statistics) Hardcover
ISBN 978-0-387-75970-8

SAS for Data Analysis Intermediate Statistical Methods



Mervyn Marasinghe William J. Kennedy

This book is an integrated treatment of applied statistical methods, presented at an intermediate level, and the SAS programming language. It serves as an advanced introduction to SAS as well as how to use SAS for the analysis of data arising from many different experimental and observational studies. Particular attention is devoted to discussions of models used in each analysis because the authors believe that it is important for users to have not only an understanding of how these models are represented in SAS but also because it helps in the interpretation of the SAS output produced.

2008. XII, 557 p. (Statistics and Computing) Hardcover
ISBN: 978-0-387-77371-1

Easy Ways to Order ►

Call: Toll-Free 1-800-SPRINGER • E-mail: orders-ny@springer.com • Write: Springer, Dept. S8113, PO Box 2485, Secaucus, NJ 07096-2485 • Visit: Your local scientific bookstore or urge your librarian to order.