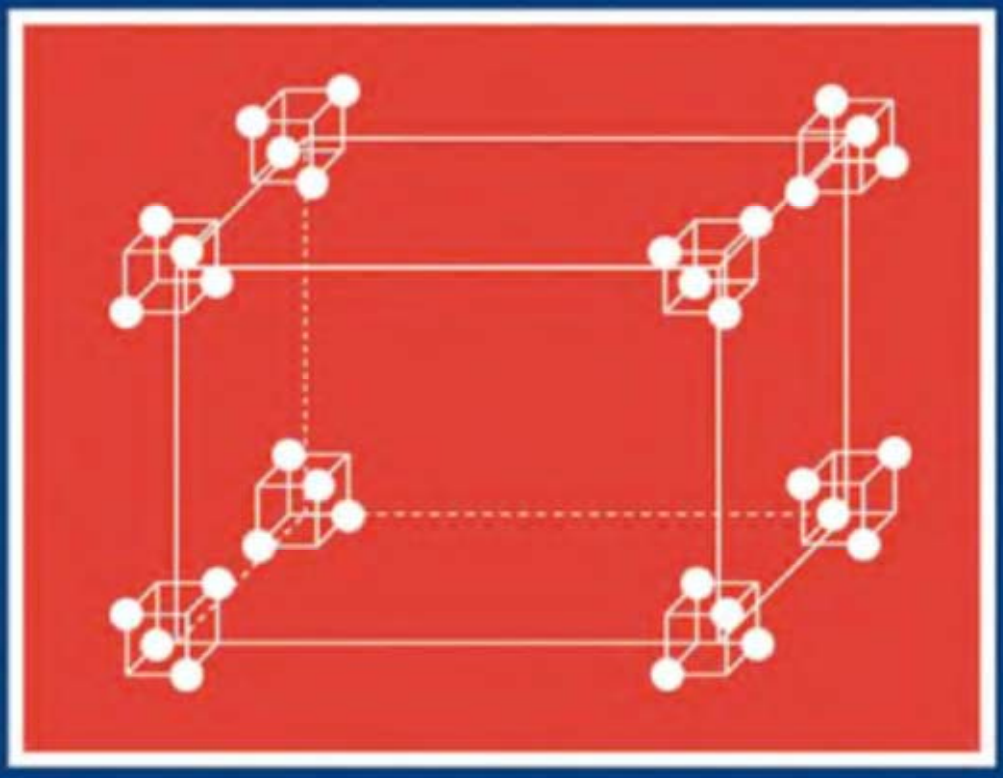
WILEY

Statistical Design and Analysis of Experiments

With Applications to Engineering and Science

Second Edition



Robert L. Mason Richard F. Gunst James L. Hess

WILEY SERIES IN PROBABILITY AND STATISTICS

Statistical Design and Analysis of Experiments

With Applications to Engineering and Science

Second Edition

Robert L. Mason
Southwest Research Institute
San Antonio, Texas

Richard F. Gunst
Department of Statistical Science
Southern Methodist University
Dallas, Texas

James L. Hess
Leggett and Platt, Inc.
Carthage, Missouri



A JOHN WILEY & SONS PUBLICATION

Statistical Design and Analysis of Experiments

Statistical Design and Analysis of Experiments

With Applications to Engineering and Science

Second Edition

Robert L. Mason
Southwest Research Institute
San Antonio, Texas

Richard F. Gunst
Department of Statistical Science
Southern Methodist University
Dallas, Texas

James L. Hess
Leggett and Platt, Inc.
Carthage, Missouri



A JOHN WILEY & SONS PUBLICATION

This book is printed on acid-free paper.

Copyright © 2003 by John Wiley & Sons, Inc. All rights reserved.

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

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

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

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data is available

ISBN 0-471-37216-1

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

of concepts by considering the topics on experimental design in close proximity to one another. In addition, alternatives to the various designs are easily cross-referenced, making the distinctions between the designs clearer. Following the concentrated attention on experimental-design issues, separate chapters immediately provide for the analysis of data from these designs. All too often, texts devote a paragraph to the design of an experiment and several pages to the analysis of the resulting data. Our experiences with this approach are that the material on experimental design is slighted when designs and analyses are presented in the same chapter. A much clearer understanding of proper methods for designing experiments is achieved by separating the topics.

The chapters in Part II concentrate on the design and analysis of experiments with factorial structures. New in the second edition is expanded coverage of statistical graphics (e.g., trellis plots in Chapter 6), three-level and combined two- and three-level fractional factorial experiments (Chapter 7), and expanded coverage on the analysis of data from unbalanced experiments (Chapter 8).

The chapters in Part III stress the design and analysis of data from designed experiments with random factor effects. Added to the second edition is additional material on the analysis of data from incomplete block designs (Chapter 9) and split-plot designs (Chapter 11), new analyses for data from process improvement designs (Chapter 12), and analyses of data from gage R&R studies and data from some designs popularized by Genichi Taguchi (Chapter 13).

Throughout the analysis chapters in Parts II and III, confidence-interval and hypothesis-testing procedures are detailed for single-factor and multifactor experiments. Statistical models are used to describe responses from experiments, with careful attention to the specification of the terms of the various models and their relationship to the possible individual and joint effects of the experimental factors.

Part IV consists of Chapters 14 to 19 and is devoted to the analysis of experiments containing quantitative predictors and factors. Linear regression modeling using least-squares estimators of the model parameters is detailed, along with various diagnostic techniques for assessing the assumptions typically made with both regression and analysis-of-variance models. Analysis-of-covariance procedures are introduced, and the design and analysis needed for use in fitting response surfaces are presented. Identification of influential observations and the concepts of model assessment and variable selection are also discussed.

We are grateful to the Literary Executor of the late Sir Ronald A. Fisher, F. R. S., to Dr. Frank Yates, F. R. S., and to the Longman Group, Ltd., London, for permission to reprint part of Table XXIII from their book *Statistical Tables for Biological, Agricultural, and Medical Research* (6th edition, 1974).

To Carmen, Ann, Janis Sue

Preface

Statistical Design and Analysis of Experiments is intended to be a practitioner's guide to statistical methods for designing and analyzing experiments. The topics selected for inclusion in this book represent statistical techniques that we feel are most useful to experimenters and data analysts who must either collect, analyze, or interpret data. The material included in this book also was selected to be of value to managers, supervisors, and other administrators who must make decisions based in part on the analyses of data that may have been performed by others.

The intended audience for this book consists of two groups. The first group covers a broad spectrum of practicing engineers and scientists, including those in supervisory positions, who utilize or wish to utilize statistical approaches to solving problems in an experimental setting. This audience includes those who have little formal training in statistics but who are motivated by industrial or academic experiences in laboratory or process experimentation. These practicing engineers and scientists should find the contents of this book to be self-contained, with little need for reference to other sources for background information.

The second group for whom this book is intended is students in introductory statistics courses in colleges and universities. This book is appropriate for courses in which statistical experimental design and the analysis of data are the main topics. It is appropriate for upper-level undergraduate or introductory graduate-level courses, especially in disciplines for which the students have had or will have laboratory or similar data-collection experiences. The focus is on the use of statistical techniques, not on the theoretical underpinnings of those techniques. College algebra is the only prerequisite. A limited amount of supplemental material makes use of vector and matrix operations, notably the coverage of multiple linear regression. This material has been placed in appendices and is not essential for an understanding of the methods and applications contained in this book.

The emphasis in this book is on the strategy of experimentation, data analysis, and the interpretation of experimental results. The text features numerous examples using actual engineering and scientific studies. It presents statistics as an integral component of experimentation from the planning stage to the presentation of the conclusions.

This second edition constitutes a significant revision. A number of users of the first edition were surveyed and their feedback was incorporated in the revision. This resulted in deleting some material that wasn't intimately connected to the main thrust of the book, adding some new topics that supplemented existing topical coverage, and rearranging the presentation. For example, some introductory material was eliminated in order to introduce experimental design topics more quickly. A number of new examples were included in several of the chapters. New exercises were added to each of the chapters. In decisions regarding topics, we were guided by our collective experiences as statistical consultants and by our desire to produce a book that would be informative and readable. The topics selected for inclusion in both editions of this book can be implemented by practitioners and do not require a high level of training in statistics.

A key feature of the book, one that was cited as pedagogically beneficial by reviewers, is the depth and concentration of experimental design coverage, with equivalent but separate emphasis on the analysis of data from the various designs. In contrast to the previous edition, however, in the second edition chapters on the analysis of designed experiments have been placed immediately following the corresponding chapters on the respective designs. This was viewed as especially beneficial for classroom use. Instructors and readers can still emphasize design issues in a cohesive manner and can now have the analysis of the data resulting from the use of the respective designs reinforce the important features of the designs by having both the design and the analysis covered in close proximity to one another.

This second edition of *Statistical Design and Analysis of Experiments* is divided into four sections. Part I consists of Chapters 1 to 3 and presents a quick overview of many conceptual foundations of modern statistical practice. These three chapters introduce the reader to the basic issues surrounding the statistical analysis of data. The distinctions between populations or processes and samples, parameters and statistics, and mathematical and statistical modeling are discussed. In addition, elementary descriptive statistics and graphical displays are presented. Throughout the presentation, the informational content of simple graphical and numerical methods of viewing data is stressed.

Chapters 4 to 8 constitute Part II and Chapters 9–13 constitute Part III. These are the heart of the experimental design and analysis portions of the book. Unlike many other statistics books, this book intentionally separates discussions of the design of an experiment from those of the analysis of the resulting data from these experiments. Readers benefit from the reinforcement

In the first edition, Bea Schube was the John Wiley editor who helped initiate this project, and later Kate Roach was the editor who completed it. We are thankful to both of them as well as to the current Wiley editor, Steve Quigley, for their contributions. For this second edition, we also express our appreciation to Andrew Prince of John Wiley and Joan Wolk of Joan Wolk Editorial Services for their excellent work during the editorial and production process.

We are indebted to many individuals for contributing to this work. Several colleagues read earlier versions of the first edition and made many valuable suggestions on content and readability. We also are thankful to many users of the first editon of this book. Their comments and suggestions, as well as those received from several anonymous reviewers, have been very useful as we developed the second edition.

Contents

Preface vii

PART I FUNDAMENTAL STATISTICAL CONCEPTS 1

1. Statistics in Engineering and Science 3

1.1. The Role of Statistics in Experimentation, 5
1.2. Populations and Samples, 9
1.3. Parameters and Statistics, 19
1.4. Mathematical and Statistical Modeling, 24
Exercises, 28

2. Fundamentals of Statistical Inference 33

2.1. Traditional Summary Statistics, 33
2.2. Statistical Inference, 39
2.3. Probability Concepts, 42
2.4. Interval Estimation, 48
2.5. Statistical Tolerance Intervals, 50
2.6. Tests of Statistical Hypotheses, 52
2.7. Sample Size and Power, 56
Appendix: Probability Calculations, 59
Exercises, 64

3. Inferences on Means and Standard Deviations **69**

3.1. Inferences on a Population or Process Mean, 72

3.1.1. Confidence Intervals, 73

3.1.2. Hypothesis Tests, 76

3.1.3. Choice of a Confidence Interval or a Test, 78

3.1.4. Sample Size, 79

3.2. Inferences on a Population or Process Standard Deviation, 81

3.2.1. Confidence Intervals, 82

3.2.2. Hypothesis Tests, 84

3.3. Inferences on Two Populations or Processes Using Independent Pairs of Correlated Data Values, 86

3.4. Inferences on Two Populations or Processes Using Data from Independent Samples, 89

3.5. Comparing Standard Deviations from Several Populations, 96

Exercises, 99

PART II DESIGN AND ANALYSIS WITH FACTORIAL STRUCTURE **107**

4. Statistical Principles in Experimental Design **109**

4.1. Experimental-Design Terminology, 110

4.2. Common Design Problems, 115

4.2.1. Masking Factor Effects, 115

4.2.2. Uncontrolled Factors, 117

4.2.3. Erroneous Principles of Efficiency, 119

4.2.4. One-Factor-at-a-Time Testing, 121

4.3. Selecting a Statistical Design, 124

4.3.1. Consideration of Objectives, 125

4.3.2. Factor Effects, 126

4.3.3. Precision and Efficiency, 127

4.3.4. Randomization, 128

4.4. Designing for Quality Improvement, 128

Exercises, 132

5. Factorial Experiments in Completely Randomized Designs **140**

5.1. Factorial Experiments, 141
5.2. Interactions, 146
5.3. Calculation of Factor Effects, 152
5.4. Graphical Assessment of Factor Effects, 158
Appendix: Calculation of Effects for Factors with More than Two Levels, 160
Exercises, 163

6. Analysis of Completely Randomized Designs **170**

6.1. Balanced Multifactor Experiments, 171
6.1.1. Fixed Factor Effects, 171
6.1.2. Analysis-of-Variance Models, 173
6.1.3. Analysis-of-Variance Tables, 176
6.2. Parameter Estimation, 184
6.2.1. Estimation of the Error Standard Deviation, 184
6.2.2. Estimation of Effects Parameters, 186
6.2.3. Quantitative Factor Levels, 189
6.3. Statistical Tests, 194
6.3.1. Tests on Individual Parameters, 194
6.3.2. F-Tests for Factor Effects, 195
6.4. Multiple Comparisons, 196
6.4.1. Philosophy of Mean-Comparison Procedures, 196
6.4.2. General Comparisons of Means, 203
6.4.3. Comparisons Based on t-Statistics, 209
6.4.4. Tukey's Significant Difference Procedure, 212
6.5. Graphical Comparisons, 213
Exercises, 221

7. Fractional Factorial Experiments **228**

7.1. Confounding of Factor Effects, 229
7.2. Design Resolution, 237
7.3. Two-Level Fractional Factorial Experiments, 239

7.3.1. Half Fractions, 239
7.3.2. Quarter and Smaller Fractions, 243
7.4. Three-Level Fractional Factorial Experiments, 247
7.4.1. One-Third Fractions, 248
7.4.2. Orthogonal Array Tables, 252
7.5. Combined Two- and Three-Level Fractional Factorials, 254
7.6. Sequential Experimentation, 255
7.6.1. Screening Experiments, 256
7.6.2. Designing a Sequence of Experiments, 258
Appendix: Fractional Factorial Design Generators, 260
Exercises, 266

8. Analysis of Fractional Factorial Experiments 271
8.1. A General Approach for the Analysis of Data from Unbalanced Experiments, 272
8.2. Analysis of Marginal Means for Data from Unbalanced Designs, 276
8.3. Analysis of Data from Two-Level, Fractional Factorial Experiments, 278
8.4. Analysis of Data from Three-Level, Fractional Factorial Experiments, 287
8.5. Analysis of Fractional Factorial Experiments with Combinations of Factors Having Two and Three Levels, 290
8.6. Analysis of Screening Experiments, 293
Exercises, 299

PART III Design and Analysis with Random Effects 309

9. Experiments in Randomized Block Designs 311
9.1. Controlling Experimental Variability, 312
9.2. Complete Block Designs, 317
9.3. Incomplete Block Designs, 318
9.3.1. Two-Level Factorial Experiments, 318
9.3.2. Three-Level Factorial Experiments, 323
9.3.3. Balanced Incomplete Block Designs, 325

8. Upper-Tail Critical Values for the F-Max Test, 703
9. Orthogonal Polynomial Coefficients, 705
10. Critical Values for Outlier Test Using L_k and S_k, 709
11. Critical Values for Outlier Test Using E_k, 711
12. Coefficients Used in the Shapiro–Wilk Test for Normality, 713
13. Critical Values for the Shapiro–Wilk Test for Normality, 716
14. Percentage Points of the Studentized Range, 718

INDEX **723**

DATA INDEX **729**

9.4. Latin-Square and Crossover Designs, 328
9.4.1. Latin Square Designs, 328
9.4.2. Crossover Designs, 331
Appendix: Incomplete Block Design Generators, 332
Exercises, 342

10. Analysis of Designs with Random Factor Levels 347
10.1. Random Factor Effects, 348
10.2. Variance-Component Estimation, 350
10.3. Analysis of Data from Block Designs, 356
10.3.1. Complete Blocks, 356
10.3.2. Incomplete Blocks, 357
10.4. Latin-Square and Crossover Designs, 364
Appendix: Determining Expected Mean Squares, 366
Exercises, 369

11. Nested Designs 378
11.1. Crossed and Nested Factors, 379
11.2. Hierarchically Nested Designs, 381
11.3. Split-Plot Designs, 384
11.3.1. An Illustrative Example, 384
11.3.2. Classical Split-Plot Design Construction, 386
11.4. Restricted Randomization, 391
Exercises, 395

12. Special Designs for Process Improvement 400
12.1. Assessing Quality Performance, 401
12.1.1. Gage Repeatability and Reproducibility, 401
12.1.2. Process Capability, 404
12.2. Statistical Designs for Process Improvement, 406
12.2.1. Taguchi's Robust Product Design Approach, 406
12.2.2. An Integrated Approach, 410
Appendix: Selected Orthogonal Arrays, 414
Exercises, 418

13. Analysis of Nested Designs and Designs for Process Improvement **423**

13.1. Hierarchically Nested Designs, 423

13.2. Split-Plot Designs, 428

13.3. Gage Repeatability and Reproducibility Designs, 433

13.4. Signal-to-Noise Ratios, 436

Exercises, 440

PART IV Design and Analysis with Quantitative Predictors and Factors **459**

14. Linear Regression with One Predictor Variable **461**

14.1. Uses and Misuses of Regression, 462

14.2. A Strategy for a Comprehensive Regression Analysis, 470

14.3. Scatterplot Smoothing, 473

14.4. Least-Squares Estimation, 475

14.4.1. Intercept and Slope Estimates, 476

14.4.2. Interpreting Least-Squares Estimates, 478

14.4.3. No-Intercept Models, 480

14.4.4. Model Assumptions, 481

14.5. Inference, 481

14.5.1. Analysis-of-Variance Table, 481

14.5.2. Tests and Confidence Intervals, 484

14.5.3. No-Intercept Models, 485

14.5.4. Intervals for Responses, 485

Exercises, 486

15. Linear Regression with Several Predictor Variables **496**

15.1. Least Squares Estimation, 497

15.1.1. Coefficient Estimates, 497

15.1.2. Interpreting Least-Squares Estimates, 499

15.2. Inference, 503

15.2.1. Analysis of Variance, 503

15.2.2. Lack of Fit, 505

15.2.3. Tests on Parameters, 508

15.2.4. Confidence Intervals, 510

15.3. Interactions Among Quantitative Predictor Variables, 511
15.4. Polynomial Model Fits, 514
Appendix: Matrix Form of Least-Squares Estimators, 522
Exercises, 525

16. Linear Regression with Factors and Covariates as Predictors 535
16.1. Recoding Categorical Predictors and Factors, 536
16.1.1. Categorical Variables: Variables with Two Values, 536
16.1.2. Categorical Variables: Variables with More Than Two Values, 539
16.1.3. Interactions, 541
16.2. Analysis of Covariance for Completely Randomized Designs, 542
16.3. Analysis of Covariance for Randomized Complete Block Designs, 552
Appendix: Calculation of Adjusted Factor Averages, 556
Exercises, 558

17. Designs and Analyses for Fitting Response Surfaces 568
17.1. Uses of Response-Surface Methodology, 569
17.2. Locating an Appropriate Experimental Region, 575
17.3. Designs for Fitting Response Surfaces, 580
17.3.1. Central Composite Design, 582
17.3.2. Box–Behnken Design, 585
17.3.3. Some Additional Designs, 586
17.4. Fitting Response-Surface Models, 588
17.4.1. Optimization, 591
17.4.2. Optimization for Robust Parameter Product-Array Designs, 594
17.4.3. Dual Response Analysis for Quality Improvement Designs, 597
Appendix: Box–Behnken Design Plans; Locating Optimum Responses, 600
Exercises, 606

18. Model Assessment **614**

18.1. Outlier Detection, 614
18.1.1. Univariate Techniques, 615
18.1.2. Response-Variable Outliers, 619
18.1.3. Predictor-Variable Outliers, 626
18.2. Evaluating Model Assumptions, 630
18.2.1. Normally Distributed Errors, 630
18.2.2. Correct Variable Specification, 634
18.2.3. Nonstochastic Predictor Variables, 637
18.3. Model Respecification, 639
18.3.1. Nonlinear-Response Functions, 640
18.3.2. Power Reexpressions, 642
Appendix: Calculation of Leverage Values and Outlier Diagnostics, 647
Exercises, 651

19. Variable Selection Techniques **659**

19.1. Comparing Fitted Models, 660
19.2. All-Possible-Subset Comparisons, 662
19.3. Stepwise Selection Methods, 665
19.3.1. Forward Selection, 666
19.3.2. Backward Elimination, 668
19.3.3. Stepwise Iteration, 670
19.4. Collinear Effects, 672
Appendix: Cryogenic-Flowmeter Data, 674
Exercises, 678

APPENDIX: Statistical Tables **689**

1. Table of Random Numbers, 690
2. Standard Normal Cumulative Probabilities, 692
3. Student t Cumulative Probabilities, 693
4. Chi-Square Cumulative Probabilities, 694
5. F Cumulative Probabilities, 695
6. Factors for Determining One-sided Tolerance Limits, 701
7. Factors for Determining Two-sided Tolerance Limits, 702

PART I

Fundamental Statistical Concepts

CHAPTER 1

Statistics in Engineering and Science

In this chapter we introduce basic statistical concepts and terminology that are fundamental to the use of statistics in experimental work. These concepts include:

- *the role of statistics in engineering and scientific experimentation,*
- *the distinction between samples and populations,*
- *relating sample statistics to populations parameters, and*
- *characterizing deterministic and empirical models.*

The term *scientific* suggests a process of objective investigation that ensures that valid conclusions can be drawn from an experimental study. Scientific investigations are important not only in the academic laboratories of research universities but also in the engineering laboratories of industrial manufacturers. *Quality* and *productivity* are characteristic goals of industrial processes, which are expected to result in goods and services that are highly sought by consumers and that yield profits for the firms that supply them. Recognition is now being given to the necessary link between the scientific study of industrial processes and the quality of the goods produced. The stimulus for this recognition is the intense international competition among firms selling similar products to a limited consumer group.

The setting just described provides one motivation for examining the role of statistics in scientific and engineering investigations. It is no longer satisfactory just to monitor on-line industrial processes and to ensure that products are within desired specification limits. Competition demands that a better product be produced within the limits of economic realities. Better products are initiated in academic and industrial research laboratories, made feasible in

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

pilot studies and new-product research studies, and checked for adherence to design specifications throughout production. All of these activities require experimentation and the collection of data. The definition of the discipline of statistics in Exhibit 1.1 is used to distinguish the field of statistics from other academic disciplines and is oriented toward the experimental focus of this text. It clearly identifies statistics as a scientific discipline, which demands the same type of rigor and adherence to basic principles as physics or chemistry. The definition also implies that when problem solving involves the collection of data, the science of statistics should be an integral component of the process.

EXHIBIT 1.1

Statistics. Statistics is the science of problem-solving in the presence of variability.

Perhaps the key term in this definition is the last one. The problem-solving process involves a degree of uncertainty through the natural variation of results that occurs in virtually all experimental work.

When the term *statistics* is mentioned, many people think of games of chance as the primary application. In a similar vein, many consider statisticians to be "number librarians," merely counters of pertinent facts. Both of these views are far too narrow, given the diverse and extensive applications of statistical theory and methodology.

Outcomes of games of chance involve uncertainty, and one relies on probabilities, the primary criteria for statistical decisions, to make choices. Likewise, the determination of environmental standards for automobile emissions, the forces that act on pipes used in drilling oil wells, and the testing of commercial drugs all involve some degree of uncertainty. Uncertainty arises because the level of emissions for an individual automobile, the forces exerted on a pipe in one well, and individual patient reactions to a drug vary with each observation, even if the observations are taken under "controlled" conditions. These types of applications are only a few of many that could be mentioned. Many others are discussed in subsequent chapters of this book.

Figure 1.1 symbolizes the fact that statistics should play a role in every facet of data collection and analysis, from initial problem formulation to the drawing of final conclusions. This figure distinguishes two types of studies: experimental and observational. In experimental studies the variables of interest often can be controlled and fixed at predetermined values for each test run in the experiment. In observational studies many of the variables of interest cannot be controlled, but they can be recorded and analyzed. In this book we emphasize experimental studies, although many of the analytic procedures discussed can be applied to observational studies.

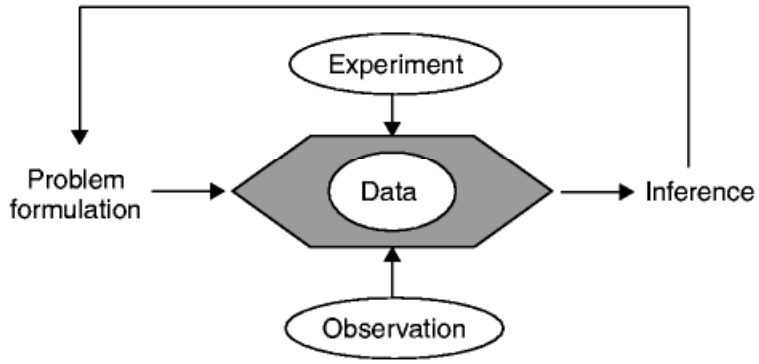


Figure 1.1 Critical stages of statistical input in scientific investigations.

Data are at the center of experimental and observational studies. As will be stressed in Section 1.1, all data are subject to a variety of sources that induce variation in measurements. This variation can occur because of fixed differences among machines, random differences due to changes in ambient conditions, measurement error in instrument readings, or effects due to many other known or unknown influences.

Statistical experimental design will be shown to be effective in eliminating known sources of bias, guarding against unknown sources of bias, ensuring that the experiment provides precise information about the responses of interest, and guaranteeing that excessive experimental resources are not needlessly wasted through the use of an uneconomical design. Likewise, whether one simply wishes to describe the results of an experiment or one wishes to draw inferential conclusions about a process, statistical data-analysis techniques aid in clearly and concisely summarizing salient features of experimental data.

The next section of this chapter discusses the role of statistics in the experimental process, and illustrates how a carefully designed experiment and straightforward statistical graphics can clearly identify major sources of variation in a chemical process. The last three sections of this chapter introduce several concepts that are fundamental to an understanding of statistical inference.

1.1 THE ROLE OF STATISTICS IN EXPERIMENTATION

Statistics is a scientific discipline devoted to the drawing of valid inferences from experimental or observational data. The study of variation, including the construction of experimental designs and the development of models which describe variation, characterizes research activities in the field of statistics. A basic principle that is the cornerstone of the material covered in this book is the following:

All measurements are subject to variation.

The use of the term *measurement* in this statement is not intended to exclude qualitative responses of interest in an experiment, but the main focus of this text is on designs and analyses that are appropriate for quantitative measurements.

In most industrial processes there are numerous sources of possible variation. Frequently studies are conducted to investigate the causes of excessive variation. These studies could focus on a single source or simultaneously examine several sources. Consider, for example, a chemical analysis that involves different specimens of raw materials and that is performed by several operators. Variation could occur because the operators systematically differ in their method of analysis. Variation also could occur because one or more of the operators do not consistently adhere to the analytic procedures, thereby introducing uncontrolled variability to the measurement process. In addition, the specimens sent for analysis could differ on factors other than the ones under examination.

To investigate sources of variability for a chemical analysis similar to the one just described, an experiment was statistically designed and analyzed to ensure that relevant sources of variation could be identified and measured. A test specimen was treated in a combustion-type furnace, and a chemical analysis was performed on it. In the experiment three operators each analyzed two specimens, made three combustion runs on each specimen, and titrated each run in duplicate. The results of the experiment are displayed in Table 1.1 and graphed in Figure 1.2.

Figure 1.2 is an example of a *scatterplot*, a two-dimensional graph of individual data values for pairs of quantitative variables. In Figure 1.2, the abscissa (horizontal axis) is simply the specimen/combustion run index and the ordinate (vertical axis) is the chemical analysis result. Scatterplots can be made for any pair of variables so long as both are quantitative. A scatterplot is constructed by plotting the (x_i, y_i) pairs as indicated in Exhibit 1.2.

EXHIBIT 1.2 SCATTERPLOTS

1. Construct horizontal and vertical axes that cover the ranges of the two variables.
2. Plot (x_i, y_i) points for each observation in the data set.

Figure 1.2 highlights a major problem with the chemical analysis procedure. There are definite differences in the analytic results of the three operators. Operator 1 exhibits very consistent results for each of the two specimens and each of the three combustion runs. Operator 2 produces analytic results that

TABLE 1.1 Results of an Experiment to Identify Sources of Variation in Chemical Analyses[a]

Operator	Specimen	Combustion Run	Chemical Analysis 1	Chemical Analysis 2
1	1	1	156	154
		2	151	154
		3	154	160
	2	4	148	150
		5	154	157
		6	147	149
2	3	7	125	125
		8	94	95
		9	98	102
	4	10	118	124
		11	112	117
		12	98	110
3	5	13	184	184
		14	172	186
		15	181	191
	6	16	172	176
		17	181	184
		18	175	177

[a] Adapted from Snee, R. D. (1983). "Graphical Analysis of Process Variation Studies," *Journal of Quality Technology*, **15**, 76–88. Copyright, American Society for Quality Control, Inc., Milwaukee, WI. Reprinted by permission.

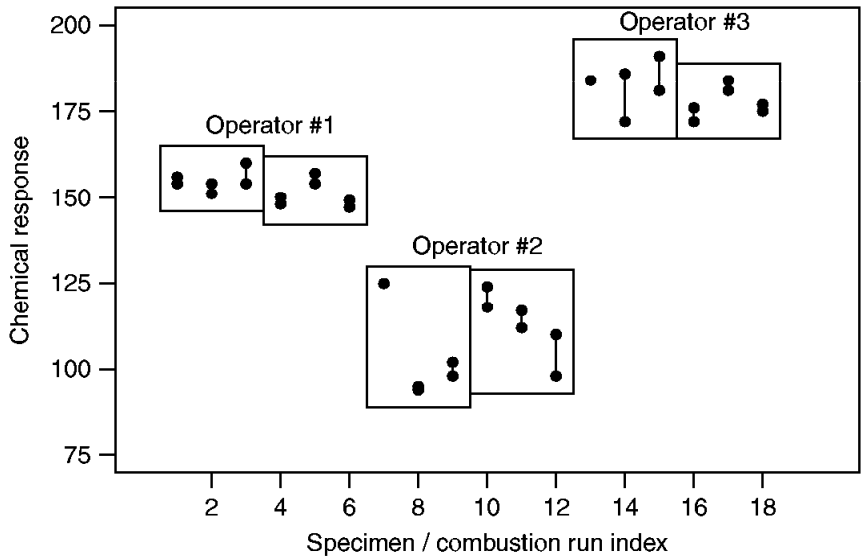


Figure 1.2 Results of a study of variation in a chemical analysis. (Combustion runs are boxed; duplicate analyses are connected by vertical lines.)

are lower on the average than those of the other two operators. Operator 3 shows good consistency between the two specimens, but the repeat analyses of two of the combustion runs on specimen 5 appear to have substantially larger variation than for most of the other repeat analyses in the data set. Operator 2 likewise shows good average consistency for the two specimens, but large variation both for the triplicate combustion runs for each specimen and for at least one of the repeat analyses for the fourth specimen.

Thus, the experimental results indicate that the primary sources of variation in this chemical analysis are the systematic differences (biases) among operators and, in some instances, the (random) inconsistency of the chemical analyses performed by a single operator. In reaching these conclusions statistics played a role in both the design of the experiment and the formal analysis of the results, the foregoing graphical display being one component of the analysis. The quality of this data-collection effort enables straightforward, unambiguous conclusions to be drawn. Such clear-cut inferences are often lacking when data are not collected according to a detailed statistical experimental design.

This example illustrates three general features of the statistical design and analysis of experiments. First, statistical considerations should be included in the project design phase of any experiment. At this stage of a project one should consider the nature of the data to be collected, including what measurements are to be taken, what is known about the likely variation to be encountered, and what factors might influence the variation in the measurements.

Second, a statistical design should be selected that controls, insofar as possible, variation from known sources. The design should allow the estimation of the magnitude of uncontrollable variation and the modeling of relationships between the measurements of interest and factors (sources) believed to influence these measurements.

Uncontrollable variation can arise from many sources. Two general sources of importance to the statistical design of experiments are experimental error and measurement error. Experimental error is introduced whenever test conditions are changed. For example, machine settings are not always exact enough to be fixed at precisely the same value or location when two different test runs call for identical settings. Batches of supposedly identical chemical solutions do not always have exactly the same chemical composition. Measurement errors arise from the inability to obtain exactly the same measurement on two successive test runs when all experimental conditions are unchanged.

Third, a statistical analysis of the experimental results should allow inferences to be drawn on the relationships between the design factors and the measurements. This analysis should be based on both the statistical design

TABLE 1.2 Role of Statistics in Experimentation

Project Planning Phase

- What is to be measured?
- How large is the likely variation?
- What are the influential factors?

Experimental Design Phase

- Control known sources of variation
- Allow estimation of the size of the uncontrolled variation
- Permit an investigation of suitable models

Statistical Analysis Phase

- Make inferences on design factors
- Guide subsequent designs
- Suggest more appropriate models

and the model used to relate the measurements to the sources of variation. If additional experimentation is necessary or desirable, the analysis should guide the experimenter to an appropriate design and, if needed, a more appropriate model of the measurement process.

Thus, the role of statistics in engineering and scientific experimentation can be described using three basic categories: project planning, experimental design, and data analysis. These three basic steps in the statistical design and analysis of experimental results are depicted in Table 1.2.

1.2 POPULATIONS AND SAMPLES

Experimental data, in the form of a representative sample of observations, enable us to draw inferences about a phenomenon, population, or process of interest. These inferences are obtained by using sample statistics to draw conclusions about postulated models of the underlying data-generating mechanism.

All possible items or units that determine an outcome of a well-defined experiment are collectively called a "population" (see Exhibit 1.3). An item or a unit could be a measurement, or it could be material on which a measurement is taken. For example, in a study of geopressure as an alternative source of electric power, a population of interest might be all geographical locations for which characteristics such as wellhead fluid temperature, pressure, or gas content could be measured. Other examples of populations are:

- all 30-ohm resistors produced by a particular manufacturer under specified manufacturing conditions during a fixed time period;
- all possible fuel-consumption values obtainable with a four-cylinder, 1.7-liter engine using a 10%-methanol, 90%-gasoline fuel blend, tested under controlled conditions on a dynamometer stand;
- all measurements on the fracture strength of one-inch-thick underwater welds on a steel alloy base plate that is located 200 feet deep in a specified salt-water environment; or
- all 1000-lb containers of pelletized, low-density polyethylene produced by a single manufacturing plant under normal operating conditions.

EXHIBIT 1.3

Population. A statistical population consists of all possible items or units possessing one or more common characteristics under specified experimental or observational conditions.

These examples suggest that a population of observations may exist only conceptually, as with the population of fracture-strength measurements. Populations also may represent processes for which the items of interest are not fixed or static; rather, new items are added as the process continues, as in the manufacture of polyethylene.

Populations, as represented by a fixed collection of units or items, are not always germane to an experimental setting. For example, there are no fixed populations in many studies involving chemical mixtures or solutions. Likewise, ongoing production processes do not usually represent fixed populations. The study of physical phenomena such as aging, the effects of drugs, or aircraft engine noise cannot be put in the context of a fixed population of observations. In situations such as these it is a physical process rather than a population that is of interest (see Exhibit 1.4).

EXHIBIT 1.4

Process. A process is a repeatable series of actions that results in an observable characteristic or measurement.

The concepts and analyses discussed in this book relative to samples from populations generally are applicable to processes. For example, one samples both populations and processes in order to draw inferences on models appropriate for each. While one models a fixed population in the former case, one

models a "state of nature" in the latter. A simple random sample may be used to provide observations from which to estimate the population model. A suitably conducted experiment may be used to provide observations from which to estimate the process model. In both situations it is the representative collection of observations and the assumptions made about the data that are important to the modeling procedures.

Because of the direct analogies between procedures for populations and for processes, the focus of the discussions in this book could be on either. We shall ordinarily develop concepts and experimental strategies with reference to only one of the two, with the understanding that they should readily be transferrable to the other. In the remainder of this section, we concentrate attention on developing the relationships between samples and populations.

When defining a relevant population (or process) of interest, one must define the exact experimental conditions under which the observations are to be collected. Depending on the experimental conditions, many different populations of observed values could be defined. Thus, while populations may be real or conceptual, they must be explicitly defined with respect to all known sources of variation in order to draw valid statistical inferences.

The items or units that make up a population are usually defined to be the smallest subdivisions of the population for which measurements or observations can take on different values. For the populations defined above, for example, the following definitions represent units of interest. An individual resistor is the natural unit for studying the actual (as opposed to specified) resistance of a brand of resistors. A measurement of fuel consumption from a single test sequence of accelerations and decelerations is the unit for which data are accumulated in a fuel economy study. Individual welds are the appropriate units for investigating fracture strength. A single container of pellets is the unit of interest in the manufacture of polyethylene.

Measurements on a population of units can exhibit many different statistical properties, depending on the characteristic of interest. Thus, it is important to define the fundamental qualities or quantities of interest in an experiment. We term these qualities or quantities *variables* (see Exhibit 1.5).

EXHIBIT 1.5

Variable. A property or characteristic on which information is obtained in an experiment.

An *observation*, as indicated in Exhibit 1.6, refers to the collection of information in an experiment, and an *observed value* refers to an actual measurement or attribute that is the result of an individual observation. We often

use "observation" in both senses; however, the context of its use should make it clear which meaning is implied.

EXHIBIT 1.6

Observation. The collection of information in an experiment, *or* actual values obtained on variables in an experiment.

A delineation of variables into two categories, response variables (see Exhibit 1.7) and factors (see Exhibit 1.8), is an important consideration in the modeling of data. In some instances response variables are defined according to some probability model which is only a function of certain (usually unknown) constants. In other instances the model contains one or more factors in addition to (unknown) constants.

EXHIBIT 1.7

Response Variable. Any outcome or result of an experiment.

EXHIBIT 1.8

Factors. Controllable experimental variables that can influence the observed values of response variables.

The response variable in a resistor study is the actual resistance measured on an individual resistor. In a study of fuel economy one might choose to model the amount of fuel consumed (response variable) as some function of vehicle type, fuel, driver, ambient temperature, and humidity (factors). In the underwater weld study the response variable is the fracture strength. In the manufacture of polyethylene the response variable of interest might be the actual weight of a container of pellets.

Most of the variables just mentioned are quantitative variables, because each observed value can be expressed numerically. There also exist many qualitative or nonnumerical variables that could be used as factors. Among those variables listed above, ones that could be used as qualitative factors include vehicle type, fuel, and driver.

Populations often are too large to be adequately studied in a specified time period or within designated budgetary constraints. This is particularly true when the populations are conceptual, as in most scientific and engineering experiments, when they represent every possible observation that could be

obtained from a manufacturing process under specified conditions, or when the collection of data requires the destruction of the item. If it is not feasible to collect information on every item in a population, inferences on the population can be made by studying a representative subset of the data, a *sample* (see Exhibit 1.9). Figure 1.3 illustrates one of the primary goals of scientific experimentation and observation: induction from a sample to a population or a process.

EXHIBIT 1.9

Sample. A sample is a group of observations taken from a population or a process.

There are many ways to collect samples in experimental work. A *convenience sample* is one that is chosen simply by taking observations that are easily or inexpensively obtained. The key characteristic of a convenience sample is that all other considerations are secondary to the economic or rapid collection of data. For example, small-scale laboratory studies often are necessary prior to the implementation of a manufacturing process. While this type of pilot study is an important strategy in feasibility studies, the results are generally inadequate for inferring characteristics of the full-scale manufacturing process. Sources of variation on the production line may be entirely different from those in the tightly controlled environment of a laboratory.

Similarly, simply entering a warehouse and conveniently selecting a number of units for inspection may result in a sample of units which exhibits less variation than the population of units in the warehouse. From a statistical viewpoint, convenience samples are of dubious value because the population that they represent may have substantially different characteristics than the population of interest.

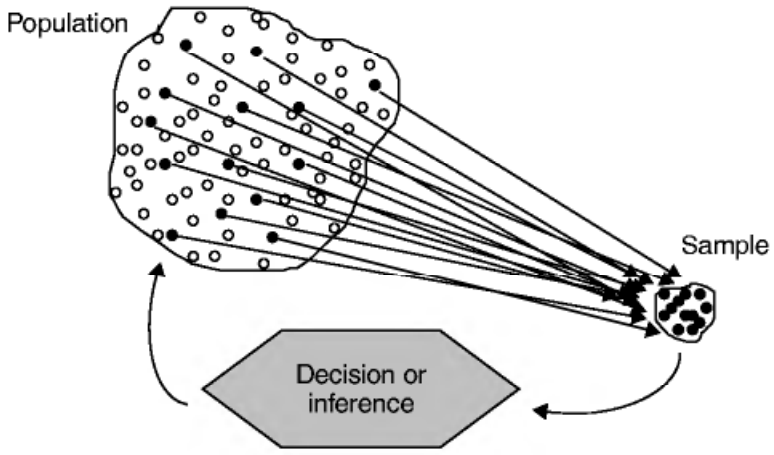


Figure 1.3 Representative samples permit inductive inferences on populations.

Another sampling technique that is frequently used in scientific studies is termed *judgmental sampling*. Here one's experience and professional judgment are used to select representative observations from a population of interest. In the context of a fuel-economy study, an example would be the selection of a particular engine, one fuel blend, and specific laboratory conditions for conducting the study. If the conditions selected for study are not truly representative of typical engine, fuel, and operating conditions, it is difficult to define the relevant population to which the observed fuel-consumption values pertain. A current example of this problem is the E.P.A. fuel-economy ratings posted on automobile stickers. These figures are comparable only under the laboratory conditions under which the estimates are made, not on any typical vehicle, fuel, or operating conditions.

Convenience and judgmental samples are important in exploratory research. The difficulty with these types of samples is that important sources of variation may be held constant or varied over a narrower range than would be the case for the natural occurrence of experimental units from the population of interest. In addition, these sampling schemes may mask the true effects of influential factors. This is an especially acute problem if two or more factors jointly influence a response. Holding one or more of these joint factors constant through convenience or judgmental sampling could lead to erroneous inferences about the effects of the factors on the response.

One of the most important sampling methodologies in experimental work is the *simple random sample*, defined in Exhibit 1.10. In addition to its use in the sampling of observations from a population, simple random sampling has application in the conduct of scientific and engineering experiments. Among the more prominent uses of simple random sampling in experimental work are the selection of experimental units and the randomization of test runs.

EXHIBIT 1.10

Simple Random Sample. In an experimental setting, a simple random sample of size n is obtained when items are selected from a fixed population or a process in such a manner that every group of items of size n has an equal chance of being selected as the sample.

If one wishes to sample 100 resistors from a warehouse, simple random sampling requires that every possible combination of 100 resistors present in the warehouse have an equal chance of being included in the selected sample. Although the requirements of simple random sampling are more stringent than most other sampling techniques, unintentional biases are avoided.

Simple random samples can be obtained in many ways. For example, in the selection of experimental units to be included in an experiment, a common approach is to enumerate or label each item from 1 to N and then use a

table of random numbers to select n of the N units. If a test program is to consist of n test runs, the test runs are sequentially numbered from 1 to n and a random-number table is used to select the run order. Equivalently, one can use random-number generators, which are available on computers. Use of such tables or computer algorithms removes personal bias from the selection of units or the test run order.

Simple random samples can be taken *with or without replacement*. Sampling with replacement allows an experimental unit to be selected more than once. One simply obtains n numbers from a random-number table without regard to whether any of the selected numbers occur more than once in the sample. Sampling without replacement prohibits any number from being selected more than once. If a number is sampled more than once, it is discarded after the first selection. In this way n unique numbers are selected. The sequencing of test runs is always performed by sampling without replacement. Ordinarily the selection of experimental units is also performed by sampling without replacement.

Inspection sampling of items from lots in a warehouse is an example for which a complete enumeration of experimental units is possible, at least for those units that are present when the sample is collected. When a population of items is conceptual or an operating production process is being studied, this approach is not feasible. Moreover, while one could conceivably sample at random from a warehouse full of units, the expense suffered through the loss of integrity of bulk lots of product when a single item is selected for inclusion in a sample necessitates alternative sampling schemes.

There are many other types of random sampling schemes besides simple random sampling. *Systematic* random samples are obtained by sampling every kth (e.g., every 5th, 10th, or 100th) unit in the population. *Stratified random samples* are based on subdividing a heterogeneous population into groups, or *strata*, of similar units and selecting simple random samples from each of the strata. *Cluster sampling* is based on subdividing the population into groups, or clusters, of units in such a way that it is convenient to randomly sample the clusters and then either randomly sample or completely enumerate all the observations in each of the sampled clusters. More details on these and other alternatives to simple random sampling are given in the recommended readings at the end of this chapter.

Regardless of which sampling technique is used, the key idea is that the sample should be representative of the population under study. In experimental settings for which the sampling of populations or processes is not germane, the requirement that the data be representative of the phenomenon or the "state of nature" being studied is still pertinent and necessary. Statistics, as a science, seeks to make inferences about a population, process, or phenomenon based on the information contained in a representative sample or collection of observations.

TABLE 1.3 Employee Identification Numbers

1	A11401	41	B09087	81	G07704	121	B04256
2	P04181	42	B00073	82	K20760	122	K05170
3	N00004	43	J08742	83	W00124	123	R07790
4	C03253	44	W13972	84	T00141	124	G15084
5	D07159	45	S00856	85	M25374	125	C16254
6	M00079	46	A00166	86	K03911	126	R20675
7	S15552	47	S01187	87	W01718	127	G06144
8	G01039	48	D00022	88	T04877	128	T12150
9	P00202	49	Z01194	89	M22262	129	R07904
10	R22110	50	M32893	90	C00011	130	M24214
11	D00652	51	K00018	91	W23233	131	D00716
12	M06815	52	H16034	92	K10061	132	M27410
13	C09071	53	F08794	93	K11411	133	J07272
14	S01014	54	S71024	94	B05848	134	L02455
15	D05484	55	G00301	95	L06270	135	D06610
16	D00118	56	B00103	96	K08063	136	M31452
17	M28883	57	B29884	97	P07211	137	L25264
18	G12276	58	G12566	98	F28794	138	M10405
19	M06891	59	P03956	99	L00885	139	D00393
20	B26124	60	B00188	100	M26882	140	B52223
21	D17682	61	J21112	101	M49824	141	M16934
22	B42024	62	J08208	102	R05857	142	M27362
23	K06221	63	S11108	103	L30913	143	B38384
24	C35104	64	M65014	104	B46004	144	H08825
25	M00709	65	M07436	105	R03090	145	S14573
26	P00407	66	H06098	106	H09185	146	B23651
27	P14580	67	S18751	107	J18200	147	S27272
28	P13804	68	W00004	108	W14854	148	G12636
29	P23144	69	M11028	109	S01078	149	R04191
30	D00452	70	L00213	110	G09221	150	D13524
31	B06180	71	J06070	111	M17174	151	G00154
32	B69674	72	B14514	112	L04792	152	B19544
33	H11900	73	H04177	113	S23434	153	V01449
34	M78064	74	B26003	114	T02877	154	F09564
35	L04687	75	B26193	115	K06944	155	L09934
36	F06364	76	H28534	116	E14054	156	A10690
37	G24544	77	B04303	117	F00281	157	N02634
38	T20132	78	S07092	118	H07233	158	W17430
39	D05014	79	H11759	119	K06204	159	R02109
40	R00259	80	L00252	120	K06423	160	C18514

To illustrate the procedures involved in randomly sampling a population, consider the information contained in Table 1.3. The table enumerates a portion of the world-wide sales force of a manufacturer of skin products. The employees are identified in the table by the order of their listing (1–160) and by their employee identification numbers. Such a tabulation might be obtained from a computer printout of personnel records. For the purposes of the study to be described, these 160 individuals form a population that satisfies several criteria set forth in the experimental protocol.

Suppose the purpose of a study involving these employees is to investigate the short-term effects of certain skin products on measurements of skin elasticity. Initial skin measurements are available for the entire population of employees (see Table 1.4). However, the experimental protocol requires that skin measurements be made on a periodic basis, necessitating the transportation of each person in the study to a central measuring laboratory. Because of the expense involved, the researchers would like to limit the participants included in the study to a simple random sample of 25 of the employees listed in Table 1.3.

Because the population of interest has been completely enumerated, one can use a random-number table (e.g., Table A1 of the Appendix) or a computer-generated sequence of random numbers to select 25 numbers between 1 and 160. One such random number sequence is

57, 77, 8, 83, 92, 18, 63, 121, 19, 115, 139, 96, 133,
131, 122, 17, 79, 2, 68, 59, 157, 138, 26, 70, 9.

Corresponding to this sequence of random numbers is the sequence of employee identification numbers that determines which 25 of the 160 employees are to be included in the sample:

B29884, B04303, G01039, W00124, K10061, G12276, S11108,
B04256, M06891, K06944, D00393, K08063, J07272, D00716,
K05170, M28883, H11759, P04181, W00004, P03956, N02634,
M10405, P00407, L00213, P00202.

With periodic measurements taken on only this random sample of employees the researchers wish to draw conclusions about skin elasticity for the population of employees listed in Table 1.3. This statement suggests that a distinction must be made between measured characteristics taken on a population and those taken on a sample. This distinction is made explicit in the next section.

TABLE 1.4 Skin Elasticity Measurements

1	31.9	41	36.0	81	36.3	121	33.0
2	33.1	42	28.6	82	36.3	122	37.4
3	33.1	43	38.0	83	41.5	123	33.8
4	38.5	44	39.1	84	33.0	124	35.3
5	39.9	45	39.4	85	36.3	125	37.5
6	36.5	46	30.6	86	36.3	126	31.6
7	34.8	47	34.1	87	30.9	127	33.1
8	38.9	48	40.8	88	32.3	128	38.2
9	40.3	49	35.1	89	39.2	129	31.4
10	33.6	50	34.1	90	35.2	130	35.9
11	36.4	51	36.3	91	35.1	131	37.6
12	34.4	52	35.1	92	33.9	132	35.5
13	35.7	53	35.0	93	42.0	133	34.2
14	33.9	54	39.0	94	35.1	134	34.0
15	36.6	55	34.0	95	34.5	135	31.3
16	36.0	56	35.3	96	35.0	136	32.6
17	30.8	57	36.0	97	35.1	137	34.9
18	31.1	58	34.7	98	35.7	138	35.3
19	37.6	59	39.8	99	36.4	139	35.1
20	35.7	60	35.8	100	39.6	140	35.7
21	29.6	61	35.7	101	35.2	141	32.3
22	37.3	62	39.8	102	37.2	142	38.1
23	31.4	63	36.4	103	33.3	143	36.8
24	31.6	64	36.1	104	33.7	144	38.7
25	34.6	65	37.7	105	37.8	145	40.0
26	34.6	66	32.3	106	34.4	146	35.4
27	33.7	67	35.6	107	36.9	147	34.0
28	30.9	68	38.2	108	31.8	148	34.3
29	34.6	69	39.0	109	35.3	149	32.8
30	37.0	70	34.3	110	38.1	150	30.7
31	35.3	71	40.6	111	34.1	151	34.4
32	36.3	72	37.4	112	35.8	152	34.3
33	31.8	73	37.3	113	33.3	153	35.8
34	38.2	74	36.9	114	33.8	154	37.5
35	34.6	75	29.0	115	36.4	155	34.4
36	36.0	76	39.0	116	36.9	156	35.8
37	40.8	77	33.7	117	35.3	157	31.9
38	39.2	78	32.9	118	37.0	158	36.9
39	33.4	79	33.8	119	33.5	159	34.4
40	34.0	80	36.2	120	40.3	160	30.1

1.3 PARAMETERS AND STATISTICS

Summarization of data can occur in both populations and samples. Parameters, as defined in Exhibit 1.11, are constant population values that summarize the entire collection of observations. Parameters can also be viewed in the context of a stable process or a controlled experiment. In all such settings a parameter is a fixed quantity that represents a characteristic of interest. Some examples are:

- the mean fill level for twelve-ounce cans of a soft drink bottled at one plant,
- the minimum compressive strength of eight-foot-long, residential-grade, oak ceiling supports, and
- the maximum wear on one-half-inch stainless-steel ball bearings subjected to a prescribed wear-testing technique.

EXHIBIT 1.11

Parameters and Statistics. A parameter is a numerical characteristic of a population or a process. A statistic is a numerical characteristic that is computed from a sample of observations.

Parameters often are denoted by Greek letters, such as μ for the mean and σ for the standard deviation (a measure of the variability of the observations in a population), to reinforce the notion that they are (generally unknown) constants. Often population parameters are used to define specification limits or tolerances for a manufactured product. Alternatively they may be used to denote hypothetical values for characteristics of measurements that are to be subjected to scientific or engineering investigations.

In many scientific and engineering contexts the term *parameter* is used as a synonym for variable (as defined in the previous section). The term *parameter* should be reserved for a constant or fixed numerical characteristic of a population and not used for a measured or observed property of interest in an experiment. To emphasize this distinction we will henceforth use Greek letters to represent population parameters and Latin letters to denote variables. Sample statistics, in particular estimates of population parameters, also will generally be denoted by Latin letters.

The term *distribution* (see Exhibit 1.12) is used throughout this text to refer to the possible values of a variable along with some measure of how frequently they occur. In a sample or a population the frequency could be measured by counts or percentages. Often when dealing with populations or processes the frequency is measured in terms of a probability model specifying the likelihood of occurrence of the values.

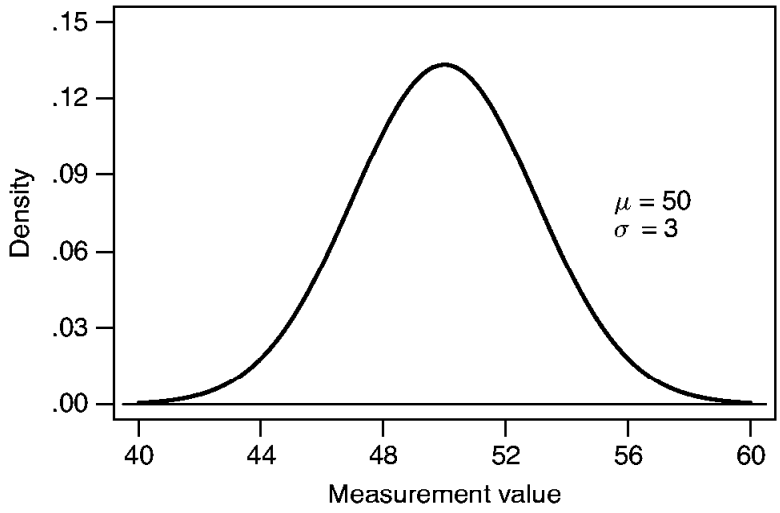


Figure 1.4 Normal distribution of measurement values.

The curve in Figure 1.4 often is used as a probability model, the *normal* distribution, to characterize populations and processes for many types of measurements. The *density* or height of the curve above the axis of measurement values, represents the likelihood of obtaining a value. Probabilities for any range of measurement values can be calculated from the probability model once the model parameters are specified. For this distribution, only the mean and the standard deviation are needed to completely specify the probability model.

EXHIBIT 1.12

Distribution. A tabular, graphical, or theoretical description of the values of a variable using some measure of how frequently they occur in a population, a process, or a sample.

The peak of the curve in Figure 1.4 is located above the measurement value 50, which is the mean μ of the distribution of data values. Because the probability density is highest around the mean, measurement values around the mean are more likely than measurement values greatly distant from it. The standard deviation σ of the distribution in Figure 1.4 is 3. For normal distributions (Section 2.3), approximately 68% of the measurement values lie between $\mu \pm \sigma$ (47 to 53), approximately 95% between $\mu \pm 2\sigma$ (44 to 56), and approximately 99% between $\mu \pm 3\sigma$ (41 to 59). The mean and the standard deviation are very important parameters for the distribution of measurement values for normal distributions such as that of Figure 1.4.

Statistics are sample values that generally are used to estimate population parameters. For example, the average of a sample of observations can be used

to estimate the mean of the population from which the sample was drawn. Similarly, the standard deviation of the sample can be used to estimate the population standard deviation. As we shall see in subsequent chapters, there are often several sample statistics that can be used to estimate a population parameter.

While parameters are fixed constants representing an entire population of data values, statistics are "random" variables and their numerical values depend on which particular observations from the population are included in the sample. One interesting feature about a statistic is that it has its own probability, or *sampling*, distribution: the sample statistic can take on a number of values according to a probability model, which is determined by the probability model for the original population and by the sampling procedure (see Exhibit 1.13). Hence, a statistic has its own probability model as well as its own parameter values, which may be quite different from those of the original population.

EXHIBIT 1.13

Sampling Distribution. A sampling distribution is a theoretical model that describes the probability of obtaining the possible values of a sample statistic.

Histograms are among the most common displays for illustrating the distribution of a set of data. They are especially useful when large numbers of data must be processed. *Histograms* (see Exhibit 1.14) are constructed by dividing the range of the data into several intervals (usually of equal length), counting the number of observations in each interval, and constructing a bar chart of the counts. A by-product of the construction of the histogram is the *frequency distribution*, which is a table of the counts or frequencies for each interval.

EXHIBIT 1.14 FREQUENCY DISTRIBUTIONS AND HISTOGRAMS

1. Construct intervals, ordinarily equally spaced, which cover the range of the data values.
2. Count the number of observations in each of the intervals. If desirable, form proportions or percentages of counts in each interval.
3. Clearly label all columns in tables and both axes on histograms, including any units of measurement, and indicate the sample or population size.
4. For histograms, plot bars whose
   (a) widths correspond to the measurement intervals, and
   (b) heights are (proportional to) the counts for each interval (e.g., heights can be counts, proportions, or percentages).

Both histograms and the tables of counts that accompany them are sometimes referred to as frequency distributions, because they show how often the data occur in various intervals of the measured variable. The intervals for which counts are made are generally chosen to be equal in width, so that the size (area) of the bar or count is proportional to the number of observations contained in the interval. Selection of the interval width is usually made by simply dividing the range of the data by the number of intervals desired in the histogram or table. Depending on the number of observations, between 8 and 20 intervals are generally selected—the greater the number of observations, the greater the number of intervals.

When the sample size is large, it can be advantageous to construct *relative-frequency* histograms. In these histograms and frequency distributions either the proportions (counts/sample size) or the percentages (proportions $\times$ 100%) of observations in each interval are calculated and graphed, rather than the frequencies themselves. Use of relative frequencies (or percentages) in histograms ensures that the total area under the bars is equal to one (or 100%). This facilitates the comparison of the resultant distribution with that of a theoretical probability distribution, where the total area of the distribution also equals one.

A frequency distribution and histogram for the skin elasticity measurements in Table 1.4 are shown in Table 1.5 and Figure 1.5. The histogram in Figure 1.5 is an example of a relative-frequency histogram. The heights of the bars suggest a shape similar to the form of the normal curve in Figure 1.4. On the basis of these data one might postulate a normal probability model for the skin measurements.

Figure 1.6 shows a normal probability model that has the same mean ($\mu = 35.4$) and standard deviation ($\sigma = 2.65$) as the population of values

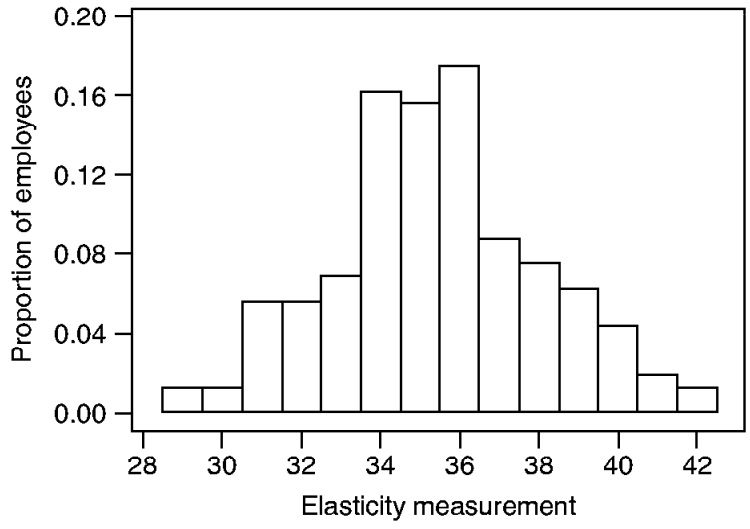


Figure 1.5 Distribution of elasticity measurements ($n = 160$).

TABLE 1.5 Frequency Distribution for Skin Elasticity Data Set

Skin Elasticity*	Interval Midpoint	Frequency	Proportion
28.5–29.5	29	2	0.01
29.5–30.5	30	2	0.01
30.5–31.5	31	9	0.06
31.5–32.5	32	9	0.06
32.5–33.5	33	11	0.07
33.5–34.5	34	26	0.16
34.5–35.5	35	25	0.16
35.5–36.5	36	28	0.18
36.5–37.5	37	14	0.09
37.5–38.5	38	12	0.08
38.5–39.5	39	10	0.06
39.5–40.5	40	7	0.04
40.5–41.5	41	3	0.02
41.5–42.5	42	2	0.01
		160	1.00

*Intervals include lower limits, exclude upper ones.

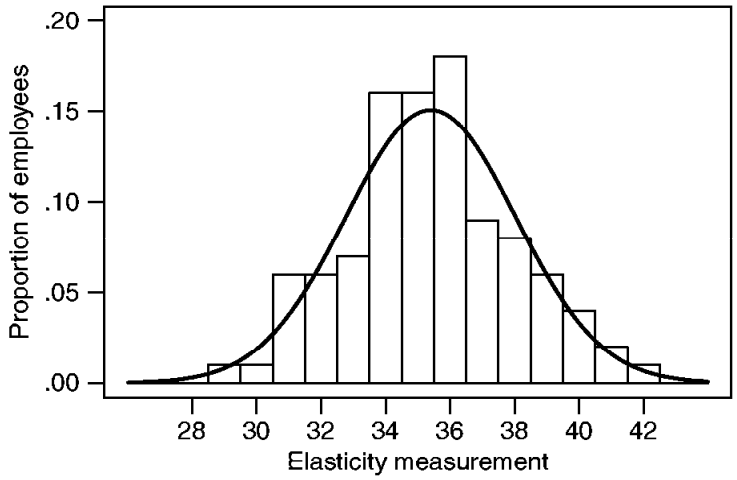


Figure 1.6 Normal approximation to elasticity distribution.

in Table 1.4. Observe that the curve for the theoretical normal model provides a good approximation to the actual distribution of the population of measurements, represented by the vertical bars.

One of the features of a normal model is that averages from simple random samples of size n also follow a normal probability model with the same

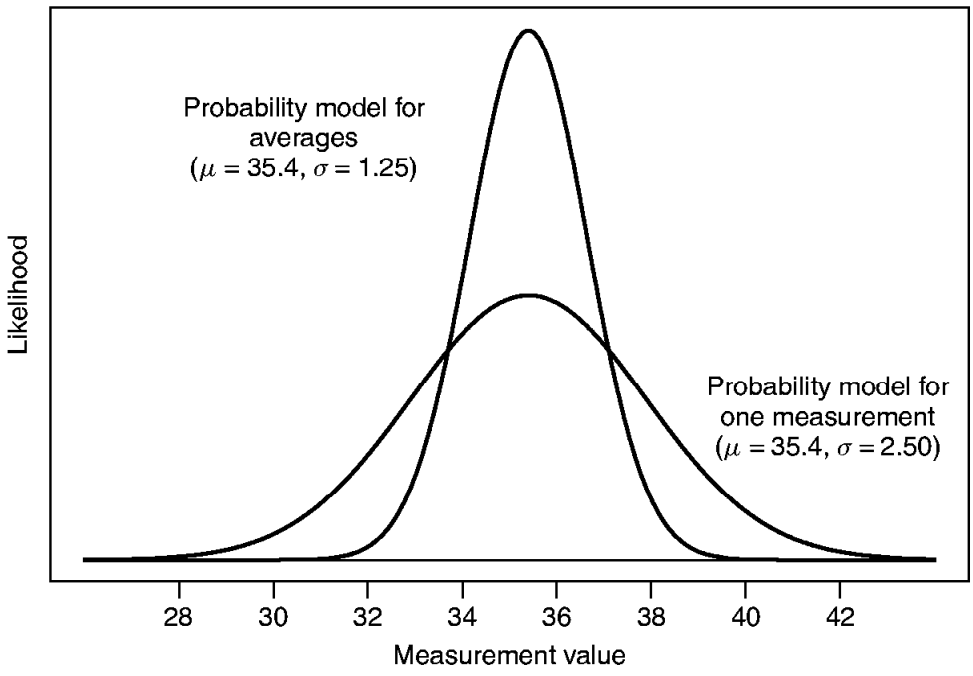


Figure 1.7 Comparison of theoretical normal distributions.

population mean but with a standard deviation that is reduced by a factor of $\sqrt{n}$ from that of the original population. Thus, averages of random samples of size 4 have standard deviations that are half that of the original population. Figure 1.7 shows the relationship between a normal probability model for individual measurements that have a population mean of $\mu = 35.4$ and a standard deviation of $\sigma = 2.5$ and one for the corresponding population of sample averages of size 4. Note that the latter distribution has $\mu = 35.4$ but $\sigma = 2.5/\sqrt{4} = 1.25$. The distribution of the averages is more concentrated around the population mean than is the distribution of individual observations. This indicates that it is much more likely to obtain a sample average that is in a fixed interval around the population mean than it is to obtain a single observation in the same fixed interval.

This discussion is intended to highlight the informational content of population parameters and to shed some light on the model-building processes involved in drawing inferences from sample statistics. The final section in this chapter focuses on one additional issue, which helps to distinguish statistical from mathematical problem solving.

1.4 MATHEMATICAL AND STATISTICAL MODELING

Models and model building are commonplace in the engineering and physical sciences. A research engineer or scientist generally has some basic knowledge about the phenomenon under study and seeks to use this information

to obtain a plausible model of the data-generating process. Experiments are conducted to characterize, confirm, or reject models — in particular, through hypotheses about those models. Models take many shapes and forms, but in general they all seek to characterize one or more response variables, perhaps through relationships with one or more factors.

Mathematical models, as defined in Exhibit 1.15, have the common trait that the response and predictor variables are assumed to be free of specification error and measurement uncertainty. Mathematical models may be poor descriptors of the physical systems they represent because of this lack of accounting for the various types of errors included in statistical models. Statistical models, as defined in Exhibit 1.16, are approximations to actual physical systems and are subject to specification and measurement errors.

EXHIBIT 1.15

Mathematical Model. A model is termed *mathematical* if it is derived from theoretical or mechanistic considerations that represent exact, error-free assumed relationships among the variables.

EXHIBIT 1.16

Statistical Model. A model is termed *statistical* if it is derived from data that are subject to various types of specification, observation, experimental, and/or measurement errors.

An example of a mathematical model is the well-known fracture mechanics relation:

$$K_{IC} = \gamma S a^{1/2}, \tag{1.1}$$

where K_{IC} is the critical stress intensity factor, S is the fracture strength, a is the size of the flaw that caused the fracture, and γ is a constant relating to the flaw geometry. This formula can be utilized to relate the flaw size of a brittle material to its fracture strength. Its validity is well accepted by mechanical engineers because it is based on the theoretical foundations of fracture mechanics, which have been confirmed through extensive experimental testing.

Empirical studies generally do not operate under the idealized conditions necessary for a model like equation (1.1) to be valid. In fact, it often is not possible to postulate a mathematical model for the mechanism being studied. Even when it is known that a model like equation (1.1) should be valid, experimental error may become a nontrivial problem. In these situations statistical models

are important because they can be used to approximate the response variable over some appropriate range of the other model variables. For example, additive or multiplicative *errors* can be included in the fracture-mechanics model, yielding the statistical models

$$K_{IC} = \gamma S a^{1/2} + e \quad \text{or} \quad K_{IC} = \gamma S a^{1/2} e \tag{1.2}$$

where e is the error. Note that the use of "error" in statistical models is not intended to indicate that the model is incorrect, only that unknown sources of uncontrolled variation, often measurement error, are present.

A mathematical model, in practice, can seldom be proved with data. At best, it can be concluded that the experimental data are consistent with a particular hypothesized model. The chosen model might be completely wrong and yet this fact might go unrecognized because of the nature of the experiment; e.g., data collected over a very narrow range of the variables would be consistent with any of a vast number of models. Hence, it is important that proposed mathematical models be sufficiently "strained" by the experimental design so that any substantial discrepancies from the postulated model can be identified.

In many research studies there are mathematical models to guide the investigation. These investigations usually produce statistical models that may be partially based on theoretical considerations but must be validated across wide ranges of the experimental variables. Experimenters must then seek "lawlike relationships" that hold under a variety of conditions rather than try to build separate statistical models for each new data base. In this type of model generalization, one may eventually evolve a "theoretical" model that adequately describes the phenomenon under study.

REFERENCES

Text References

The following books provide excellent case studies. Brief summaries of each are provided below.

Andrews, D. F. and Herzberg, A. M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*, New York: Springer-Verlag, Inc. *This book is a collection of 71 data sets with descriptions of the experiment or the study from which the data were collected. The data sets exemplify the rich variety of problems for which the science of statistics can be used as an integral component in problem-solving.*

Peck, R., Haugh, L. D., and Goodman, A. (1998). *Statistical Case Studies: A Collaboration Between Academe and Industry. ASA-SIAM Series on Statistics and Applied Probability*. Philadelphia: Society for Industrial and Applied Mathematics.

This collection of 20 case studies is unique in that each is co-authored by at least one academic and at least one industrial partner. Each case study is based on an actual project with specific research goals. A wide variety of data collection methods and data analysis techniques are presented.

Snee, R. D., Hare, L. B., and Trout, J. R. (1985). *Experiments in Industry: Design, Analysis, and Interpretation of Results*, Milwaukee, WI: American Society for Quality Control.

This collection of eleven case histories focuses on the design, analysis, and interpretation of scientific experiments. The stated objective is "to show scientists and engineers not familiar with this methodology how the statistical approach to experimentation can be used to develop and improve products and processes and to solve problems."

Tanur, J. M., Mosteller, F., Kruskal, W. H., Link, R. F., Pieters, R. S., and Rising, G. R. (1972). *Statistics: A Guide to the Unknown*, San Francisco: Holden-Day, Inc.

This collection of 44 essays on applications of statistics presents excellent examples of the uses and abuses of statistical methodology. The authors of these essays present applications of statistics and probability in nontechnical expositions which for the most part do not require previous coursework in statistics or probability. Individual essays illustrate the benefits of carefully planned experimental designs as well as numerous examples of statistical analysis of data from designed experiments and observational studies.

The following books contain excellent discussions on the impact of variation on processes and on achieving business objectives:

Leitnaker, M. G., Sanders, R. D., and Hild, C. (1996). *The Power of Statistical Thinking: Improving Industrial Processes*. New York: Addison Wesley Publishing Company.

Joner, B. L. (1994). *Fourth Generation Management: The New Business Consciousness*. New York: McGraw-Hill, Inc.

The distinction between populations and samples and between population parameters and sample statistics is stressed in most elementary statistics textbooks. The references at the end of Chapter 2 provide good discussions of these concepts. There are many excellent textbooks on statistical sampling techniques, including:

Cochran, W. G. (1977). *Sampling Techniques, Third Edition*, New York: John Wiley and Sons, Inc.

Scheaffer, R. L., Mendenhall, W., and Ott, L. (1996). *Elementary Survey Sampling, Fifth Edition*, North Scituate, MA: Duxbury Press.

The first of these texts is a classic in the statistical literature. The second one is more elementary and a good reference for those not familiar with sampling techniques.

Explicit instructions for constructing frequency distributions, histograms, and scatterplots can be found in most elementary statistics texts, including:

Freedman, D., Pisani, R., and Purves, R. (1997). *Statistics Third Edition*, New York: W. W. Norton & Company, Chapters 3 and 7.

Koopmans, L. (1981). *An Introduction to Contemporary Statistics*, Belmont, CA: Duxbury Press, Chapters 1 and 4.

Ott, L. (1977). *An Introduction to Statistical Methods and Data Analysis*, Belmont, CA: Duxbury Press, Chapters 1 and 6.

Mathematical and statistical modeling is not extensively covered in introductory statistics textbooks. The following text does devote ample space to this important topic:

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*. New York: John Wiley & Sons, Inc., Chapters 9 and 16.

EXERCISES

1 The Department of Transportation (DOT) was interested in evaluating the safety performance of motorcycle helmets manufactured in the United States. A total of 264 helmets were obtained from the major U.S. manufacturers and supplied to an independent research testing firm where impact penetration and chin retention tests were performed on the helmets in accordance with DOT standards.

(a) What is the population of interest?

(b) What is the sample?

(c) Is the population finite or infinite?

(d) What inferences can be made about the population based on the tested samples?

2 List and contrast the characteristics of population parameters and sample statistics.

3 A manufacturer of rubber wishes to evaluate certain characteristics of its product. A sample is made from a warehouse containing bales of synthetic rubber. List some of the possible candidate populations from which this sample can be taken.

4 It is known that the bales of synthetic rubber described in Exercise 3 are stored on pallets with a total of 15 bales per pallet. What type of sampling methodology is being implemented under the following sample scenarios?

(a) Five pallets of bales are randomly chosen; then eight bales of rubber are randomly selected from each pallet.

(b) Forty bales are randomly selected from the 4500 bales in the warehouse.

(c) All bales are sampled on every fifth pallet in the warehouse.

(d) All bales that face the warehouse aisles and can be reached by a forklift truck are selected.

5 Recall the normal distribution discussed in Section 1.3. What is the importance of $\mu \pm 3\sigma$?

6 The population mean and standard deviation of typical cetane numbers measured on fuels used in compression–ignition engines is known to be

$\mu = 30$ and $\sigma = 5$. Fifteen random samples of these fuels were taken from the relevant fuel population, and the sample means and standard deviations were calculated. This random sampling procedure was repeated (replicated) nine times.

Replicate No.	n	Sample Mean	Sample Standard Deviation
1	15	32.61	4.64
2	15	28.57	6.49
3	15	29.66	4.68
4	15	30.09	5.35
5	15	30.11	6.39
6	15	28.02	4.05
7	15	30.09	5.35
8	15	29.08	3.56
9	15	28.91	4.88

Consider the population of all sample means of size $n = 15$. What proportion of means from this population should be expected to be between the 30 ± 5 limits? How does this sample of nine averages compare with what should be expected?

7 A research program was directed toward the design and development of self-restoring traffic-barrier systems capable of containing and redirecting large buses and trucks. Twenty-five tests were conducted in which vehicles were driven into the self-restoring traffic barriers. The range of vehicles used in the study included a 1800-lb car to a 40,000-lb intercity bus. Varying impact angles and vehicle speeds were used, and the car damage, barrier damage, and barrier containment were observed.

(a) What is an observation in this study?

(b) Which variables are responses?

(c) Which variables are factors?

8 Space vehicles contain fuel tanks that are subject to liquid sloshing in low-gravity conditions. A theoretical basis for a model of low-gravity sloshing was derived and used to predict the slosh dynamics in a cylindrical tank. Low-gravity simulations were performed in which the experimental results were used to verify a statistical relationship. It was shown in this study that the statistical model closely resembled the theoretical model. What type of errors are associated with the statistical model? Why aren't the statistical and theoretical models exactly the same?

9 Use a table of random numbers or a computer-generated random number sequence to draw 20 simple random samples, each of size $n = 10$, from the population of employees listed in Table 1.3. Calculate the average of

the skin elasticity measurements (Table 1.4) for each sample. Graph the distribution of these 20 averages in a form similar to Figure 1.5. Would this graph be sufficient for you to conclude that the sampling distribution of the population of averages is a normal distribution? Why (not)?

10 Use the table of random numbers in the Appendix to choose starting points between 1 and 16 for ten systematic random samples of the population of employees listed in Table 1.3. Select every 10th employee. Calculate the average of the skin elasticity measurements for each sample. Does a graph of the distribution of these averages have a similar shape to that of Exercise 9?

11 Simple random samples and systematic random samples often result in samples that have very similar characteristics. Give three examples of populations that you would expect to result in similar simple and systematic random samples. Explain why you expect the samples to be similar. Give three examples for which you expect the samples to be different. Explain why you expect them to be different.

12 A series of valve closure tests were conducted on a 5-inch-diameter speed-control valve. The valve has a spring-loaded poppet mechanism that allows the valve to remain open until the flow drag on the poppet is great enough to overcome the spring force. The poppet then closes, causing flow through the valve to be greatly reduced. Ten tests were run at different spring locking-nut settings, where the flow rate at which the valve poppet closed was measured in gallons per minute. Produce a scatterplot of these data. What appears to be the effect of nut setting on flow rate?

Nut Setting	Flow Rate	Nut Setting	Flow Rate
0	1250	10	2085
2	1510	12	1503
4	1608	14	2115
6	1650	16	2350
8	1825	18	2411

13 A new manufacturing process is being implemented in a factory that produces automobile spark plugs. A random sample of 50 spark plugs is selected each day over a 15-day period. The spark plugs are examined and the number of defective plugs is recorded each day. Plot the following data in a scatterplot with the day number on the horizontal axis. A scatterplot with time on the horizontal axis is often called a *sequence plot*. What does the plot suggest about the new manufacturing process?

Day	No. of Defectives	Day	No. of Defectives
1	3	9	4
2	8	10	3
3	4	11	6
4	5	12	4
5	5	13	4
6	3	14	3
7	4	15	1
8	5		

14 Construct two histograms from the solar-energy data in Exercise 3 of Chapter 2. Use the following interval widths and starting lower limits for the first class. What do you conclude about the choices of the interval width for this data set?

	Histogram 1	Histogram 2
Interval width	8	2
Starting lower limit	480	480

15 The following data were taken from a study of red-blood-cell counts before and after major surgery. Counts were taken on 23 patients, all of whom were of the same sex (female) and who had the same blood type (O+).

	Count			Count	
Patient	Pre-op	Post-op	Patient	Pre-op	Post-op
1	14	0	13	5	6
2	13	26	14	4	0
3	4	2	15	15	3
4	5	4	16	4	2
5	18	8	17	0	3
6	3	1	18	7	0
7	6	0	19	2	0
8	11	3	20	8	13
9	33	23	21	4	24
10	11	2	22	4	6
11	3	2	23	5	0
12	3	2			

(a) Construct histograms of the pre-op and the post-op blood counts. What distinguishing features, if any, are there in the distributions of the blood counts?

(b) Make a scatter diagram of the two sets of counts. Is there an apparent relationship between the two sets of counts?

16 Satellite sensors can be used to provide estimates of the amounts of certain crops that are grown in agricultural regions of the United States. The following data consist of two sets of estimates of the proportions of each of 33 5 × 6-nautical-mile segments of land that are growing corn during one time period during the crop season (the rest of the segment may be growing other crops or consist of roads, lakes, houses, etc.). Use the graphical techniques discussed in this chapter to assess whether these two estimation methods are providing similar information on the proportions of these segments that are growing corn.

	Proportion Growing Corn			Proportion Growing Corn	
Segment	**Method 1**	**Method 2**	**Segment**	**Method 1**	**Method 2**
1	0.49	0.24	18	0.61	0.33
2	0.63	0.32	19	0.50	0.20
3	0.60	0.51	20	0.62	0.65
4	0.63	0.36	21	0.55	0.51
5	0.45	0.23	22	0.27	0.31
6	0.64	0.26	23	0.65	0.36
7	0.67	0.36	24	0.70	0.33
8	0.66	0.95	25	0.52	0.27
9	0.62	0.56	26	0.60	0.30
10	0.59	0.37	27	0.62	0.38
11	0.60	0.62	28	0.26	0.22
12	0.50	0.31	29	0.46	0.72
13	0.60	0.56	30	0.68	0.76
14	0.90	0.90	31	0.42	0.36
15	0.61	0.32	32	0.68	0.34
16	0.32	0.33	33	0.61	0.28
17	0.63	0.27			

CHAPTER 2

Fundamentals of Statistical Inference

In this chapter, basic summary statistics that are commonly used to summarize important characteristics of data are introduced. Accompanying these basic statistics are interval estimation and hypothes testing is methods that can be used to draw inferences on a population or process from data collected in an experiment. Specifically, this chapter:

- *describes traditional summary statistics: averages, medians, standard deviations, and quartiles;*
- *discusses sampling distributions and their role in statistical inference;*
- *develops the fundamental inferential concepts of interval estimation and hypothesis testing; and*
- *illustrates basic techniques for sample size determination.*

2.1 TRADITIONAL SUMMARY STATISTICS

Describing the numerical results of a sequence of tests is a fundamental requirement of most experimental work. This description may utilize a variety of quantitative values, such as averages or medians. In this chapter, the calculation and display of summary statistics that describe the center (location) and spread (variation) of a set of data values are introduced.

The most commonly used numerical measure of the center of a set of data values is the sample mean, or average. The sample mean is easy to calculate and is readily interpretable as a "typical" value of the data set. Although the sample mean is familiar to most readers of this text, we include its definition in Exhibit 2.1 for completeness. Inferences about "typical" values of populations

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

are frequently made by using the sample mean to draw inferences on the population mean.

EXHIBIT 2.1

Sample Mean or Average. The sample mean or average of a set of data values is the total of the data values divided by the number of observations. Symbolically, if $y_1, y_2, \ldots, y_n$ denote n data values, the sample mean, denoted $\bar{y}$, is

$$\bar{y} = n^{-1} \sum y_i = \frac{y_1 + y_2 + \cdots + y_n}{n}.$$

An alternative to the average as a measure of the center of a set of data values is the sample median (Exhibit 2.2). In the definition in Exhibit 2.2, we distinguish the raw data values $y_1, y_2, \ldots, y_n$ from the ordered data values $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$. The sample median is either the middle data value (if n is odd), or it is the average of the two middle data values (if n is even), after the data have been ordered from smallest to largest. The sample median also is referred to as the 50th percentile of the data, because 50% of the data values are less than or equal to the median and 50% are greater than or equal to it.

The sample mean is generally preferred to the sample median as a summary measure of a set of data values. It is intuitively more reasonable than the median in most data analyses because it is computed using all the data values whereas the median only uses the middle one or two data values. Because the sample median is only based on the middle one or two observations, however, it is less susceptible to the influence of extreme data values than is the sample mean.

EXHIBIT 2.2

Sample Median. The sample median M is a number that divides ordered data values into two groups of equal size. It is determined as follows:

(i) order the data from the smallest to the largest values, denoting the ordered data values by

$$y_{(1)} \leqslant y_{(2)} \leqslant \cdots \leqslant y_{(n)};$$

(ii) determine the median as

(a) $M = y_{(q)}$ if n is odd, where $q = (n + 1)/2$.

(b) $M = [y_{(q)} + y_{(q+1)}]/2$ if n is even, where $q = n/2$.

Measures of spread are especially important as measures of typical or expected random variation in data values. It is not sufficient to describe a set of measurements only by reporting the average or some other measure of the center of the data. For example, it is not sufficient to know that bolts that are designed to have thread widths of 1.0 mm have, on the average, a width of exactly 1.0 mm. Half of the bolts could have a thread width of 0.5 mm and the other half a thread width of 1.5 mm. If nuts that are manufactured to fit these bolts all have thread widths of 1.0 ± 0.1 mm, none of the bolts would be usable with the nuts.

Perhaps the simplest measures of the spread of a set of data values are indicated by the extremes of the data set, that is, the minimum and maximum data values. Although the most frequent use of the extremes of a data set is to indicate limits of the data, there are other important uses. For example, with the use of computers to analyze large numbers of data there is a great likelihood that mistakes in data entry will go unnoticed. Routine examination of the extremes of a data set often aid in the identification of errors in the data entry — for example, percentages that exceed 100% or negative entries for variables that can only take on positive values.

Perhaps the two most common measures of the spread of a set of data values are the range (Exhibit 2.3) and the sample standard deviation (Exhibit 2.4). The range is quick and easy to compute and is especially valuable where small amounts of data are to be collected and a quick measure of spread is needed, such as by quality-control personnel on a production line. A disadvantage of the range is that it only makes use of two of the data values and can, therefore, be severely influenced by a single erratic observation.

EXHIBIT 2.3

Range. Range = maximum data value − minimum data value.

The sample standard deviation s is calculated from all the data values using either of the formulas in Exhibit 2.4. The sample standard deviation is based on *deviations*, or differences in the individual observations from the sample mean [see formula (a)]. These deviations, $y_i - \bar{y}$, are squared, the squares are averaged, and the square root is taken of the result. The squaring is performed because a measure of spread should not be affected by whether deviations are positive ($y_i > \bar{y}$) or negative ($y_i < \bar{y}$) but only by the magnitude (the size) of the difference from the mean. Averaging by dividing by $n - 1$ rather than n is done for technical reasons, which need not concern us here; in any event, for large sample sizes the difference in divisors has a negligible effect on the calculation. Because the square root has been taken, the sample standard deviation is measured in units identical to the original observations.

EXHIBIT 2.4

Sample Standard Deviation. The sample standard deviation of a set of data values $y_1, y_2, \ldots, y_n$ can be calculated in either of two equivalent ways:

$$\text{(a)}\quad s = \left\{\sum (y_i - \bar{y})^2/(n-1)\right\}^{1/2}$$

$$= \{[(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2]/(n-1)\}^{1/2}$$

$$\text{(b)}\quad s = \left\{\left[\sum y_i^2 - n^{-1}\left(\sum y_i\right)^2\right]/(n-1)\right\}^{1/2}$$

Formula (b) is easier to use than formula (a) when calculations are made on a desk calculator rather than on a computer. The two expressions for the sample standard deviation are algebraically equivalent, and apart from roundoff error they give the same result.

In view of formula (a) one can interpret the sample standard deviation as a measure of the "typical" variation of data values around the sample mean. The larger the standard deviation, the more the variation in the data values. The smaller the standard deviation, the more concentrated the data values are around the mean. The sample standard deviation often is used as a measure of the precision of a measurement process.

The question of what constitutes excessive spread of a sample of data values is difficult to answer. The interpretation of a measure of spread is enhanced when it can be meaningfully compared either with a standard value such as a specification limit or with values previously obtained from similar measurements. Occasionally several data sets of similar measurements are to be compared and the relative magnitudes of the standard deviations provide valuable information on differences in variability of the processes that generated the data sets.

Quartiles (Exhibit 2.5) are another important set of measures of the spread of data values. They are usually unaffected by a few extreme measurements. The first quartile (25th percentile) is a numerical value that divides the (ordered) data so that 25% of the data values are less than or equal to it, the second quartile (50th percentile) is the sample median, and the third quartile (75th percentile) divides the data so that 75% of the (ordered) data values are less than or equal to its value. Half the difference between the third and the first quartile, referred to as the *semi-interquartile range* (Exhibit 2.5), is a measure of the spread of the data values that is quick to compute and is less affected by extremes in the data than is the sample standard deviation.

EXHIBIT 2.5 QUARTILES, SEMI-INTERQUARTILE RANGE

Quartiles Q_1, Q_2, and Q_3 are numerical values that divide a sample of observations into groups so that one-fourth (25%) of the data values are less than or equal to Q_1, half (50%) of the values are less than or equal to Q_2, and three-fourths (75%) of the values are less than or equal to Q_3. The second quartile Q_2 is the sample median M. Quartiles are determined as follows:

1. Order the data values:
$$y_{(1)} \leqslant y_{(2)} \leqslant \cdots \leqslant y_{(n)}.$$

2. Let $q = (n + 1)/2$ if n is odd and $q = n/2$ if n is even. Then $Q_2 = M =$
   (a) $y_{(q)}$ if n is odd,
   (b) $(y_{(q)} + y_{(q+1)})/2$ if n is even.
3. **(a)** If q is odd, let $r = (q + 1)/2$. Then
$$Q_1 = y_{(r)} \quad \text{and} \quad Q_3 = y_{(n+1-r)}.$$

   (b) If q is even, let $r = q/2$. Then
$$Q_1 = \frac{y_{(r)} + y_{(r+1)}}{2}$$
and
$$Q_3 = \frac{y_{(n+1-r)} + y_{(n-r)}}{2}$$

The semi-interquartile range (SIQR) is
$$\text{SIQR} = \frac{Q_3 - Q_1}{2}.$$

Figure 2.1 displays the results of a fuel-economy study of four diesel-powered automobiles. All four vehicles were tested under controlled conditions in a laboratory using dynamometers to control the speed and load conditions. For each vehicle, Figure 2.1 graphs the fuel economy (mi/gal) for eight tests using eight different test fuels. The eight mileage values for the Mercedes and the Peugeot exhibit similar distributions, but the values for the Oldsmobile and the Volkswagen are clearly different from the other two. Note particularly the low test result for one of the fuels on the Volkswagen. This erratic mileage value occurred because one of the test fuels could not be blended according to

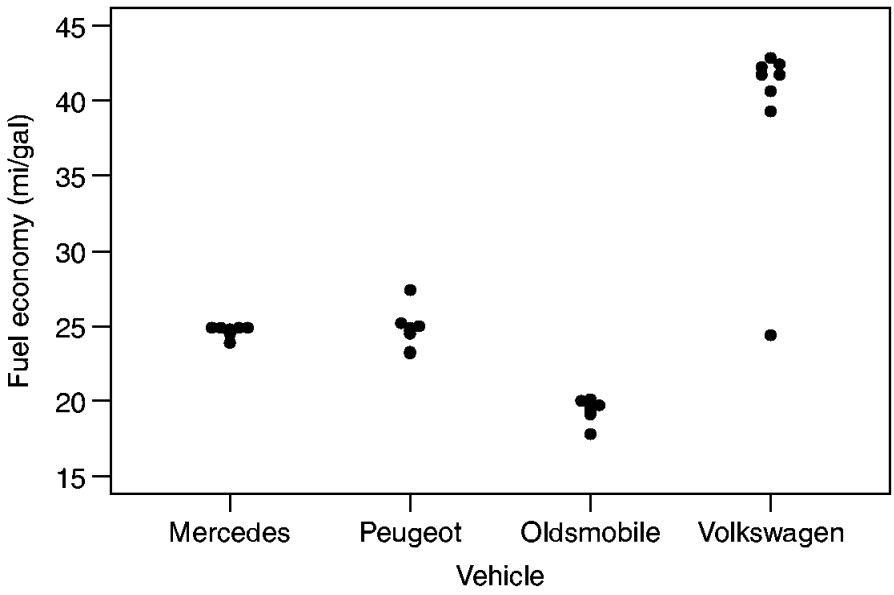


Figure 2.1 Fuel-economy measurements.

TABLE 2.1 Fuel Economy for Four Test Vehicles Using Eight Different Fuels

	Fuel Economy (mi/gal)			
Fuel Number	Mercedes	Peugeot	Oldsmobile	Volkswagen
1	24.8	24.5	19.6	41.7
2	24.7	25.2	20.0	42.2
3	24.9	27.4	19.3	41.7
4	23.9	23.3	19.1	39.3
5	24.9	25.0	19.7	42.8
6	24.9	24.7	20.1	42.4
7	24.9	23.2	17.8	24.4
8	24.5	24.9	19.6	40.6
		Summary Statistics		
Average	24.7	24.8	19.4	39.4
Median	24.8	24.8	19.6	41.7
S.D.	0.3	1.3	0.7	6.2
Min	23.9	23.2	17.8	24.4
Max	24.9	27.4	20.1	42.8
Range	1.0	4.2	2.3	18.4
Quartiles:				
First	24.6	23.9	19.2	40.0
Third	24.9	25.1	19.8	42.3
SIQR	0.20	0.85	0.30	1.45

the experimental design; that is, it would not produce satisfactory combustion in diesel engines. The test fuel then was altered with additives until it did produce satisfactory combustion.

The visual impressions left by Figure 2.1 are quantified in Table 2.1. The averages and medians listed in Table 2.1 depict the typical performance of the four vehicles. The Volkswagen is seen to have higher average fuel economy than the other three vehicles for these tests, with the Oldsmobile having a slightly lower average and median fuel economy than the Mercedes and the Peugeot. Note the averages and sample medians are very similar for the Mercedes, Peugeot, and Oldsmobile, but markedly different for the Volkswagen because of the one extreme mileage reading. Because this low mileage value has no effect on the sample median, compared to the sample mean, the sample median is the preferred indicator of the data center.

The Mercedes test results are most consistent in that the eight fuel-economy values exhibit the least spread as indicated by the ranges, the semi-interquartile ranges, and the sample standard deviations. The Volkswagen results are the least consistent as a result of the one extreme reading.

The statistics discussed in this section are the most commonly used measures of the center and spread of a set of data. They allow one to summarize a set of data with only a few key statistics. They are especially valuable when large numbers of data prevent the display of an entire data set or when the characteristics that are summarized by the sample statistics are themselves the key properties of interest in the data analysis.

2.2 STATISTICAL INFERENCE

Most statistically designed experiments are conducted with the intent of better understanding populations, observable phenomena, or stable processes. In some instances, sample statistics are computed to draw conclusions about the corresponding population parameters. In others, factor effects are calculated to compare the influences of different factor levels. In both of these examples and in many other applications of statistical methodology, statistical models are postulated and inferences are drawn relative to the specification of the model. Any such inference or conclusion involves some degree of uncertainty because of the uncontrolled experimental variability of the observed responses.

The data displayed in Table 2.2 are tensile-strength measurements on wire that is used in industrial equipment. The wire is manufactured by drawing the base metal through a die that has a specified diameter. The data in Table 2.2 are tensile strengths of eighteen samples taken from large spools of wire made with each of three dies. It is of interest to determine whether the use of different dies will affect the tensile strength of the wire.

Figure 2.2 displays *point plots* of the tensile strength measurements for each of the three dies. Point plots are simple graphs of small data sets that

TABLE 2.2 Tensile Strengths of Wire from Three Different Dies

Tensile Strength (10^3 psi)					
Die 1		Die 2		Die 3	
85.769	86.725	79.424	82.912	82.423	81.768
86.725	84.292	81.628	83.185	81.941	83.078
87.168	84.513	82.692	86.725	81.331	83.515
84.513	86.725	86.946	84.070	82.205	82.423
84.513	84.513	86.725	86.460	81.986	83.078
83.628	82.692	85.619	83.628	82.860	82.641
82.912	83.407	84.070	84.513	82.641	82.860
84.734	84.070	85.398	84.292	82.592	82.592
82.964	85.337	86.946	85.619	83.026	81.507

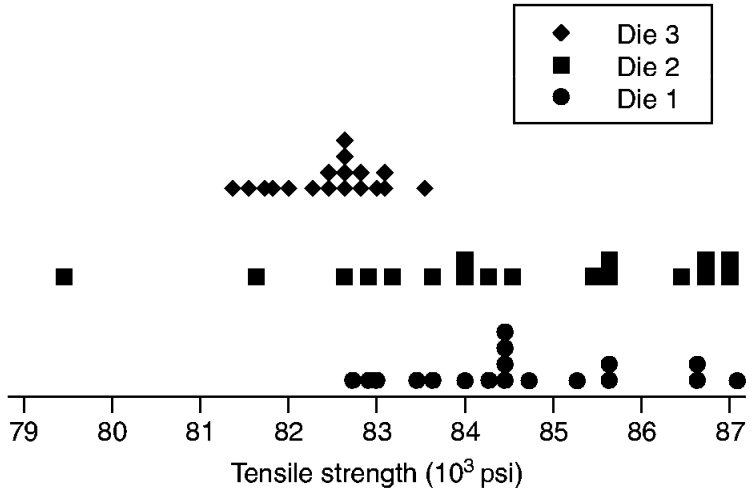


Figure 2.2 Tensile strength measurements of wire from three dies.

enable one to readily visualize location and spread characteristics of small data sets. Point plots are constructed as indicated in Exhibit 2.6. It is apparent from Figure 2.2 that the centers of the measurements of the three dies differ and that the measurements from die 3 appear to be less variable than those of the other two dies.

EXHIBIT 2.6 POINT PLOTS

1. Construct a horizontal axis covering the range of data values.
2. Vertically stack repeated data values.

TABLE 2.3 Summary Information on Tensile-Strength Data

Statistic	Die 1	Die 2	Die 3
Average	84.733	84.492	82.470
Median	84.513	84.403	82.592
Standard deviation	1.408	2.054	0.586
n	18	18	18

The characteristics that are so visually apparent from Figure 2.2 are quantified in Table 2.3. This table contains summary statistics calculated from the measurements for the wires from each die.

The tensile strengths for the wires from die 3 average about 2000 psi lower than those of the other dies. The point plot suggests that this difference in average tensile strengths is due to generally smaller tensile strengths for wire from die 3. Several interesting questions emerge from the information provided in Table 2.3. One question is whether this information is sufficient to conclude that the third die produces wire with a smaller mean tensile strength than the other two dies.

This question is difficult to answer if one only examines the information shown in Table 2.3. While the average tensile strength for the wires from die 3 is lower than those for the other two dies, the wires from the other two dies exhibit greater variability, as measured by the sample standard deviations, than those from die 3. In addition, the wire sample with the lowest tensile strength measurement, 79.424, is made from die 2 not die 3. Under some circumstances comparisons among dies are fairly straightforward; for instance, when all the measurements for one die are lower than the minimum tensile strengths for the other dies. When measurements overlap, as they do in the point plots in Figure 2.2, decisions on the relative performance of the dies are more difficult. It is in these circumstances that statistical analyses that incorporate measures of uncontrolled variability are needed.

Most of the statistical procedures used in this book to analyze data from designed experiments are based on modeling the response variable. The statistical models used include various parameters (constants) that, depending on the experimental situation, characterize either location or the influence of factor levels. These models also include variables that represent either random assignable causes or uncontrolled experimental variability. These variables are assumed to follow certain probability distributions, most frequently a normal probability distribution.

In Section 2.3, probability distributions and the calculation of probabilities are discussed. Also discussed are sampling distributions for sample means

and variances. Sections 2.4 and 2.5 are devoted to an exposition of interval estimation in which measures of variability are explicitly included in inferential estimation procedures. Section 2.6 details the use of statistical hypothesis-testing principles for drawing inferences. The determination of sample-size requirements for certain types of experimental settings is presented in Section 2.7.

2.3 PROBABILITY CONCEPTS

A probability is a number between zero and one that expresses how likely an event, an action, or a response is to occur. A probability close to one indicates that an event is very likely to occur, whereas a probability close to zero indicates that an event is very unlikely to occur. Probabilities sometimes are expressed as percentages, in which case we shall refer to them as chances; that is chance = probability $\times$ 100%.

Probabilities are calculated in a variety of ways. One of the simplest methods of calculating probabilities occurs when there are N equally likely outcomes (responses) that could occur. The probability of observing any specified outcome is then $1/N$. If m of these outcomes constitute an event of interest, the probability p of observing the event of interest is then $p = m/N$. Games of chance (e.g., drawing cards from a well-shuffled deck) often satisfy the requirements for calculating probabilities this way.

Probabilities can be calculated for any population or process if a probability distribution can be assumed for the response variables. Probability distributions specify the possible values of a response variable and either the probability of observing each of the response values or a density function for the values.

Probabilities for many probability distributions are obtainable, as in the above illustration, from formulas that relate an individual response to its probability. These distributions are for discrete response variables, those whose permissible values form a finite or countable infinite set. The probability distribution for N equally likely outcomes is expressed as a listing of each of the outcomes and a tabular or graphical display of the probabilities, $1/N$, for each.

Using appropriate probability rules for combining outcomes, discrete probability distributions can be determined for phenomena as diverse as games of chance and lotteries, multiple choice tests, public opinion polls, and pharmaceutical drug testing.

A *density* is a function that defines the likelihood of obtaining various response values for *continuous* response variables (those whose permissible values are any real number in a finite or infinite interval). Probabilities for continuous response variables are obtained as areas under the curve defined by the density function.

The normal probability distribution was introduced in Section 1.3 as an appropriate reference distribution for many types of response variables. The form of the normal density function is

$$f(y) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \tag{2.1}$$

for

$$-\infty < y < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0.$$

In the density function (2.1), y is a response variable, μ is the population mean, and σ is the population standard deviation.

The degree to which the probability distribution of the responses assumed by a model (the reference distribution) must agree with the actual distribution of the responses in a population or from a process depends on the statistical methodology used. Some statistical procedures are very robust to specific differences between the assumed and the actual distribution of the responses. Others are very sensitive to such differences. The researcher must identify any assumptions needed to implement statistical procedures and assess whether the needed assumptions are reasonable for the data being analyzed. The assessment of model assumptions is dealt with in detail in Chapter 18 for linear regression models. These techniques can be applied to experimental-design models if factor levels are recoded to appropriate numerical values (see Section 16.1).

The normal probability distribution is a possible reference distribution for the tensile-strength measurements of the wires from any one of the dies in the example discussed in the previous section. In this setting, μ is the mean tensile strength for all wire produced by a particular die. The standard deviation σ controls the spread in the distribution of the responses (see Figure 1.7). This spread represents the cumulative effects of a variety of sources contributing to the uncontrolled variability of the measurements.

Probabilities for continuous response variables are obtained as areas under the plotted density function. In some cases these areas can be obtained analytically through integration of the density function. In other cases, the integral has no closed-form solution, but numerical solutions for the integrals can be obtained using appropriate computer software. Many of the density functions used in this text have been conveniently tabulated so that no integration is needed. Such is the case for the standard normal distribution (see Table A2 in the appendix). Use of probability tables for the standard normal probability distribution is explained in the appendix to this chapter. Probabilities for any normal distribution can be calculated from those of a standard normal distribution, if the mean μ and standard deviation σ are known, by transforming the original normal variable to a standard normal variable (see Exhibit 2.7).

EXHIBIT 2.7 STANDARD NORMAL VARIABLES

If y is a response variable following a normal distribution with population or process mean μ and standard deviation σ, the variate

$$z = \frac{y - \mu}{\sigma} \tag{2.2}$$

follows a standard normal distribution; that is, z is normally distributed with mean 0 and standard deviation 1.

In Chapter 1 the distinctions between populations and samples and between parameters and statistics are stressed. Experiments are performed to collect data (samples) from which to draw inferences on populations or processes. Frequently these inferences are conclusions drawn about parameters of models or probability distributions, conclusions that are based on the calculation of statistics from the individual responses. Many of the statistics used and the resulting inferences require an assumption that the individual responses are independent (see Exhibit 2.8).

EXHIBIT 2.8

Statistical Independence. Observations are statistically independent if the value of one of the observations does not influence the value of any other observation. Simple random sampling produces independent observations.

Sample statistics follow sampling distributions (see Section 1.3) just as individual response variables follow probability distributions. This is because there are many samples (perhaps an infinite number) of a fixed size n that could be drawn from a population or a process. Each sample that could be drawn results in a value for a sample statistic. The collection of these possible values forms the sampling distribution.

It is important to the proper understanding of a sampling distribution to distinguish between a statistic and a realization of a statistic. Prior to collecting data, a statistic has a sampling distribution from which a realization will be taken. The sampling distribution is determined by how the data are to be collected, that is, the experimental design or the sampling procedure. The realization is simply the observed value of the statistic once it is calculated from the sample of n responses. This distinction is often stressed in the terminology used in the estimation of parameters. An *estimator* is the statistic used to estimate the parameter, and an *estimate* is the realized or calculated value from the sample responses.

If individual response variables are statistically independent and follow a normal probability distribution, the population of all possible sample means of a fixed sample size also follows a normal probability distribution (see Exhibit 2.9).

EXHIBIT 2.9 SAMPLING DISTRIBUTION OF SAMPLE MEANS

If independent observations $y_1, y_2, y_3, \ldots$ follow a normal probability distribution with mean μ and standard deviation σ, then the distribution of all possible sample means $\bar{y}_1, \bar{y}_2, \ldots$ of size n from this population is also normal with population mean μ but with a standard error (standard deviation) of $\sigma/n^{1/2}$.

The term *standard error* (see Exhibit 2.10) is used to distinguish the standard deviation of a sample statistic from that of an individual observation, just as the term *sampling distribution* is used to distinguish the distribution of the statistic from that of the individual observation. The standard error is the most frequently used measure of the precision of a parameter estimator. The standard error measures the precision of an estimator as a function of both the standard deviation of the original population of measurements and the sample size. For example, the standard error of a sample mean of n independent observation is

$$\sigma_{\bar{y}} = \sigma/n^{1/2}. \tag{2.3}$$

Note that this standard error can be small if either the population standard deviation σ is small or the sample size n is large. Thus, both the standard deviation of the measurements and the sample size contribute to the standard error, and hence to the precision, of the sample mean.

EXHIBIT 2.10

Standard Error. The standard error of a statistic is the standard deviation of its sampling distribution. The standard error is usually obtained by taking the positive square root of the variance of the statistic.

The distribution of sample means of independent observations from a normal distribution can be standardized as in (2.2):

$$z = \frac{\bar{y} - \mu}{\sigma/n^{1/2}} = \frac{n^{1/2}(\bar{y} - \mu)}{\sigma} \tag{2.4}$$

follows a standard normal distribution. If σ is known in (2.4), inferences on an unknown mean μ can be made using the standard normal distribution and the inference procedures described in Sections 2.4 to 2.6 (see also Chapter 3).

When the population standard deviation is unknown, especially if the sample size is not large, the standard normal variate (2.4) cannot be used to draw inferences on the population mean. On replacing the standard deviation σ in (2.4) with the sample standard deviation s (Section 2.1), the sampling distribution of the resulting variate is not a standard normal distribution. Rather, the sampling distribution of the variate

$$t = \frac{n^{1/2}(\bar{y} - \mu)}{s} \tag{2.5}$$

is a *Student's t-distribution.* Probabilities for Student's t-distribution depend on the sample size n through its *degrees of freedom.* If n independent observations are taken from a normal distribution with mean μ and standard deviation σ, the degrees of freedom are $n - 1$. Probability calculations for Student's t-distribution are described in the appendix to this chapter. Probability tables for this distribution are given in Table A3 of the appendix to this book.

Frequently interest in an industrial process centers on measuring variation rather than on estimates of a population mean. Inferences on population or process standard deviations for independent observations from normal probability distributions can be made using the variate

$$X^2 = (n - 1)s^2/\sigma^2. \tag{2.6}$$

The sampling distribution of the variate (2.6) is a *chi-square* distribution. This distribution, like Student's t, depends on the sample size through the degrees of freedom. The form of the statistic (2.6) is appropriate for inferences on σ based on independent observations from a normal distribution, in which case the degrees of freedom again are $n - 1$. Probability calculations for this distribution are explained in the appendix to this chapter and probability tables are given in Table A4 in the appendix to this book.

Another frequently used sampling distribution is needed when one wishes to compare the variation in two populations or processes. The ratio of two sample variances (s^2), each divided by its population variance (σ^2), is termed an *F-variate*:

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{s_1^2\sigma_2^2}{s_2^2\sigma_1^2}. \tag{2.7}$$

Note that this form of the F-variate is the ratio of two independent chi-square variates, $(n_i - 1)s_i^2/\sigma_i^2$, each divided by its number of degrees of freedom, $n_i - 1$. The sampling distribution of this F-variate depends on the numbers of degrees of freedom, $n_1 - 1$ and $n_2 - 1$, of both of the samples. Probability calculations for the F-distribution are illustrated in the appendix to this chapter. Probability tables for the F-distribution are given in Table A5 of the appendix to this book.

The sampling distributions for the variates (2.4)–(2.7) have density functions from which probabilities are obtained as areas under the respective curves. These density functions are used to tabulate the probabilities in the various tables in the appendix. There are many different forms for normal, t, chi-square, and F statistics, depending on how the response variables are obtained (including the statistical design used, if appropriate) and the model used to define the relationship between the response variable and the assignable causes and uncontrolled random variation. While the forms of these statistics may change (and the degrees of freedom for the last three), the tables in the appendix can be used for all the forms discussed in this book.

Each of the sampling distributions for the variates in (2.4)–(2.7) requires that the individual response variables in a sample be independent and that they come from a single normal probability distribution. In practice there are many situations in which the assumption of a normal probability distribution is not reasonable. Fortunately there is a powerful theoretical result, the central limit property (see Exhibit 2.11), which enables the use of these statistics in certain situations when the individual response variables are not normally distributed. The implications of the central limit property are important for inferences on population means. Regardless of the true (usually unknown) probability distribution of individual observations, the standard normal distribution and Student's t-distribution are sampling distributions that can be used with the sample mean if the conditions listed in Exhibit 2.11 are met. As a practical guide, if underlying distributions are not badly skewed, samples of size 5 to 10 are usually adequate to invoke the central limit property. The more radical the departure from the form of a normal distribution (e.g., highly skewed distributions), the larger the sample size that is required before the normal approximation will be adequate.

The central limit property can be applied to variables that are discrete (e.g., frequency counts) or continuous (e.g., measurements). In the next several chapters, guidelines for the application of the central limit property will be provided for many important inference procedures.

EXHIBIT 2.11 CENTRAL LIMIT PROPERTY

If a sample of n observations $y_1, y_2, \ldots, y_n$ are

(a) independent, and

(b) from a single probability distribution having mean μ and standard deviation σ,

then the sampling distribution of the sample mean is well approximated by a normal distribution with mean μ and standard deviation $\sigma/n^{1/2}$, if the sample size is sufficiently large.

One note of caution about the central limit property is needed. It is stated in terms of the sample mean. It can be extended to include the sample variance and thereby permit the use of the chi-square distribution (2.6); however, larger sample sizes are needed than for the sample mean. It is also applicable to any statistic that is a linear function of the response variables, for example, factor effects and least-squares regression estimators (see Chapters 14 and 15). There are many functions of sample statistics, such as the ratio of sample variances in the F-statistic (2.7), to which the central limit property does not apply.

Randomization affords a second justification for the use of the above sampling distributions when the assumption of normality is not met. The mere process of randomization induces sampling distributions for the statistics in (2.4)–(2.7) that are quite similar to the sampling distributions of these statistics derived under the assumption that the responses are statistically independent and normally distributed. Thus, randomization not only affords protection against possible bias effects due to unexpected causes during the course of an experiment, it also provides a justification for the sampling distributions of the statistics (2.4)–(2.7) when the assumption of normality is not satisfied.

2.4 INTERVAL ESTIMATION

Parameter estimates do not by themselves convey any information about the adequacy of the estimation procedure. In particular, individual estimates do not indicate the precision with which the parameter is estimated. For example, the sample mean of the tensile-strength measurements for die 1 in Table 2.2 is 84.733. Can one state with any assurance that this estimate is close to the population mean tensile strength for die 1? By itself, this average provides no such information.

Standard errors provide information on the precision of estimators. Estimates of standard errors should therefore be reported along with the parameter estimates. A convenient and informative way to do so is with the use of *confidence intervals*. Confidence intervals are intervals formed around parameter estimates. The length of a confidence interval provides a direct measure of the precision of the estimator: the shorter the confidence interval, the more precise the estimator. Confidence intervals also allow one to be more specific about plausible values of the parameter of interest than just reporting a single value.

Consider now the construction of a confidence interval for the population mean of a normal probability distribution when the standard deviation is known. Let $z_{\alpha/2}$ denote the standard normal deviate (value) corresponding to an upper-tail probability of $\alpha/2$. This value is obtained from Table A2 in

the appendix. Because the variate in (2.4) has a standard normal probability distribution,

$$\Pr\left\{-z_{\alpha/2} < \frac{n^{1/2}(\bar{y} - \mu)}{\sigma} < z_{\alpha/2}\right\} = 1 - \alpha. \tag{2.8}$$

If one solves for μ in the inequality, the probability statement can be rewritten as

$$\Pr\{\bar{y} - z_{\alpha/2}\sigma_{\bar{y}} < \mu < \bar{y} + z_{\alpha/2}\sigma_{\bar{y}}\} = 1 - \alpha, \tag{2.9}$$

where $\sigma_{\bar{y}} = \sigma/n^{1/2}$ is the standard error of $\bar{y}$, equation (2.3). The limits in the probability statement (2.9) are said to form a $100(1 - \alpha)\%$ confidence interval on the unknown population mean. A general definition of a confidence interval is given in Exhibit 2.12. The lower and upper limits, $L(\hat{\theta})$ and $U(\hat{\theta})$, in this definition are functions of the estimator $\hat{\theta}$ of the parameter θ. Thus, it is the limits that are random, not the parameter enclosed by the limits. These limits can conceivably be calculated for a large number of samples. In the long run, $100(1 - \alpha)\%$ of the intervals will include the unknown parameter.

EXHIBIT 2.12

Confidence Interval. A $100(1 - \alpha)\%$ confidence interval for a parameter θ consists of limits $L(\hat{\theta})$ and $U(\hat{\theta})$ that will bound the parameter with probability $1 - \alpha$.

The $100(1 - \alpha)\%$ confidence interval for the mean of a normal distribution with known standard deviation is, from (2.9),

$$\bar{y} - z_{\alpha/2}\sigma_{\bar{y}} < \mu < \bar{y} + z_{\alpha/2}\sigma_{\bar{y}} \quad \text{or} \quad \bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}}. \tag{2.10}$$

Note that the limits are a function of the sample mean. The midpoint of the interval in (2.10) is the parameter estimator $\bar{y}$. The length of the confidence interval depends on the size of the standard error $\sigma_{\bar{y}}$ and on the confidence level $100(1 - \alpha)\%$. Confidence levels are ordinarily desired to be high, for example, 90–99%. A narrow confidence interval, one that indicates a precise estimate of the population mean, ordinarily requires a small standard error.

Interval estimates for means and standard deviations of normal probability models can be constructed using the technique described above and the appropriate sampling distribution. The construction of confidence intervals for means and standard deviations will be considered more fully in Chapter 3, including illustrations with the analysis of data.

It is important to stress that confidence intervals provide bounds on a population parameter: it is the bounds that are random, not the parameter value. Confidence intervals can be interpreted in several useful ways. In Exhibit 2.13, we present some statements about confidence intervals that are valuable when interpreting the numerical bounds obtained when sample values are inserted into confidence interval formulas. These interpretations will be reinforced in the next several chapters. The first of the interpretations in Exhibit 2.13 is commonly used to state the results of the confidence interval procedure. The second and third interpretations follow from the probability statement on which the confidence interval is based, e.g., equation (2.9). Ordinarily a single sample is collected and a single interval is calculated using (2.10). Based on the probability statement (2.9), if a large number of samples could be taken, $100(1-\alpha)\%$ of them would contain the unknown population parameter, in this case the population mean. Because of this, we state that we are $100(1-\alpha)\%$ confident that the one interval we have computed using the inequality (2.10) does indeed contain the unknown population mean.

EXHIBIT 2.13 INTERPRETATION OF CONFIDENCE INTERVALS

- With $100(1-\alpha)\%$ confidence, the confidence interval includes the parameter.
- The procedures used (including the sampling technique) provide bounds that include the parameter with probability $1-\alpha$.
- In repeated sampling, the confidence limits will include the parameter $100(1-\alpha)\%$ of the time.

2.5 STATISTICAL TOLERANCE INTERVALS

Tolerance intervals are extremely important in evaluating production, quality, and service characteristics in many manufacturing and service industries. Two types of tolerance intervals must be distinguished in this context: engineering tolerance intervals and statistical tolerance intervals.

Engineering tolerance intervals define an acceptable range of performance for a response. Often expressed as specification limits, engineering tolerance intervals dictate a range of allowable variation for the response variable within which the product or service will meet stated requirements.

Statistical tolerance intervals (see Exhibit 2.14) reflect the actual variability of the product or service. This actual variability may or may not be within the desired engineering tolerances. Statistical tolerance intervals are not intervals around parameter values; they are intervals that include a specified portion of the observations from a population or a process.

EXHIBIT 2.14

Statistical Tolerance Intervals. A statistical tolerance interval establishes limits that include a specified proportion of the responses in a population or a process with a prescribed degree of confidence.

Statistical tolerance intervals are derived from probability statements similar to the derivation of confidence intervals. For example, if a response variable y can be assumed to follow a normal probability distribution with known mean and standard deviation, one can write the following probability statement using the standard normal response variable (2.2):

$$\Pr(\mu - z_\alpha\sigma < y < \mu + z_\alpha\sigma) = p.$$

From this probability statement, it immediately follows that $100p\%$ of the distribution of responses are contained within the interval

$$(\mu - z_\alpha\sigma, \mu + z_\alpha\sigma), \tag{2.11}$$

where z_α is the critical value from Table A2 in the appendix corresponding to an upper-tail probability of $\alpha = (1 - p)/2$. The limits in (2.11) are referred to as *natural process limits* because they are limits that include $100p\%$ of the responses from a process that is in statistical control. Intervals similar to (2.11) form the basis for many control-charting techniques (e.g., Grant and Leavenworth 1988).

No "confidence" is attached to the interval (2.11), because all the quantities are known and fixed. This tolerance interval is special in that regard. Most applications of tolerance intervals, especially for new products and services, require the estimation of means and standard deviations.

When responses follow a normal probability distribution but the mean and standard deviation are unknown, it is natural to consider intervals of the form

$$(\bar{y} - ks, \bar{y} + ks), \tag{2.12}$$

where k is a critical value from some appropriate sampling distribution corresponding to an upper-tail area of $\alpha = (1 - p)/2$. This interval, unlike (2.11), is not an exact $100p\%$ tolerance interval for the responses. One reason this interval is not exact is that the sample mean and standard deviation are only estimates of the corresponding parameters. Because of this, the interval itself is random, whereas the true interval (2.11) is constant.

Procedures have been devised to incorporate not only the randomness of the response variable y but also the randomness of the sample statistics $\bar{y}$ and

s in the construction of statistical tolerance intervals. In doing so, one can only state with a prescribed degree of confidence that the calculated interval contains the specified proportion of the responses.

Tables A6 and A7 in the appendix provide factors for determining one- and two-sided tolerance intervals, respectively. Factors for upper tolerance limits, $\bar{y} + ks$, are obtained from Table A6. Lower tolerance limits, $\bar{y} - ks$, use the same factors. Two-sided tolerance intervals have the form (2.12), where k is obtained from Table A7. These tolerance factors can be used when the sample mean and standard deviation are calculated from n independent normally distributed observations $y_1, y_2, \ldots, y_n$. The tolerance factor k is determined by selecting the confidence coefficient γ, the proportion p of the population of observations around which the tolerance interval is desired, and the sample size n.

2.6 TESTS OF STATISTICAL HYPOTHESES

Statistical hypotheses are statements about theoretical models or about probability or sampling distributions. In this section we introduce the concepts needed to effectively conduct tests of statistical hypotheses. To clarify the basic principles involved in the testing of statistical hypotheses, we focus attention on tests involving the mean of a normal probability distribution when the population or process standard deviation is known.

There are two hypotheses that must be specified in any statistical testing procedure: the *null* hypothesis and the *alternative* hypothesis (see Exhibit 2.15). In the context of a statistically designed experiment, the null hypothesis, denoted H_0, defines parameter values or other distributional properties that indicate no experimental effect. The alternative hypothesis, denoted H_a, specifies values that indicate change or experimental effect for the parameter or distributional property of interest.

EXHIBIT 2.15 TYPES OF STATISTICAL HYPOTHESES

Null Hypothesis (H_0). Hypothesis of no change or experimental effect.
Alternative hypothesis (H_a). Hypothesis of change or experimental effect.

In the tensile-strength example discussed in Section 2.2, the key parameters of interest might be the mean tensile-strength measurements for all wire samples drawn from each die. Suppose die 1 is a standard die, and dies 2 and 3 are new designs that are being examined. One hypothesis of interest is whether the new dies have mean tensile-strength measurements that differ

from the mean for the standard die. Thus one set of hypotheses of interest in a comparison of the first two dies is

$$H_0: \mu_1 = \mu_2 \quad \text{vs} \quad H_a: \mu_1 \neq \mu_2,$$

where μ_1 denotes the average tensile strength measurement for die 1, and μ_2 is the average for die 2. Note that H_a is a hypothesis of experimental effect; that is, that the mean tensile-strength measurement for wire drawn from die 2 differs from that for the standard die. The hypothesis of no effect (no difference) is the null hypothesis.

Hypotheses can be one-sided if the experimental effect is believed to be one-sided or if interest is only in one-sided effects. For example, one might only be interested in whether the two experimental dies increase the mean of the tensile-strength measurements. If so, the only comparison of interest between the first two dies might be posed as follows:

$$H_0: \mu_1 \geqslant \mu_2 \quad \text{vs} \quad H_a: \mu_1 < \mu_2.$$

Again, it is the alternative hypothesis that designates the hypothesis of experimental effect. The null hypothesis in this case could be written as $H_0: \mu_1 = \mu_2$, but the above statement makes it explicit that only an increase in the mean tensile-strength measurement for die 2 over die 1 constitutes an experimental effect of interest.

The procedures used in statistical hypothesis testing have an interesting analogy with those used in the U.S. judicial system. In a U.S. court of law the defendant is presumed innocent (H_0). The prosecution must prove the defendant guilty (H_a). To do so, the evidence must be such that the innocence of the defendant can be rejected "beyond a reasonable doubt." Thus, the burden of proof is on the prosecution to prove the defendant guilty. Failure to do so will result in a verdict of "not guilty." The defendant does not have to prove innocence; the failure to prove guilt suffices.

Table 2.4 shows the possible true status of the defendant and the possible decisions a judge or jury could make. If the defendant is innocent and the decision is "not guilty," a correct decision is made. Similarly, if the defendant is guilty and the decision is "guilty," a correct decision is made. The other two possibilities result in incorrect decisions.

Table 2.5 displays the corresponding situation for a test of a statistical hypothesis. The decision, based on data collection and analysis, corresponds to the hearing of evidence and judgment by a judge or jury. Again, there are two possible correct decisions and two possible incorrect ones. Note that the two decisions are labeled "do not reject H_0" and "reject H_0." There is no decision labeled "accept H_0," just as there is no judicial decision labeled "innocent."

TABLE 2.4 Relation of Statistical Hypothesis Testing to U.S. Judicial System

		Defendant's True Status	
		Innocent	Guilty
Decision	Not Guilty	Correct Decision	Error
	Guilty	Error	Correct decision

TABLE 2.5 Statistical Hypothesis Testing

		True Hypothesis	
		H_0 True	H_a True
Decision	Do Not Reject H_0	Correct decision	Type II error (probability $= \beta$)
	Reject H_0	Type I error (probability $= \alpha$)	Correct decision

Consider testing the hypotheses

$$H_0: \mu \leqslant 10 \quad \text{vs} \quad H_a: \mu > 10,$$

where μ is the mean of a normal probability distribution with known standard deviation. A sample mean sufficiently greater than 10 would lead one to reject the null hypothesis and accept the alternative hypothesis. How large must the sample mean be to reject the null hypothesis?

Assuming that the null hypothesis is true, we can use the transformation (2.4) to convert the sample mean to a standard normal variate, $z = n^{1/2}(\bar{y} - 10)/\sigma$. We could then decide on the following decision rule:

$$\text{Decision:} \quad \text{Reject } H_0 \text{ if} \quad z > z_\alpha, \tag{2.13}$$

where z_α is a *critical value* from the standard normal table corresponding to an upper-tail probability of α. Note that if the null hypothesis is true, there is only a probability of α that the standard normal variate z will be in the rejection region defined by (2.13). This is the probability of a Type I error in Table 2.5. The Type I error probability is termed the *significance level* of the test. The term *confidence level* is also used. The confidence level, $100(1 - \alpha)\%$, actually denotes the chance of failing to reject the null hypothesis when it is true.

Table 2.6 lists several of the terms in common usage in statistical hypothesis-testing procedures. The significance level of a test is controlled by an experimenter. The significance level is set sufficiently small so that if the test statistic falls in the rejection region, the experimenter is willing to

TABLE 2.6 Terms Used in Statistical Hypothesis Testing

Alternative hypothesis. Hypothesis of experimental effect or change.
Confidence level. Probability of failing to reject H_0 when H_0 is true $(1 - \alpha)$.
Critical value(s). Cutoff value(s) for a test statistic used as limits for the rejection region; determined by the alternative hypothesis and the significance level.
Null hypothesis. Hypothesis of no experimental effect or change.
Operating characteristic curve. Graph of the probability of a Type II error (β) as a function of the hypothetical values of the parameter being tested.
Power. Probability of correctly rejecting H_0 when H_0 is false $(1 - \beta)$; usually unknown.
Rejection (critical) region. Large and/or small values of a test statistic that lead to rejection of H_0.
Significance level. Probability of a Type I error (α); controlled by the experimenter.
Significance probability (p-value). Probability of obtaining a value for a test statistic that is as extreme as or more extreme than the observed value, assuming the null hypothesis is true.
Type I error. Erroneous rejection of H_0; probability $= \alpha$.
Type II error. Erroneous failure to reject H_0; probability $= \beta$.

conclude that the null hypothesis is false. Once the significance level is set, the test procedure is objective: the test statistic is compared with the critical value, and the null hypothesis is either rejected or not, depending on whether the test statistic is in the rejection region.

The significance probability of a test statistic quantifies the degree of discrepancy between the estimated parameter value and its hypothesized value. The significance probability is the probability of obtaining a value of the test statistic that is as extreme as or more extreme than the observed value, assuming the null hypothesis is true. In the above test on the mean of a normal distribution, suppose the standard normal variate had a value of 3.10. Because the alternative hypothesis specifies that large values of the sample mean (and therefore the standard normal variate) lead to rejection of the null hypothesis, the significance probability for this test statistic would be

$$p = \Pr\{Z > 3.10\} = 0.001.$$

Note that if the p-value is smaller than the significance level, the null hypothesis is rejected. Comparison of the p-value with the significance level is equivalent to comparing the test statistic to the critical value.

Common values of the significance level are 0.05 and 0.01, although any small value can be selected for α. The more serious the consequences of a Type I error, the smaller one will choose the significance level. The choice of a significance level will be discussed further in the examples presented in the next several chapters.

In the analogy with the U.S. judicial system, the determination of a rejection region is equivalent to the judge's charge to the jury, especially regarding points of law and requirements needed to find the defendant guilty. The significance level is equivalent to a definition of what constitutes "reasonable doubt." If after an evaluation of the evidence and comparison with the points of law the jurors are confident "beyond a reasonable doubt" that the defendant committed the crime, they are required to find the defendant guilty; otherwise, the defendant is found not guilty.

The significance probability is equivalent to the "strength of conviction" with which the jury finds the defendant guilty. If the p-value is greater than the significance level, the jury does not have sufficient evidence to convict the defendant. If the p-value is much smaller than the significance level, the evidence is overwhelmingly in favor of conviction. If the p-value is only marginally smaller than the significance level, the jury has sufficient evidence for conviction, but the strength of their conviction is less than if the significance probability were very small.

As mentioned above, the term *confidence* is often used in place of significance level or significance probability. It is not the preferred usage when testing hypotheses, but it has become common in some disciplines. The confidence generally quoted for a statistical test is $100(1 - p)\%$. Thus, one might quote statistical significance for the above test "with 99.99% confidence." This use of "confidence" requires caution. It is incorrect to report "with 95% confidence" that the null hypothesis is true. This is equivalent to stating that a jury is "95% confident" that a defendant is innocent. Recall that both in the courtroom and when conducting statistical tests, evidence is collected to show guilt (reject the null hypothesis). Failure to prove guilt (failure to reject the null hypothesis) does not prove innocence (null hypothesis is true) at any "confidence level."

Unlike the significance level, the probability of a Type II error is almost never known to the experimenter. This is because the exact value of the parameter of interest, if it differs from that specified in the null hypothesis, is generally unknown. The probability of a Type II error can only be calculated for specific values of the parameter of interest. The probability of a Type II error, operating characteristic curves, and power are extremely important for assessing the adequacy of an experimental design, especially for determining sample sizes. These considerations are the subject of the next section.

2.7 SAMPLE SIZE AND POWER

Although the probability of a Type II error is almost never known to an experimenter, calculation of these probabilities can be made for hypothetical values

of the parameter of interest. Suppose, for example, one wishes to test the hypotheses

$$H_0: \mu \leqslant 10 \quad \text{vs} \quad H_a: \mu > 10$$

when the population standard deviation is known to equal 2. If a significance level of 0.05 is chosen for this test, the null hypothesis will be rejected if

$$z = \frac{n^{1/2}(\bar{y} - 10)}{2} > z_{0.05} = 1.645.$$

Next suppose that it is critically important to the experimenter that the null hypothesis be rejected if the true population mean is 11 rather than the hypothesized value of 10. If the experimenter wishes to be confident that the null hypothesis will be rejected when the true population mean is 11, the power $(1 - \beta)$ of the test must be acceptably large; equivalently, the probability β of a Type II error must be acceptably small. Figure 2.3 is a typical graph of the sampling distributions of the sample mean for the two cases $\mu = 10$ and $\mu = 11$. The significance level, the probability of a Type II error, and power of the test are indicated.

If $\mu = 11$, then $z = n^{1/2}(\bar{y} - 11)/2$ has a standard normal distribution, while $n^{1/2}(\bar{y} - 10)/2$ is not a standard normal variate. Using the sampling distribution of the mean under the alternative hypothesis, the power of the test can be calculated as follows:

$$1 - \beta = \Pr\left\{\frac{n^{1/2}(\bar{y} - 10)}{2} > 1.645\right\}$$

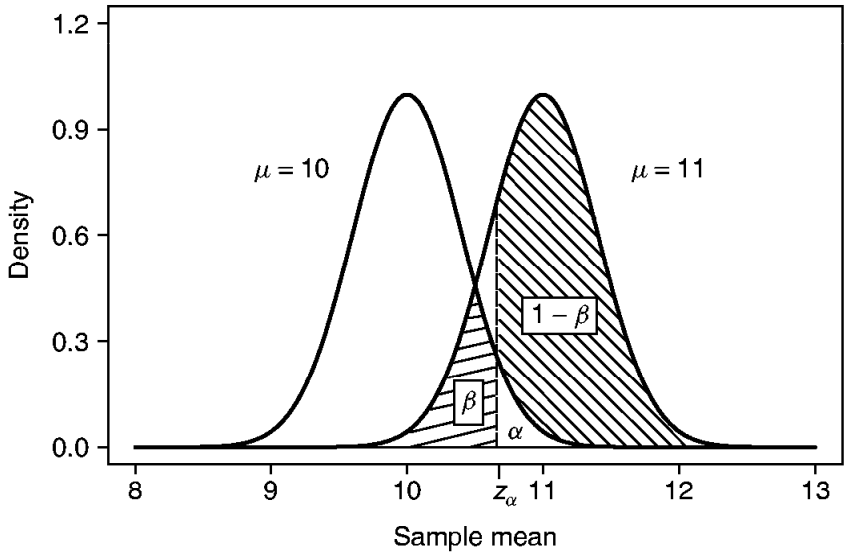


Figure 2.3 Sampling distributions and error probabilities.

$$= \Pr\left\{\frac{n^{1/2}(\bar{y} - 11)}{2} > 1.645 - n^{1/2}\frac{11 - 10}{2}\right\}$$

$$= \Pr\{z > 1.645 - 0.5n^{1/2}\}. \tag{2.14}$$

Once the sample size is determined, the power (2.14) of the above test can be determined. If the experimenter decides to use a sample size of $n = 25$, the power of the test is

$$1 - \beta = \Pr\{z > -0.855\} = 0.804.$$

Thus, the experiment has a probability of 0.804 of correctly rejecting the null hypothesis when the true population mean is $\mu = 11$. If this power is not acceptably large to the experimenter, the sample size must be increased. A sample size of $n = 50$ would produce a power of

$$1 - \beta = \Pr\{z > -1.891\} = 0.971.$$

Figure 2.4 illustrates an analogy that may assist in understanding the connections between the three quantities that determine the properties of a statistical test: sample size, significance level, and power. This schematic drawing shows the reciprocal of the sample size, the confidence level $(1 - \alpha)$ of the test, and the power of the test plotted on trilinear coordinate paper. The scales of the axes in Figure 2.4 are nonlinear and have been omitted for clarity.

Note that if one fixes any two of these quantities, the third one is necessarily determined. So too, if one fixes any one of the components, the other two

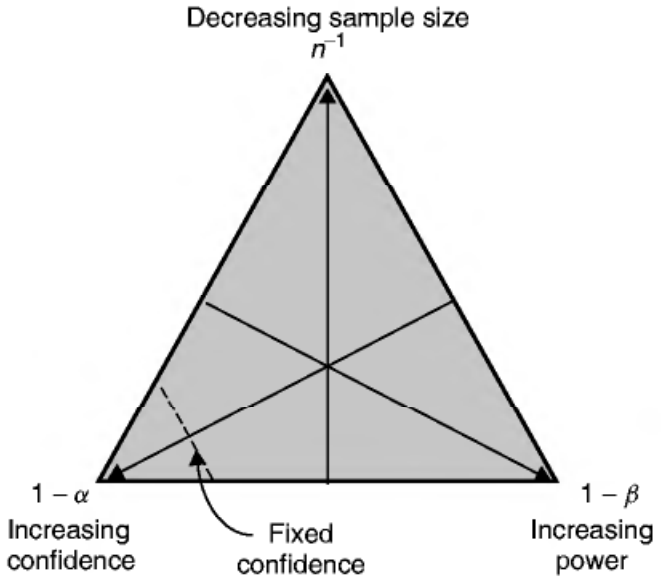


Figure 2.4 Simplex analogy between sample size (n), significance level (α), and power ($1 - \beta$).

components are restricted to a curve in the trilinear coordinate system. Again for simplicity, the curve for fixed confidence is shown in Figure 2.4 as a straight line. The important point to remember is that for any statistical test procedure one can only increase the power for a fixed significance level by increasing the sample size.

A graph of power probabilities can be made as a function of the sample size for a fixed value of the alternative hypothesis or as a function of the values of the parameter of interest for a fixed sample size. Equivalently one can plot the probability of Type II error. The latter type of graph is called an *operating-characteristic (OC) curve*. Power curves and OC curves are extremely valuable for evaluating the utility of an experimental design and for assessing the size of a sample needed to reject the null hypothesis for specified values of the alternative hypothesis.

Returning to the analogy with the judicial system, power refers to the probability of obtaining a conviction when the defendant is guilty. The resources needed (sample size) to assure conviction with a high probability depend on the magnitude of the crime relative to the other factors that might prevent discovery. For example, if a defendant embezzles a very large sum of money from a company, there might not need to be as much effort put into obtaining evidence definitively linking the defendant to the crime than if a much smaller sum of money were embezzled. For a very large sum of money, perhaps only a few people at the company would have access to the funds; therefore, only a few people would need to be investigated.

The foregoing discussion of sample size as related to significance levels and power serves two useful purposes. First, it establishes relationships among the three quantities that are useful for designing experiments. Second, it alerts the experimenter to the need to consider questions relating to the attainable power of a statistical test when sample sizes are fixed and cannot be changed. Many experiments are of that kind. Often the experiment size is predetermined by budget or other constraints. In such settings sample-size considerations are moot. Nevertheless, if great expenditures are being invested in a project, it may be critical to the success of the project to use the procedures described in this section and in Chapter 3 to evaluate the potential for determining the existence of experimental effects of importance.

APPENDIX: PROBABILITY CALCULATIONS

1 Normal Distribution

Table A2 in the appendix to this book lists probabilities for a standard normal probability distribution. The probabilities tabulated are *cumulative* probabilities: for any standard normal variate z, the probability in the body of the table is

$$\Pr\{z \leqslant z_c\},$$

where z_c is a value of that variable. For example,

$$\Pr\{z \leqslant 1.65\} = 0.9505.$$

The entire area under the standard normal distribution is 1.00. Consequently, "upper-tail" probabilities are obtained by subtraction:

$$\Pr\{z > 1.65\} = 1 - 0.9505 = 0.0495.$$

Probabilities for negative standard normal values can be obtained from the corresponding probabilities for positive values. This can be done because the normal distribution is symmetric around the mean; that is the curve to the left of $z = 0$ is the mirror image of the curve to the right of $z = 0$. Thus,

$$\Pr\{z \geqslant -1.65\} = \Pr\{z \leqslant 1.65\} = 0.9505$$

and

$$\Pr\{z < -1.65\} = \Pr\{z > 1.65\} = 0.0495.$$

Also because of the symmetry of the normal distribution, the area under the distribution to the left of $z = 0$ is 0.5000 and the area to the right of $z = 0$ is 0.5000. With this information one can compute probabilities for any range of a standard normal variable. For example,

$$\begin{aligned}
\Pr\{0 \leqslant z \leqslant 2.00\} &= \Pr\{z \leqslant 2.00\} - 0.5000 \\
&= 0.9772 - 0.5000 = 0.4772, \\
\Pr\{-1.00 \leqslant z \leqslant 0\} &= \Pr\{0 \leqslant z \leqslant 1.00\} \\
&= 0.8413 - 0.5000 = 0.3413, \\
\Pr\{0.79 \leqslant z \leqslant 1.47\} &= \Pr\{z \leqslant 1.47\} - \Pr\{z \leqslant 0.79\} \\
&= 0.9292 - 0.7852 = 0.1440.
\end{aligned}$$

2 Student's t-Distribution

Probabilities for Student's t-distribution depend on the degrees of freedom of the statistic much like probabilities for a normal distribution depend on the mean and standard deviation of the normal variable. Unlike the standard normal distribution, there is no transformation of the t-variable that will allow one t-distribution to be used for all t-variables.

The degrees of freedom for a t-variable are essentially an adjustment to the sample size based on the number of distributional parameters that must be estimated. Consider a t statistic of the form

$$t = \frac{n^{1/2}(\bar{y} - \mu)}{s},$$

where the sample mean and the sample standard deviation are estimated from n independent observations from a normal probability distribution. If the standard deviation were known, a standard normal variable, equation (2.4), could be used instead of this t-statistic by inserting σ for s. The standard normal distribution can be shown to be a t-distribution with an infinite number of degrees of freedom. The above t-statistic has $v = n - 1$ degrees of freedom, based on the stated assumptions. So the number of degrees of freedom, $v = n - 1$ instead of $v = \infty$, represents an adjustment to the sample size for having to estimate the standard deviation.

Table A3 in the appendix lists cumulative t-probabilities for several t-distributions. The values in the table are t-values that give the cumulative probabilities listed at the top of each column. Thus, for a t-variable with 10 degrees of freedom,

$$\Pr\{t \leqslant 1.812\} = 0.95.$$

If, as is generally the case, a computed t-value does not exactly equal one of the table values, the exact probability is reported as being in an interval bounded by the probabilities in the table. For example, for a t-variable with eight degrees of freedom $\Pr\{t \leqslant 2.11\}$ is not available from the table. Note that 2.11 lies between the table values 1.860 and 2.306. Denoting the desired probability by p, one would report

$$0.950 < p < 0.975.$$

Similarly, if one wishes to compute the "upper-tail" probability $\Pr\{t > 3.02\}$ for a t-variable having 18 degrees of freedom, one would report

$$p < 0.005.$$

Probabilities for negative t-values are obtained from Table A3 much the same way as are normal probabilities. The t-distribution is symmetric around $t = 0$. Thus, for a t-variable with five degrees of freedom

$$\Pr\{t \geqslant -2.015\} = \Pr\{t \leqslant 2.015\} = 0.95$$

and

$$\Pr\{t < -3.365\} = \Pr\{t > 3.365\} = 0.01.$$

Intervals for nontabulated t-values would be reported similarly to the above illustrations.

3 Chi-Square Distribution

Chi-square distributions depend on degrees of freedom in much the same way as t-distributions. Table A4 in the appendix lists cumulative and upper-tail probabilities for several chi-square distributions, each having a different number of degrees of freedom. Chi-square variates can only take nonnegative values, so there is no need to compute probabilities for negative chi-square values. Use of Table A4 for chi-square values follows the same procedures as was outlined above for nonnegative t-values.

4 F-Distribution

F-statistics are ratios of two independent chi-squares, each divided by the number of its degrees of freedom. There are two sets of degrees of freedom for an F variable, one corresponding to the numerator statistic (v_1) and one corresponding to the denominator statistic (v_2).

Table A5 in the appendix contains F-values and their corresponding cumulative and upper-tail probabilities. Due to the need to present tables for a large number of possible numerator and denominator degrees of freedom. Table A5 only contains F-values corresponding to cumulative probabilities of 0.75, 0.90, 0.95, 0.975, 0.99, and 0.995. The F-statistic is nonnegative, so the use of Table A5 is similar to the use of Table A4 for the chi-square distribution. Critical F-values for lower-tail cumulative probabilities of 0.25, 0.10, 0.05, 0.025, 0.01, and 0.005 can be obtained from these tables by using the following relationship:

$$F_{\alpha}(v_1, v_2) = 1/F_{1-\alpha}(v_2, v_1).$$

REFERENCES

Text References

Most textbooks on statistical methods include discussions of summary statistics and graphical displays. Traditional descriptive statistics (mean, median, standard deviation, quartiles) receive the most extensive treatment. A selected list of textbooks intended for scientific and engineering audiences is

Bethea, R. M. and Rhinehart, R. R. (1991). *Applied Engineering Statistics*, New York: Marcel Dekker, Inc.

Grant, E. L. and Leavenworth, R. S. (1988). *Statistical Quality Control*, Sixth Edition. New York: McGraw-Hill Book Co.

One of many texts on statistical quality control that makes extensive use of descriptive statistics and graphics for control charts.

Montgomery, D. C. and Runger, G. C. (2002). *Applied Statistics and Probability for Engineers*, Third Edition, New York: John Wiley & Sons, Inc.

Ostle, B., Turner, K., Hicks, C., and McElrath, G. (1996). *Engineering Statistics: The Industrial Experience*, Belmont, CA: Duxbury Press.

The following mathematical statistics texts provide the theory behind many of the topics introduced in this chapter. These texts are intended for scientific and engineering audiences.

Bethea, R. M., Duran, B. S., and Boullion, T. L. (1995). *Statistical Methods for Engineers and Scientists*, Third Edition, New York: Marcel Dekker, Inc.

Johnson, R. A. (1994). *Probability and Statistics for Engineers*, Fifth Edition, Englewood Cliffs, NJ: Prentice-Hall, Inc.

Hines, W. W. and Montgomery, D. C. (1990). *Probability and Statistics in Engineering and Management*, Third Edition, New York: John Wiley & Sons, Inc.

Hogg, R. V. and Ledolter, J. (1992). *Applied Statistics for Engineers and Scientists*, Second Edition, Englewood Cliffs, NJ: Prentice-Hall, Inc.

Mendenhall, W. and Sincich, T. (1995). *Statistics for Engineering and the Sciences*, Fourth Edition, Englewood Cliffs, NJ: Prentice-Hall, Inc.

Montgomery, D. C., Runger, G. C., and Hubele, N. F. (1998). *Engineering Statistics*, New York: John Wiley & Sons, Inc.

Scheaffer, R. and McClave, J. T. (1995). *Probability and Statistics for Engineers*, Fourth Edition, Belmont, CA: Duxbury Press.

Vardeman, S. B. (1994). *Statistics for Engineering Problem Solving*, Boston: PWS Publishers.

Walpole, R. E., Myers, R. H., Myers, S. L., and Yee, K. (2002). *Probability and Statistics for Engineers and Scientists*, Seventh Edition, Englewood Cliffs, NJ: Prentice-Hall, Inc.

Each of the above texts discusses confidence intervals and tests of statistical hypotheses. The treatment is more theoretical than in this book, and most of the texts provide detailed derivations. As specific confidence-interval and hypothesis-testing procedures are covered in the following chapters, references to more applied texts will be provided.

The following two books also discuss tolerance intervals. The first of these books contains extensive tables of tolerance factors.

Odeh, R. E. and Owen, D. B. (1980). *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*. New York: Marcel Dekker, Inc.

Ostle, B. and Malone, L. C. (1988). *Statistics in Research*, Fourth Edition. Ames, Iowa: Iowa State University Press.

Data References

The fuel-economy data discussed in this chapter are taken from

Hare, C. T. (1985). "Study of the Effects of Fuel Composition, Injection, and Combustion System Type and Adjustment on Exhaust Emissions from Light-Duty Diesels,"

The Coordinating Research Council, Inc., CRC-APRAC Project No. CAPE-32–80, San Antonio, TX: Southwest Research Institute.

EXERCISES

1 The diameters (mm) of 25 randomly selected piston rings are given below. Are the observed differences in the values of these diameters due to location, dispersion, or both? What is the apparent shape of the distribution?

76.7	78.7	74.5	78.9	79.6	74.4	79.7	75.4	78.7
79.5	75.1	79.9	75.3	79.8	76.8	79.6	76.0	74.2
79.0	75.9	79.0	73.8	79.2	75.9	79.2		

2 Algebraically show the equivalence between formulas (a) and (b) for the sample standard deviation in Exhibit 2.4.

3 A solar energy system was designed and constructed for a North Carolina textile company. The system was used to preheat boiler feedwater before injection in the process steam system. The energy delivered to the feedwater was observed for a 48-hour period. The values below represent the energy rate (kbtu/hr):

493	500	507	500	501	489	495	508	490	495	511	498
490	507	488	499	509	499	494	490	490	489	515	493
505	497	490	507	497	492	503	495	513	495	492	492
510	501	530	504	501	491	504	507	496	492	496	511

Calculate the following descriptive statistics for the observed energy rates:

(a) Average.
(b) Sample standard deviation.
(c) Minimum and maximum.
(d) Range.
(e) Median.
(f) Quartiles — first, second, and third.
(g) SIQR.

Which is a better measure of the center of these energy rates, the average or the median? Why? What interval of energy rates includes most typical values, as represented by the energy rates in this sample?

4 A study was conducted to examine the explosibility of M-1 propellant dust and to determine the minimum energy of electrostatic discharge needed to induce an explosion for dust concentrations of 0.4 g/liter. The ignition energies (joules) were observed for a series of eighteen experiments:

0.23	0.30	0.35	0.33	0.64	0.36
0.27	0.20	0.23	0.31	0.22	0.21
0.16	0.24	0.22	0.27	0.20	0.25

Calculate the descriptive statistics listed in Exercise 3 for these data. Why do the median and average differ from one another? Which of the measures of center do you prefer for this data set? Why?

5 Comment on the variation of ignition energy data presented in Exercise 4. How do the range and the sample standard deviation compare in describing the dispersion of the data?

6 Listed below are red-blood-cell counts for patients with various liver diseases. The patients are categorized according to their initial diagnosis upon entering the hospital. Which disease group averages are most similar? Which disease group standard deviations are most similar. Considering both sets of statistics, if high red-blood-cell counts are desirable, which liver disease appears to be most serious? Why?

RED BLOOD CELL COUNTS

Cirrhosis	**Hepatitis**	**Tumor**	**Other**
18	14	3	14
66	17	1	36
18	10	3	4
7	5	6	13
15	27	4	4
6	11	7	5
4	30		3
8	18		6
5	4		11
28	39		6
4	25		33
49			
33			

7 An automobile manufacturer is comparing the failure times (in hours) of two different makes of alternators. The data are listed below for five units of each type. Compute and compare descriptive statistics for each group of data. What conclusion can you make about the lifetime of the two makes of alternators? Comment on both location and dispersion measures.

Make	Time to Failure (hrs)				
Manufacturer A	131	127	110	76	40
Manufacturer B	65	51	18	12	5

8 Use the normal probability tables to determine the following probabilities for a standard normal response variable z:

(a) $\Pr\{z > 1.98\}$

(b) $\Pr\{z < 3\}$

(c) $\Pr\{-2.5 < z < -1.2\}$

(d) $\Pr\{1.6 < z < 3.0\}$

(e) $\Pr\{z < 0\}$

(f) $\Pr\{z > 2\}$

(g) $\Pr\{z < 2\}$

(h) $\Pr\{z < 1.64\}$

(i) $\Pr\{z > 1.96\}$

(j) $\Pr\{z < 2.58\}$

9 For the following t and chi-square (X) variates, determine the probabilities indicated. In each case use $n - 1$ as the number of degrees of freedom.

(a) $\Pr\{t < 2.467\}, n = 29$

(b) $\Pr\{t > 2.074\}, n = 23$

(c) $\Pr\{t < -1.65\}, n = 12$

(d) $\Pr\{t > 3.011\}, n = 30$

(e) $\Pr\{t > 1.5\}, n = 7$

(f) $\Pr\{X < 53.5\}, n = 79$

(g) $\Pr\{X > 85.7\}, n = 56$

(h) $\Pr\{X < 16.2\}, n = 32$

(i) $\Pr\{X > 27.1\}, n = 15$

(j) $\Pr\{X > 42.8\}, n = 6$

10 For the following F-variates, determine the probabilities indicated. Use the numbers of degrees of freedom shown.

(a) $\Pr\{F < 4.19\}, \upsilon_1 = 3, \upsilon_2 = 4.$

(b) $\Pr\{F > 4.39\}, \upsilon_1 = 9, \upsilon_2 = 12.$

(c) $\Pr\{F < 3.50\}, \upsilon_1 = 3, \upsilon_2 = 7.$

11 A manufacturer of thin polyester film used to fabricate scientific balloons reports the standard deviation of a run of film to be 0.01 mm. A sample of nine strips of the film was taken, and the thickness of each strip was measured. The average thickness was found to be 0.5 mm. Construct a 98% confidence interval for the mean thickness of this type of film. Interpret

the confidence interval in the context of this exercise. What assumptions did you make to construct the interval?

12 Suppose the mean film thickness in the previous exercise is known to be 0.46 mm and the standard deviation is known to be 0.01 mm. Determine natural process limits for the film thickness that will include 99% of all film-thickness measurements. Interpret the process limits in the context of this exercise.

13 A container manufacturer wishes to ensure that cartons manufactured for a furniture-moving company have sufficient strength to protect the contents during shipment. Twenty-five cartons are randomly selected from the manufacturer's large inventory, and the crushing strength of each is measured. The average crushing strength is 126.2 psi and the standard deviation is calculated to be 5.0 psi. Construct a 95% tolerance interval for the crushing strengths of these cartons that will include 99% of the individual carton measurements. What assumptions are you making to construct this interval?

14 Suppose in the previous exercise that the desired minimum crushing strength is 100 psi for an individual carton. What can you conclude from the statistical tolerance interval calculated in Exercise 13 about the ability of the manufacturer's cartons to meet this engineering specification limit?

15 It has been determined that the distribution of lengths of nails produced by a particular machine can be well represented by a normal probability distribution. A random sample of ten nails produced the following lengths (in inches):

$$1.14, 1.15, 1.11, 1.16, 1.13, 1.15, 1.18, 1.12, 1.15, \ 1.12.$$

Calculate a 90% tolerance interval that includes 95% of the nail lengths produced by this machine.

16 A rocket was designed by an aerospace company to carry an expensive payload of scientific equipment on a satellite to be launched in orbit around the earth. On the launch date, scientists debated whether to launch due to possible adverse weather conditions. In the context of a test of a statistical hypothesis, the relevant hypotheses might be posed as follows:

H_0: the satellite will launch successfully,

H_a: the satellite will not launch successfully.

Describe the Type I and the Type II errors for these hypotheses. Which error appears to be the more serious? Why? How, in the context of this exercise, would you decide whether to launch to make the Type I error acceptably small? What about the Type II error?

17 In Exercise 11, suppose interest was in testing whether the mean film thickness is less than 0.48 mm versus the alternative that it is greater than 0.48. Using a significance level of 0.01, conduct an appropriate test of this hypothesis. Draw a conclusion in the context of this exercise; that is, interpret the results of the statistical test.

18 Construct OC curves and power curves for the test in the previous exercise. In each case, draw separate curves for samples of size $n = 10$, 15, 20, and 25. Use 0.48 mm as the hypothesized mean and 0.10 as the standard deviation. Use a significance level of 0.01. Evaluate the Type II error probability and the power for hypothesized mean values in an interval from 0.48 to 0.55. Suppose one wishes to ensure that a mean of 0.53 will be detected with a probability of at least 0.90. What is the minimum sample size that will suffice?

19 A manufacturer of ball bearings collects a random sample of 11 ball bearings made during a day's production shift. The engineering specifications on the bearings are that they should have diameters that are 3.575 ± 0.001 mm. Do the data below indicate that the diameters of the ball bearings are meeting this specification?

3.573	3.571	3.575	3.576	3.570	3.580
3.577	3.572	3.571	3.578	3.579	

20 In Exercise 6, suppose it is of interest to test whether the mean red-blood-cell count for patients with hepatitis is greater than 25 versus the alternative that it is less than 25. Using a significance level of 0.05, conduct an appropriate test of hypothesis, draw a conclusion, and interpret the results of the statistical test.

CHAPTER 3

Inferences on Means and Standard Deviations

In this chapter we present statistical techniques for drawing inferences on means and standard deviations. The construction of confidence intervals and procedures for performing tests of statistical hypotheses involving parameters from one, two, or several populations or stable processes are discussed. Sample-size determination is also detailed. Specific topics covered include:

- *single-sample inferences on a mean or a standard deviation,*
- *comparisons of two means or two standard deviations using either independent or paired samples, and*
- *comparisons of several standard deviations when independent samples are available to estimate the individual standard deviations.*

The basic concepts associated with confidence intervals and tests of statistical hypotheses were introduced in Chapter 2. The purpose of this chapter is to expand this framework to include specific inference procedures for common experimental situations that require inferences involving means and standard deviations from one or more populations or processes.

Most of the techniques used in this chapter are derived under the assumption that the response variables are statistically independent and follow a normal probability distribution. The normality assumption is not, however, critical for most of these inference procedures. As discussed in Section 2.3, the central limit property can be used to justify the use of the normal distribution as an approximate sampling distribution for the sample mean. The t-distribution is known to be especially robust to departures from normality when inferences are desired on a mean. The randomization of test runs can also be used to justify these inference procedures.

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

An informative graphical display that visually highlights the location and spread of a set of data and that is often highly suggestive of the need for inferential comparisons of means and standard deviations is a *boxplot*. Boxplots designate the middle half of a data set by a rectangle whose lower and upper edges (or right and left edges, if drawn horizontally), respectively, are the first (Q_1) and third (Q_3) quartiles of the data set. The median is indicated by a line segment drawn within the box parallel to the edges. Thus, the box provides a visual impression that emphasizes the middle half of the data values and the line drawn at the median provides an explicit quantification of the center of the data. If desired, the average can also be indicated by a plotting symbol, for example, x. In its simplest version, the width of the box is arbitrary and is left to the discretion of the data analyst. If several boxes are plotted in the same graph, differences in sample sizes are sometimes depicted by making the widths of the boxes proportional to the square roots of the sample sizes.

Variability of the data is portrayed by the lengths of vertical lines, often dashed, drawn from the edges of the boxes to upper and lower *adjacent values* calculated from the semi-interquartile range (SIQR). The upper adjacent value is the largest observation in the data set that does not exceed $Q_3 + 3 \times \text{SIQR}$, whereas the lower adjacent value is the smallest data value that is no less than $Q_1 - 3 \times \text{SIQR}$. Any data values larger or smaller, respectively, than the upper and lower adjacent values are plotted individually because they identify possible *outliers*, extreme data values whose magnitudes are unusual relative to the bulk of the data. We summarize the construction of boxplots in Exhibit 3.1 and show a schematic boxplot in Figure 3.1.

EXHIBIT 3.1 BOXPLOTS

1. Calculate the quartiles, the SIQR, and the average of the data values.
2. Draw a box having
   - **(a)** lower and upper edges at the first (Q_1) and third (Q_3) quartiles, respectively,
   - **(b)** a convenient width (optional: proportional to the square root of the sample size),
   - **(c)** a line, parallel to the edges, at the median,
   - **(d)** a symbol (e.g., x) at the average (optional).
3. Extend line segments (often dashed) from each edge of the box to the most extreme data values that are no farther than $3 \times \text{SIQR}$ from each edge.
4. Individually plot all observations that are more extreme than the adjacent values.

Figure 3.2 displays comparative boxplots for two data sets previously introduced in Chapter 2. Figure 3.2a contains boxplots for the tensile strength measurements listed in Table 2.2 and presented in a point plot in Figure 2.2. The comparisons in Figure 3.2a strikingly emphasize the location difference

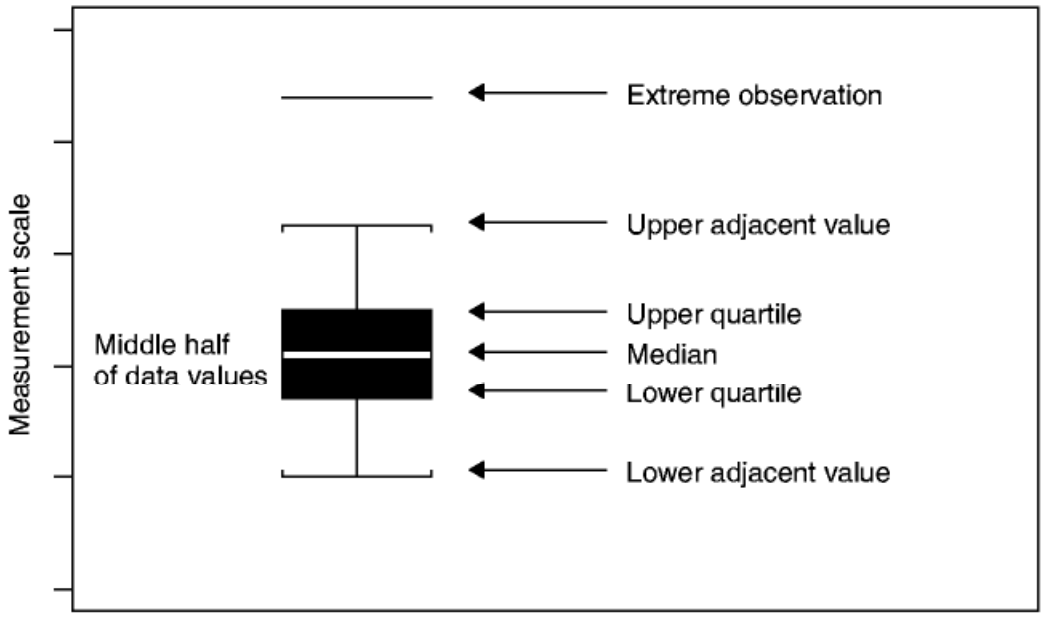


Figure 3.1 Schematic boxplot.

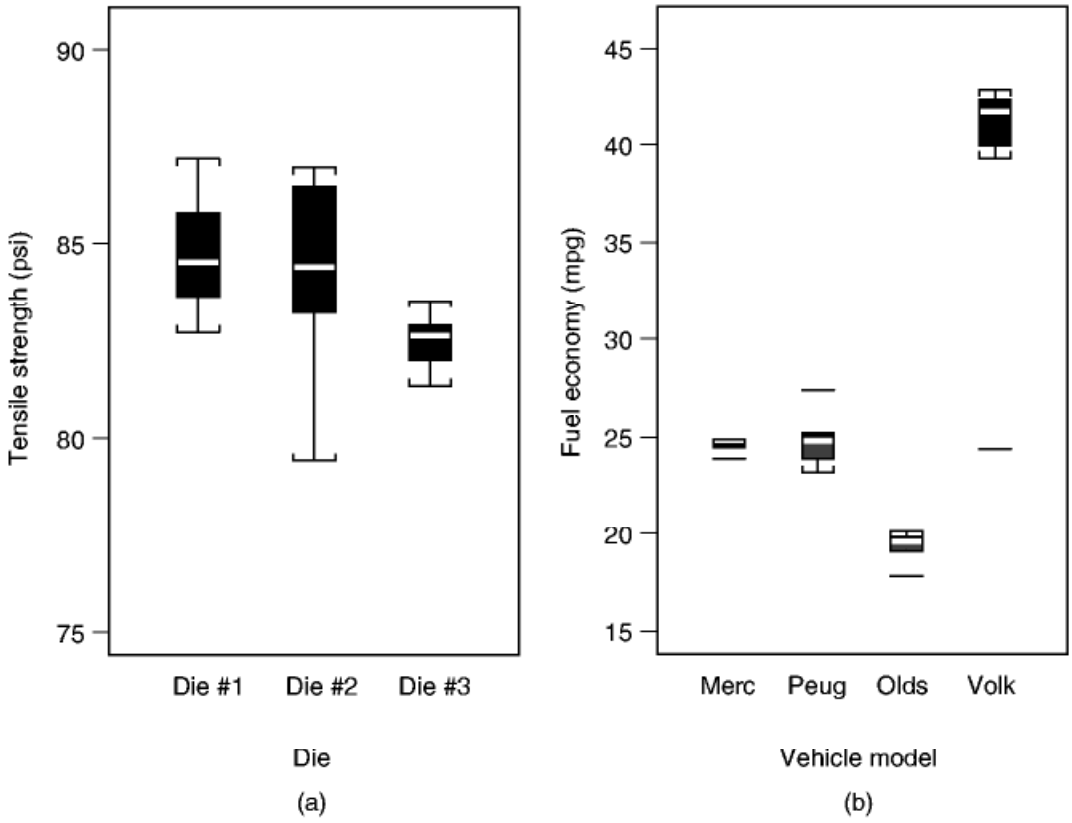


Figure 3.2 Boxplots of tensile strength and fuel economy measurements. (a) Tensile strength. (b) Fuel economy.

and the smaller variation for the measurements from Die 3. These location and spread characteristics are highlighted at the expense of some loss of detail from the point plot. Often it is precisely location and spread issues that are central to an investigation. Boxplots provide uncluttered clarity in making such comparisons.

Figure 3.2b shows comparative boxplots for the fuel economy data listed in Table 2.1 and displayed graphically in Figure 2.1. Once again, the location and variability differences in the measurements are strikingly evident. So too is the extremely low fuel economy measurement for the Volkswagen, as are one measurement for each of the other vehicles. Ordinarily one would want more than eight observations from which to form a boxplot; however, we present this plot so its features can be compared with the summary statistics in Table 2.1. Another reason is to point out that as long as interpretations are drawn with due recognition of the sampling procedure and the sample size, boxplots can be visually compelling even with small sample sizes.

The information displayed in the boxplots in Figure 3.2 suggest a number of comparisons of location and spread that would be informative. Using the probability-based inference procedures introduced in the last chapter, such comparisons are possible in spite of the relatively small sample sizes in these examples. The remaining sections of this chapter provide procedures for drawing inferences on means and standard deviations. Inferences on means and standard deviations in a more general experimental setting than is covered in this chapter are provided in Parts II and III of this text.

3.1 INFERENCES ON A POPULATION OR PROCESS MEAN

When independent observations are obtained from one process or population in an experiment, one can often model the responses as

$$y_i = \mu + e_i, \qquad i = 1, 2, \ldots, n. \tag{3.1}$$

In this representation y_i is a measurement on a continuous variate, μ is the unknown mean of the process or population, and e_i is a random error component associated with the variation in the observations. These errors are assumed to be statistically independent and to have a common probability distribution possessing a mean of zero and a constant but unknown standard deviation of σ. In the tensile-strength experiment, y_i represents a tensile-strength measurement on one of the dies, μ is the population mean tensile strength for all wire produced from that die, and e_i is the difference between the observed tensile strength and the true (unknown) average tensile strength μ.

The sample mean or average $\bar{y}$ is an estimator of the unknown constant μ in equation (3.1). Under an assumption that the errors are statistically independent

and normally distributed, the response variables are independent and normally distributed with mean μ and standard deviation σ. The sampling distribution of the sample mean is then normal with mean μ and standard error $\sigma/n^{1/2}$ (see Section 2.3). If σ is known, interval-estimation and hypothesis-testing procedures using standard normal variates can be applied. These procedures were discussed in Sections 2.4 and 2.6, respectively. In the remainder of this section, we discuss the relevant procedures for the more common situation in which the standard deviation is unknown.

3.1.1 Confidence Intervals

Suppose it is of interest to construct a confidence interval for the model mean in equation (3.1) when the standard deviation σ is unknown and the response variables y_i can be considered normally distributed. The average of the sample values $y_1, y_2, \ldots, y_n$ is denoted by $\bar{y}$, and the sample standard deviation by s.

From the distributional results presented in Section 2.3,

$$t = \frac{\bar{y} - \mu}{s/n^{1/2}} = \frac{n^{1/2}(\bar{y} - \mu)}{s} \tag{3.2}$$

follows a t-distribution with $n - 1$ degrees of freedom. In this expression, $s_{\bar{y}} = s/n^{1/2}$ is the estimated standard error of the sample mean $\bar{y}$. The degrees of freedom, $n - 1$, for this t-statistic are associated with the estimation of the standard deviation (or variance).

A t-statistic is the ratio of two statistically independent quantities. The numerator is a standard normal variate. The denominator is the square root of a chi-square variate divided by its number of degrees of freedom; that is, the general form of a t-statistic is

$$t = \frac{z}{(X^2/v)^{1/2}}, \tag{3.3}$$

where z is a standard normal variate and X^2 is a chi-square variate having v degrees of freedom. The degrees of freedom of a t statistic equal those of the chi-square statistic from which it is formed.

As mentioned in Section 2.3, $z = n^{1/2}(\bar{y} - \mu)/\sigma$ is a standard normal variate. The sample mean, and hence z, is statistically independent of the sample variance s^2. The variate $X^2 = (n - 1)s^2/\sigma^2$ follows a chi-square distribution with $v = n - 1$ degrees of freedom. Inserting these expressions for z and X^2 into (3.3) results in the t-variate (3.2).

Under the normality and independence assumptions stated above, the degrees of freedom for the chi-square variate $X^2 = (n - 1)s^2/\sigma^2$ are $n - 1$. The degrees of freedom indicate how many statistically independent responses, or in some instances functions of the responses, are used to calculate the variate.

In the calculation of the sample variance not all n of the differences $y_i - \bar{y}$ are statistically independent, because they sum to zero. Knowing any $n - 1$ of them and the constraint that all n sum to zero enables one to determine the last one: the last one has the same magnitude as the sum of the $n - 1$ known differences, but opposite sign.

In many instances, as is the case for the sample variance, the number of degrees of freedom equals the number of independent observations less the number of additional parameters that must be estimated to calculate the statistic. The mean μ must be estimated (by $\bar{y}$) to calculate the sample variance (and hence) the sample standard deviation. If μ were known, the sample variance could be estimated as

$$s^2 = \sum \frac{(y_i - \mu)^2}{n}$$

and $X^2 = ns^2/\sigma^2$ would follow a chi-square distribution with $v = n$ degrees of freedom. The t-statistic obtained by inserting this chi-square variate in (3.3) would then also have n degrees of freedom.

Following the procedure outlined in Section 2.4, a $100(1 - \alpha)\%$ confidence interval for μ can be derived by starting with the probability statement

$$\Pr\left\{-t_{\alpha/2} < \frac{n^{1/2}(\bar{y} - \mu)}{s} < t_{\alpha/2}\right\} = 1 - \alpha. \tag{3.4}$$

where $t_{\alpha/2}$ is the critical value from Table A3 in the appendix for a t-variate having $v = n - 1$ degrees of freedom and an upper-tail probability of $\alpha/2$. The confidence interval is formed by algebraically manipulating the above probability statement to isolate the mean between the inequalities. The resulting confidence interval is

$$\bar{y} - t_{\alpha/2}s_{\bar{y}} < \mu < \bar{y} + t_{\alpha/2}s_{\bar{y}},$$

or

$$\bar{y} \pm t_{\alpha/2}s_{\bar{y}}.$$

The probability statement (3.4) states that the bounds computed from the sample will cover the true value of the mean with probability $1 - \alpha$. Once the sample is collected and the bounds are computed, we state that we are $100(1 - \alpha)\%$ confident that the confidence limits do cover the unknown mean.

A 95% confidence interval for the average tensile strength of wire produced from the first die using the 18 tensile-strength measurements in Table 2.2 is

$$84.733 - 2.110\frac{1.408}{(18)^{1/2}} < \mu < 84.733 + 2.110\frac{1.408}{(18)^{1/2}},$$

or

$$84.033 < \mu < 85.433.$$

As stressed in Chapter 2, the confidence limits 84.033 and 85.433 should not be used to imply that there is a 0.95 probability that μ is in the interval defined by these limits. The probability statement (3.4) is valid only prior to the collection and analysis of the data. Because μ is a fixed constant, it either is or is not in the computed confidence interval. The confidence coefficient $100(1-\alpha)\%$ refers to the statistical methodology; that is, the methodology used will result in confidence limits that include the true parameter value with a probability of $1-\alpha$.

To emphasize this point further, Figure 3.3 schematically depicts confidence intervals from several samples laid out on the same scale as the measurement of interest and its associated normal probability distribution. Suppose the true mean is 85. Each time a sample is drawn and a confidence interval is constructed, the calculated interval either includes $\mu = 85$ or the interval fails to

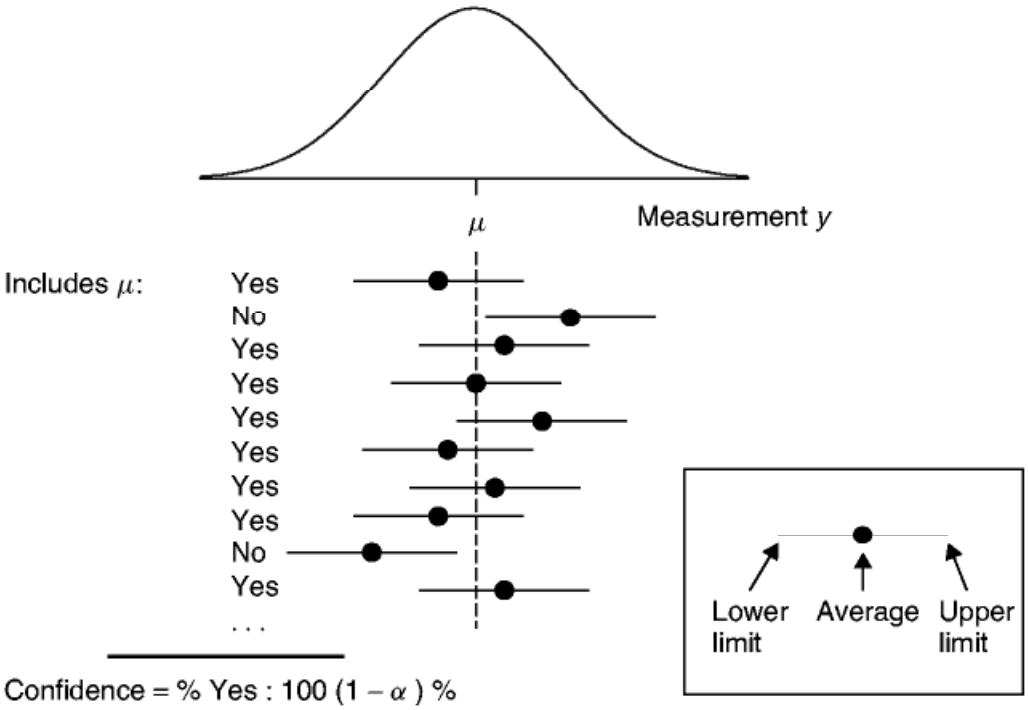


Figure 3.3 Confidence intervals for the mean of a normal population.

TABLE 3.1 Confidence Intervals for Means of Normal Probability Distributions*

σ Known	σ Unknown
(a) *Two-Sided*	
$\bar{y} - z_{\alpha/2}\sigma_{\bar{y}} < \mu < \bar{y} + z_{\alpha/2}\sigma_{\bar{y}}$	$\bar{y} - t_{\alpha/2}s_{\bar{y}} < \mu < \bar{y} + t_{\alpha/2}s_{\bar{y}}$
(b) *One-Sided Upper*	
$\mu < \bar{y} + z_{\alpha}\sigma_{\bar{y}}$	$\mu < \bar{y} + t_{\alpha}s_{\bar{y}}$
(c) *One-Sided Lower*	
$\bar{y} - z_{\alpha}\sigma_{\bar{y}} < \mu$	$\bar{y} - t_{\alpha}s_{\bar{y}} < \mu$

*$\sigma_{\bar{y}} = \sigma/n^{1/2}$, $s_{\bar{y}} = s/n^{1/2}$, z_α = standard normal critical value, $t = t$ critical value, $v = n - 1$.

include it. If one repeatedly draws samples of size n from this normal distribution and constructs a confidence interval with each sample, then approximately $100(1 - \alpha)\%$ of the intervals will contain $\mu = 85$ and $100\alpha\%$ will not.

In practice we do not take many samples; usually only one is drawn. The interval estimate obtained from this sample either brackets μ or it does not. We do not know if the interval does include μ, but our confidence rests with the procedure used: $100(1 - \alpha)\%$ of the time the mean will be included in an interval constructed in this manner. Thus, we state that we are $100(1 - \alpha)\%$ confident that the one interval we have computed does indeed include the mean.

It occasionally is of interest to an experimenter to construct a one-sided rather than a two-sided confidence interval. For example, in the tensile-strength experiment, one may only be concerned with a lower bound on the average tensile strength. One could then begin with a one-sided probability statement [using the upper limit in (3.4)] and derive the following one-sided lower confidence interval for μ:

$$\bar{y} - t_{\alpha}s_{\bar{y}} < \mu,$$

where t_α is used in place of $t_{\alpha/2}$ for one-sided intervals. Formulas for these and other one-sided and two-sided confidence intervals are given in Table 3.1.

3.1.2 Hypothesis Tests

Hypotheses testing, as introduced in Chapter 2, consists of the four basic steps given in Exhibit 3.2.

One of the most straightforward statistical tests involves determining if a mean μ differs from some hypothesized value, say μ_0. This is symbolized by

$$H_0: \mu = \mu_0 \quad \text{vs} \quad H_a: \mu \neq \mu_0.$$

EXHIBIT 3.2 STATISTICAL HYPOTHESIS TESTING

1. State the null and alternative hypotheses.
2. Draw a sample and calculate the appropriate test statistic.
3. Compare the calculated value of the test statistic to the critical value(s) corresponding to the significance level selected for the test; equivalently, compare the significance probability of the test statistic to the significance level selected for the test.
4. Draw the appropriate conclusion and interpret the results.

The appropriate test statistic for a single sample of independent observations is

$$t = \frac{\bar{y} - \mu_0}{s_{\bar{y}}}.$$

The decision rule used for these hypotheses is:

$$\text{Decision:} \quad \text{Reject } H_0 \text{ if } \quad t < -t_{\alpha/2}$$
$$\text{or if } \quad t > t_{\alpha/2},$$

where α is the chosen significance level of the test.

Consider again the tensile-strength experiment. Suppose one wishes to test

$$H_0: \mu = 84 \quad \text{vs} \quad H_a: \mu \neq 84.$$

From the summary statistics in Table 2.3, the test statistic is

$$t = \frac{84.733 - 84}{1.408/(18)^{1/2}} = 2.209.$$

If one chooses a significance level of $\alpha = 0.05$, the decision would be to reject H_0, because $t = 2.209$ exceeds $t_{0.025}(17) = 2.110$. Equivalently, since the two-tailed significance probability is $0.02 < p < 0.05$ (twice the upper-tail probability limits), which is less than the significance level selected for the test, the null hypothesis is rejected.

In stating the conclusion to a hypothesis test, some experimenters elect to state the significance level of the test, while others present the p-value associated with the test statistic. A third approach, which is less commonly used, involves relating either the significance level or the p-value to a confidence level. Hence, in the above example, one could state 95% confidence in the decision to reject the null hypothesis, because $1 - \alpha = 0.95$. One might also report that one's confidence is between 95% and 98%, because $0.95 < 1 - p < 0.98$. The dangers of having such statements misinterpreted were

explained in Section 2.6. We recommend that this terminology be used with caution and only when there is a clear explanation of the meaning.

Because hypothesis testing and the construction of confidence intervals involve the same statistics and the same probability statements, there is a unique correspondence between the two. If a computed confidence interval excludes a hypothesized value of the parameter of interest, a test of the hypothesis that the parameter equals the hypothesized value will be rejected. Similarly, if the interval includes the hypothesized parameter value, the hypothesis would not be rejected by the corresponding statistical test. For example, a two-sided 95% confidence interval for the mean of the tensile-strength measurements from die 1 was given above as (84.033, 85.433). Because this interval does contain $\mu = 85$, a test of the hypothesis $H_0: \mu = 85$ vs $H_a: \mu \neq 85$ would not be rejected. On the other hand, a test of $H_0: \mu = 84$ vs $H_a: \mu \neq 84$ would be rejected at the $\alpha = 0.05$ significance level.

Rules for testing hypotheses about the mean of a normal population using a single sample of independent observations are summarized in Table 3.2. The procedures outlined for large samples are identical to those described in Section 2.6 for models in which the standard deviations are known. The equivalence of these procedures is based on the near-equality of critical values for the normal distribution and Student's t distribution when the degrees of freedom of the latter are sufficiently many, say more than 30 (cf. Tables A2 and A3).

3.1.3 Choice of a Confidence Interval or a Test

In the previous subsection, the equivalence between confidence-interval procedures and procedures for testing statistical hypotheses was illustrated.

TABLE 3.2 Decision Rules and p-Value Calculations for Tests on the Mean of a Normal Distribution*

	Small Sample	Known σ or Large Sample
(a) $H_0: \mu = \mu_0$ vs $H_a: \mu \neq \mu_0$		
Reject H_0 if:	$\lvert t_c \rvert > t_{\alpha/2}$	$\lvert z_c \rvert > z_{\alpha/2}$
p-value:	$p = 2\Pr\{t > \lvert t_c \rvert\}$	$p = 2\Pr\{z > \lvert z_c \rvert\}$
(b) $H_0: \mu \geqslant \mu_0$ vs $H_a: \mu < \mu_0$		
Reject H_0 if:	$t_c < -t_\alpha$	$z_c < -z_\alpha$
p-value:	$p = \Pr\{t < t_c\}$	$p = \Pr\{z < z_c\}$
(c) $H_0: \mu \leqslant \mu_0$ vs $H_a: \mu > \mu_0$		
Reject H_0 if:	$t_c > t_\alpha$	$z_c > z_\alpha$
p-value:	$p = \Pr\{t > t_c\}$	$p = \Pr\{z > z_c\}$

*$t = n^{1/2}(\bar{y} - \mu_0)/s$; $z = n^{1/2}(\bar{y} - \mu_0)/\sigma$; t_c, z_c = calculated values of test statistics; z_α = standard normal critical value; $t_\alpha = t$ critical value; $v = n - 1$.

The equivalence between confidence-interval and statistical-testing procedures is a general one that can be shown to hold for the parameters of many probability distributions. Because of this equivalence, either could be used to perform a test on the stated parameter. There are circumstances, however, when one may be preferable to the other.

In general, a statistical test simply allows one to conclude whether a null hypothesis should be rejected. The p-value often provides a degree of assurance that the decision is the correct one given the observed data and the assumptions, but fundamentally a statistical test is simply a reject–no-reject decision.

When both can be used to draw inferences on a parameter of interest, confidence intervals provide more information than that afforded by a statistical test. Not only does a confidence interval allow one to assess whether a hypothesis about a parameter should be rejected, but it also provides information on plausible values of the parameter. In particular, a tight confidence interval, one whose upper and lower bounds are sufficiently close to the estimate of the parameter, implies that the parameter has been estimated with a high degree of precision. If the confidence interval is wide, the experimenter knows that the current data do not provide a precise estimate of the parameter. Either of these conclusions might cast doubt on the result of a statistical test. As input to a decision-making process, either could be more important than the result of a statistical test in the determination of a course of action to be taken.

Of what value, then, is a statistical test? Statistical testing procedures allow an assessment of the risks of making incorrect decisions. Specification of a significance level requires consideration of the consequences of a Type I error. Power curves and OC curves (see Sections 2.7 and 3.1.4) describe the risk of a Type II error in terms of possible values of the parameter of interest, the significance level of the test, and the sample size.

Confidence intervals and statistical testing procedures supplement one another. As with any statistical methodology, either of these procedures can be used inappropriately. In some settings it would be inappropriate, for example, to test a hypothesis about a mean and ignore whether it is estimated with satisfactory precision. In others, confidence-interval estimation may be inadequate without the knowledge of the risk of a Type II error. In general, it is recommended that statistical tests be accompanied by interval estimates of parameters. This philosophy is stressed throughout the remainder of this book, although for clarity of presentation many examples include only interval estimation or testing results.

3.1.4 Sample Size

An important consideration in the design of many experiments is the determination of the number of sample observations that need to be obtained. In

the estimation of parameters, sample sizes are selected to ensure satisfactory precision. In hypothesis testing, the sample size is chosen to control the value of β, the probability of a Type II error (see Section 2.7).

The sample size can be chosen to provide satisfactory precision in the estimation of parameters if some knowledge is available about the variability of responses. For example, in the estimation of the mean when responses are normally distributed and statistically independent, the length of the confidence interval is [see (2.10)]:

$$2z_{\alpha/2}\sigma/n^{1/2}. \tag{3.5}$$

If one wishes to have the length of this confidence interval be no larger than L, say, the required sample size is obtainable by setting (3.5) equal to L and solving for n:

$$n = \left(\frac{2z_{\alpha/2}\sigma}{L}\right)^2. \tag{3.6}$$

When the standard deviation σ is not known, several approximate techniques can be used. In some instances, a worst-case sample size can be determined by inserting a reasonable upper bound on σ into (3.6). If data are available from which to estimate the standard deviation, the sample estimate can be inserted into (3.6) if it is deemed to be sufficiently precise. If an estimate is available but it is not considered sufficiently precise, an upper limit from a one-sided confidence interval on σ (see Section 3.2) can be used to provide a conservative sample-size determination.

In the last section a 95% confidence interval for the mean tensile strength for wire produced by die 1 was found to be (84.033, 85.433). The length of this confidence interval is 1.400. Suppose this is not considered precise enough and in future investigations of the tensile strength of wire from dies such as this it is desired that the confidence intervals have a length of approximately 1 mm; that is, the confidence interval is desired to be approximately $\bar{y} \pm 0.5$ mm.

From an analysis of the summary statistics in Table 2.3 and perhaps analyses of other tensile-strength data, suppose the experimenters are willing to use $\sigma = 2$ mm as a conservative value for the standard deviation of wire samples made from dies similar to these three. If a 95% confidence interval is to be formed, the required sample size is, from (3.6),

$$n = \left(\frac{(2)(1.96)(2)}{(1)}\right)^2 = 61.47,$$

or approximately 62 observations.

To select sample sizes for hypothesis-testing procedures, use is made of operating-characteristic (OC) curves. OC curves are plots of β against values of the parameter of interest for several different sample sizes. Equivalently,

power curves (plots of $1 - \beta$) can be graphed versus the sample size. Separate sets of curves usually are provided for different values of α, the probability of a Type I error.

3.2 INFERENCES ON A POPULATION OR PROCESS STANDARD DEVIATION

Populations and processes are not uniquely characterized by location parameters such as the mean. In most settings the variability of measurements or observations is just as important as the mean measurement or the mean of the observations. Consider, for example, any manufacturing process where the manufactured product must meet customer specifications. It is not sufficient to claim that the average of a large number of measurements on a characteristic is equal to or sufficiently close to a specification or target — the customer wants all the measurements to be on target or sufficiently close to the target. Thus, at a minimum, the average must be sufficiently close to target and a measure of the variability of the measured characteristic must be sufficiently small. Together, these two statements imply that all the measurements of the characteristic are sufficiently close to the target value.

The variability of response variables is characterized in statistical models by the inclusion of one or more model terms denoting uncontrollable variation. This variation may be due to numerous small influences that are either unknown or unmeasurable. For example, the observable variability in measurements of the output of an electronic circuit may be due to small voltage fluctuations, minor electrical interference from other power sources, or a myriad of other possible influences.

The error components of statistical models are not fixed constants. They are assumed to be random variables that follow a probability distribution, frequently the normal distribution. The assumption that model errors follow a probability distribution is an explicit recognition of the uncontrolled nature of response variability. This assumption enables the response variability to be characterized (e.g., through the shape of the distribution) and its magnitude to be specified through one or more of the model parameters (e.g., the standard deviation). A goal of many experiments is to draw inferences on the variability of the model errors.

When the normal probability distribution is a reasonable assumption for model errors, the variability in the distribution is entirely determined by the standard deviation σ or its square, the variance σ^2. This section presents inference procedures for population or process standard deviations. These procedures are based on sampling distributions for the sample variance. Although these techniques can be used to make inferences on either variances or standard deviations, our emphasis is on the standard deviation because it is a more

natural measure of variation than the variance, in part because it is measured in the same units as the response variable.

Inference procedures, such as confidence intervals and hypothesis tests for standard deviations, are derived and interpreted similar to those for means that were presented in the previous section. For this reason our coverage of these topics for standard deviations is more concise than that given in the last section. Sample sizes for confidence intervals and statistical tests on standard deviations are derived as illustrated for means in Section 3.1.4. Sample size determination is not a main focal point of this text, so the interested reader is referred to the references at the end of this chapter and Chapter 2 for further details on this topic.

3.2.1 Confidence Intervals

Inferences on a population standard deviation from a single sample of statistically independent observations are made using the sample variance:

$$s^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2.$$

The sample standard deviation, first discussed in Section 2.1, is the positive square root of the sample variance. Note that because both the sample variance and the sample standard deviation are always nonnegative, inferences on the sample standard deviation can be made by making inferences on the sample variance. For example, a test of the hypothesis that $\sigma = 2$, say, can be made by testing that $\sigma^2 = 4$.

In Section 2.3 we noted that the variate

$$X^2 = (n-1)s^2/\sigma^2 \tag{3.7}$$

has a chi-square probability distribution with $n - 1$ degrees of freedom. The chi-square distribution for the statistic (3.7) is critically dependent on the independence and normality assumptions. To obtain an interval estimate of σ we begin with the probability statement

$$\Pr\{X^2_{1-\alpha/2} < (n-1)s^2/\sigma^2 < X^2_{\alpha/2}\} = 1 - \alpha,$$

or, after some algebraic rearrangement,

$$\Pr\{(n-1)s^2/X^2_{\alpha/2} < \sigma^2 < (n-1)s^2/X^2_{1-\alpha/2}\} = 1 - \alpha, \tag{3.8}$$

where $X^2_{1-\alpha/2}$ and $X^2_{\alpha/2}$ are, respectively, lower and upper critical points for a chi-square variate having $v = n - 1$ degrees of freedom and a probability of

$\alpha/2$ in each tail of the distribution. The latter probability statement provides the basis for the following $100(1-\alpha)\%$ two-sided confidence interval for σ^2:

$$(n-1)s^2/X^2_{\alpha/2} < \sigma^2 < (n-1)s^2/X^2_{1-\alpha/2}. \tag{3.9}$$

Taking the square root of each side of the inequality (3.9) provides a $100(1-\alpha)\%$ confidence interval for the standard deviation:

$$\left(\frac{(n-1)s^2}{X^2_{\alpha/2}}\right)^{1/2} < \sigma < \left(\frac{(n-1)s^2}{X^2_{1-\alpha/2}}\right)^{1/2}. \tag{3.10}$$

The confidence interval (3.9) for σ and the corresponding one-sided intervals are exhibited in Table 3.3.

Using the tensile-strength data for die 1 given in Table 2.2, the sample variance is $s^2 = 1.982$ and the sample standard deviation is $s = (1.982)^{1/2} = 1.408$. A 95% confidence interval for the variance is

$$\frac{(17)(1.982)}{30.19} < \sigma^2 < \frac{(17)(1.982)}{7.56},$$

or

$$1.116 < \sigma^2 < 4.457,$$

TABLE 3.3 Confidence Intervals for Standard Deviations: Normal Probability Distribution*

(a) *Two-Sided*

$$\left(\frac{(n-1)s^2}{X^2_{\alpha/2}}\right)^{1/2} < \sigma < \left(\frac{(n-1)s^2}{X^2_{1-\alpha/2}}\right)^{1/2}$$

(b) *One-Sided Upper*

$$\sigma < \left(\frac{(n-1)s^2}{X^2_{1-\alpha}}\right)^{1/2}$$

(c) *One-Sided Lower*

$$\left(\frac{(n-1)s^2}{X^2_{\alpha}}\right)^{1/2} < \sigma$$

*X^2_α = chi-square critical value, $v = n - 1$.

where 7.56 and 30.19 are the $X^2_{0.975}$ and $X^2_{0.025}$ values, respectively, based on $v = 17$ degrees of freedom. From these confidence limits, we obtain the limits for the standard deviation:

$$1.056 < \sigma < 2.111$$

Confidence intervals for the standard deviations of the measurements from the other two dies can be calculated in a similar manner. The estimates of the standard deviations for the measurements from the two dies are s = 2.054 for Die 2 and s = 0.586 for Die 3. The respective confidence intervals are $1.541 < \sigma < 3.080$ for Die 2 and $0.440 < \sigma < 0.879$ for Die 3. Note that the estimates and the confidence intervals quantify the visual impressions left by Figures 2.2 and 3.2a: Die 3 is much less variable than the other two dies, and Die 2 appears to be more variable than Die 1. Note too that the first conclusion is fairly conclusive because the confidence interval for Die 3 does not overlap those for Dies 1 and 2, while the confidence intervals for the latter two dies do overlap. Explicit pairwise comparisons of the variabilities of the dies will be made in Section 3.4.

3.2.2 Hypothesis Tests

To test a hypothesis of the form

$$H_0: \sigma^2 = \sigma_0^2 \quad \text{vs} \quad H_a: \sigma^2 \neq \sigma_0^2$$

where σ_0^2 is a hypothesized value of σ^2, use is once again made of the variate (3.7) but with the hypothesized value of the population variance inserted:

$$X^2 = (n-1)s^2/\sigma_0^2.$$

Using this statistic, the decision rule can be formulated as follows:

$$\begin{aligned} \text{Decision:} \quad \text{Reject } H_0 \text{ if} \quad & X^2 < X^2_{1-\alpha/2} \\ \text{or if} \quad & X^2 > X^2_{\alpha/2} \end{aligned}$$

where α is the chosen significance level of the test.

This test procedure is also used to test the corresponding hypotheses on the population standard deviation. The test procedure provides exact Type I and Type II error probabilities because $\sigma = \sigma_0$ if and only if $\sigma^2 = \sigma_0^2$.

Consider again the tensile-strength experiment. Suppose we wish to test the hypothesis

$$H_0: \sigma = 2 \quad \text{vs} \quad H_a: \sigma \neq 2.$$

The equivalent test for the population variance is

$$H_0: \sigma^2 = 4 \quad \text{vs} \quad H_a: \sigma^2 \neq 4.$$

The test statistic for the measurements from Die 1 is

$$X^2 = (17)(1.982)/4 = 8.424 \qquad (0.05 < p < 0.10).$$

If we select a significance level of $\alpha = 0.05$, the null hypothesis is not rejected, because X^2 is neither greater than $X^2_{0.025} = 30.19$ nor less than $X^2_{0.975} = 7.56$.

As discussed in Sections 3.1.2 and 3.1.3, there is an equivalence between many interval estimation techniques and tests of hypotheses. The test just conducted could have been performed by examining whether the hypothesized value of the standard deviation, $\sigma = 2$, is within the limits of the confidence interval for σ. If it is, the hypothesis is not rejected; otherwise, it is rejected. Because $\sigma = 2$ is in the computed interval (1.056, 2.111), the hypothesis that $\sigma = 2$ is not rejected.

Table 3.4 contains various one-sided and two-sided tests for standard deviations. Included are both decision rules and methods for calculating the appropriate significance probabilities.

TABLE 3.4 Decision Rules and *p*-value Calculations for Tests on the Standard Deviations of a Normal Distribution*

(a) $H_0: \sigma = \sigma_0$ vs $H_a: \sigma \neq \sigma_0$

Reject H_0 if:	$X_c^2 < X^2_{1-\alpha/2}$ or $X_c^2 > X^2_{\alpha/2}$
p-Value:	$p = 2\min(p_l, p_u)$
	$p_l = \Pr\{X^2 < X_c^2\}$
	$p_u = \Pr\{X^2 > X_c^2\}$

(b) $H_0: \sigma \geqslant \sigma_0$ vs $H_a: \sigma < \sigma_0$

Reject H_0 if:	$X_c^2 < X^2_{1-\alpha}$
p-Value:	$p = \Pr\{X^2 < X_c^2\}$

(c) $H_0: \sigma \leqslant \sigma_0$ vs $H_a: \sigma > \sigma_0$

Reject H_0 if:	$X_c^2 > X_\alpha^2$
p-Value:	$p = \Pr\{X^2 > X_c^2\}$

*$X^2 = (n-1)s^2/\sigma_0^2$, X_c^2 = calculated value of test statistic, X_α^2 = chi-square critical value, $v = n - 1$.

3.3 INFERENCES ON TWO POPULATIONS OR PROCESSES USING INDEPENDENT PAIRS OF CORRELATED DATA VALUES

Comparing population or process means or standard deviations using data from two samples requires consideration of whether the responses are collected on pairs of experimental units (or under similar experimental conditions), or whether the responses from the two samples are mutually independent. This section contains a discussion of the analysis of paired samples, that is, independent pairs of data values, where the two data values in a pair are not statistically independent. This setting is a special case of a randomized complete block design (Chapter 9).

The data in Table 3.5 contain pairs of measurements on the percentage of solids in a fixed volume of fluid flowing through a pipeline. The pairs of measurements are taken at two port locations along the pipeline. It is believed that measurements in different samples are statistically independent but the two measurements in each sample are not, because of the proximity of the ports. The investigators collected these data to compare the mean solids measurements from the two ports. A secondary concern is whether the variability of the measurements at the two ports is comparable.

When two populations (or processes) are being sampled in an experiment, one can often model the responses as

$$y_{ij} = \mu_i + e_{ij}, \qquad i = 1, 2, \quad j = 1, 2, \ldots, n, \tag{3.11}$$

TABLE 3.5 Percentage Solids by Volume at Two Pipeline Locations

	Solids (%)		
Sample Number	Port A	Port B	Difference $d = B - A$
1	1.03	1.98	0.95
2	0.82	1.71	0.89
3	0.77	1.63	0.86
4	0.74	1.39	0.65
5	0.75	1.33	0.58
6	0.74	1.45	0.71
7	0.73	1.39	0.66
8	0.66	1.24	0.58
9	0.65	1.30	0.65
Average	0.77	1.49	0.73
S.D.	0.11	0.24	0.14

where y_{ij} is the jth measurement taken from the ith population, μ_i is the unknown mean of the ith population, and e_{ij} is a random error component associated with the variation in the measurements. If the observations are paired, differences in the respective pairs of observations in model (3.11) can be expressed as

$$\begin{aligned} d_j = y_{1j} - y_{2j} &= (\mu_1 - \mu_2) + (e_{1j} - e_{2j}) \\ &= \mu_d + e_j, \end{aligned} \tag{3.12}$$

where $\mu_d = \mu_1 - \mu_2$ and $e_j = e_{1j} - e_{2j}$.

The model (3.12) for the differences in the pairs of observations looks very similar to the model (3.1) for a sample of independent observations from a single population. In fact, the two models are identical. The differences d_j are statistically independent because the observations on different pairs ($j = 1, 2, \ldots, n$) are assumed to be independent. If the original responses are normally distributed, the differences d_j can be shown to be independently normally distributed with common mean $\mu_d = \mu_1 - \mu_2$ and a common standard deviation σ_d. This result is valid regardless of whether the standard deviations for the two populations are equal.

Because the differences d_j can be modeled as a sample of independent observations from a normal population, the confidence-interval, hypothesis-testing, and sample-size determination procedures presented in the last section can be directly applied to drawing inferences on the mean difference μ_d. One simply replaces $\bar{y}$ and $s_{\bar{y}}$ with $\bar{d}$ and $s_{\bar{d}}$, where

$$\bar{d} = \frac{1}{n}\sum d_i = \bar{y}_1 - \bar{y}_2, \qquad s_{\bar{d}} = \frac{s_d}{n^{1/2}},$$

and

$$s_d = \left(\frac{1}{n-1}\sum (d_i - \bar{d})^2\right)^{1/2}. \tag{3.13}$$

Using the summary statistics in Table 3.5 on the differences, a 95% confidence interval on the mean difference $\mu_d = \mu_B - \mu_A$ is

$$0.635 < \mu_d < 0.817.$$

Because this interval does not include the value zero, one can infer from the confidence interval that a statistical test of the equality of the means would be rejected. Indeed, using the t-statistic (3.2) with the appropriate substitutions for $\bar{d}$ and $s_{\bar{d}}$, a test of $H_0: \mu_d = 0$ vs $H_a: \mu_d \neq 0$ is rejected, since $t = 15.71 (p < 0.001)$.

Figure 3.4(a) is a scatterplot of the measurements on the percentage solids at the two ports for the nine samples. It is clear from the difference in the

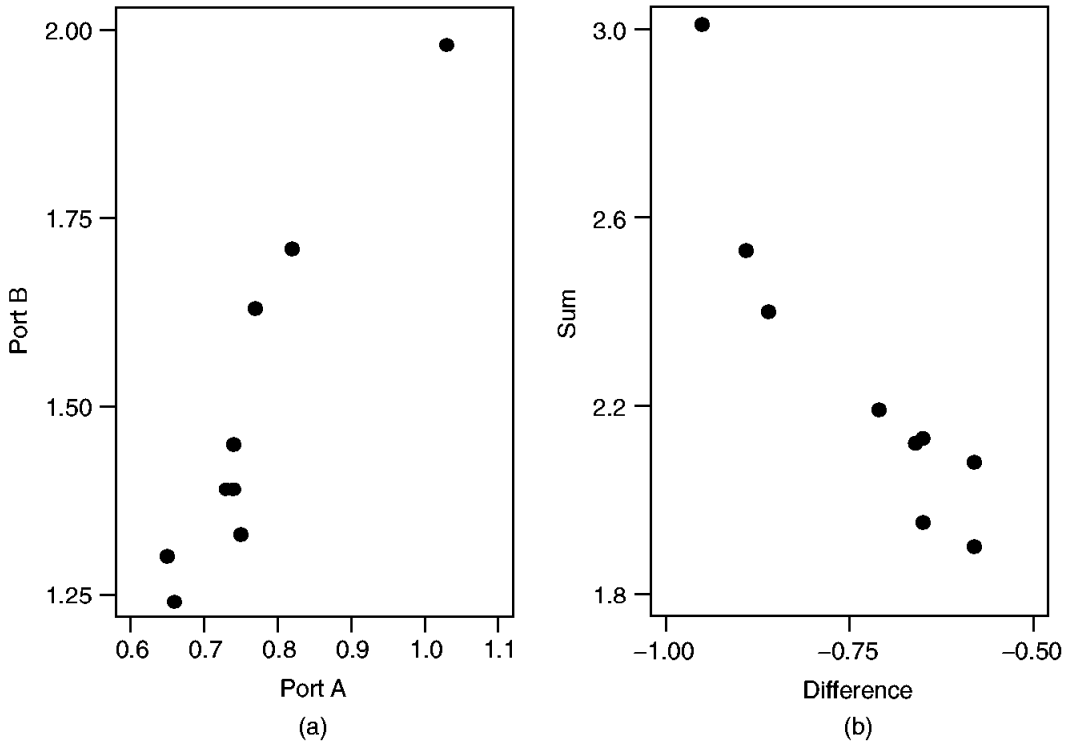


Figure 3.4 Percentage solids at two pipeline locations. (a) Percentage solids. (b) Differences and sums.

scales on the two axes in the scatterplot that the two ports do not have the same measurement. Port B typically has larger measurements than Port A, as confirmed by the test on the means that was just made. More difficult to observe from the plot is a sense of whether the measurements from the two ports are equally variable. While the scales on the axes differ, they cover approximately the same range: 0.5% for Port A and 0.75% for Port B. It is not clear whether these two ranges are sufficiently different for the variation to be declared different at the two ports.

Complicating any comparison of the variability for the two ports is the correlation between pairs of measurements. While differences between the two measurements can be used to compare the respective means, as shown above, differences cannot be used by themselves to compare the standard deviations. In addition the F statistic introduced in Section 2.3 [cf., equation (2.7)] and discussed further in the next section cannot be used to compare the standard deviations because the two sample standard deviations are not statistically independent. Fortunately, an alternative statistical analysis is available based on the concept of linear correlation.

Linear correlation is explained in detail in Chapter 14. In essence, two sets of measurements are linearly correlated if they increase or decrease together along a straight line. Figure 3.4(a) suggests that the measurements from the

two ports do increase along an approximate straight line. The linear correlation (r) between the port measurements is $r = 0.938$, which is very large considering the maximum value a linear correlation can attain is 1.0.

For pairs of measurements that are either statistically independent or that are linearly correlated, the sums and differences of the pairs are themselves linearly correlated if the standard deviations are not equal. If the standard deviations of the pairs of measurements are equal, the sums and differences are not linearly correlated. Thus, testing for statistical significance of a linear correlation between sums and differences is simultaneously a test for the equality of the standard deviations (Bradley and Blackwood 1989).

Testing the significance of a linear correlation is detailed in Chapter 14. Using the calculations presented in that chapter, one finds that the linear correlation between the sums and differences of the solids measurements at the two ports is $r = -0.923$. Figure 3.4 (b) confirms that the sums and differences are closely approximated by a straight line with a negative slope, as suggested by the large negative correlation (the lower limit on a correlation is -1.0). Calculating the t statistic for testing that the correlation is 0 yields $t = r(n-2)^{1/2}/(1-r^2)^{1/2} = -6.346$ (two-tailed p-value: $p = 0.0004$). Hence, one concludes that the standard deviations of the measurements from the two ports are not equal. As shown in Table 3.5, the estimated standard deviation of the measurements from Port B is approximately twice as large as that from Port A.

3.4 INFERENCES ON TWO POPULATIONS OR PROCESSES USING DATA FROM INDEPENDENT SAMPLES

Measurements of percent solids similar to those shown in Table 3.5 could be regarded as statistically independent if the measurements were taken differently from the description provided in the last section. For example, measurements taken at sufficiently different time intervals from the two ports, especially if the combined measurement sequence was randomized, would be regarded as statistically independent. Alternatively, measurements taken from two port locations that are not in close proximity could be regarded as statistically independent.

If two sets of measurements are statistically independent, the procedures of the previous section are not appropriate. This is because there is no physical or logical way to form pairs of observations. The sample numbers are simply arbitrary labels. One can still model the responses as in equation (3.11), but the errors, and, hence, the responses, are now mutually independent. Inferences on means based on independent samples from two populations are a special case of inferences using one-factor linear models from completely randomized experimental designs (see Chapter 6).

Denote observations from two samples as $y_{11}, y_{12}, \ldots, y_{1n_1}$ and $y_{21}, y_{22}, \ldots, y_{2n_2}$, respectively. Denote the respective sample means by $\bar{y}_1$ and $\bar{y}_2$ and the sample standard deviations by s_1 and s_2. Our discussion of inferences on the mean difference from two independent samples distinguishes two cases, equal and unequal population standard deviations. In either case, the sample sizes are not required to be equal.

Consider first the case where the population (or process) standard deviations are assumed to be equal. Inference techniques for this assumption are discussed below.

When the population standard deviations are equal, inferences on the difference in the population means are made using the following statistic:

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{[s_p^2(1/n_1 + 1/n_2)]^{1/2}}. \tag{3.14}$$

The t statistic in (3.14) uses a *pooled* estimate of the common standard deviation:

$$s_p = \left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right)^{1/2} \tag{3.15}$$

$$= \left(\frac{s_1^2 + s_2^2}{2}\right)^{1/2} \quad \text{if} \quad n_1 = n_2.$$

The denominator of the two-sample t-statistic (3.14) is the estimated standard error of the difference of the sample means:

$$\text{SE}(\bar{y}_1 - \bar{y}_2) = s_p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}.$$

The t-statistic (3.14) follows a Student's t-distribution with $v = n_1 + n_2 - 2$ degrees of freedom. Note that the number of degrees of freedom equals the sum of the numbers of degrees of freedom for the two sample variances. Using this t-statistic, the confidence interval and hypothesis testing procedures discussed in Section 3.1 can be applied to inferences on the difference of the two means.

If one wishes to obtain a confidence interval for the difference in the mean tensile strengths of wires produced from the first two dies using the sample measurements in Table 2.1, the two-sample t-statistic can be used if one can reasonably make the assumption that the variability of tensile-strength measurements is about equal for wires from the two dies. If so, the pooled

estimate of the standard deviation, using the summary statistics in Table 2.1 and equation (3.15), is

$$s_p = \left(\frac{1.408 + 2.054}{2}\right)^{1/2} = 1.761.$$

A 95% confidence interval for the difference between the mean tensile strengths for dies 1 and 2 is then

$$(84.733 - 84.492) - (2.038)(1.761)(\tfrac{1}{18} + \tfrac{1}{18})^{1/2} < \mu_1 - \mu_2$$
$$< (84.733 - 84.492) + (2.038)(1.761)(\tfrac{1}{18} + \tfrac{1}{18})^{1/2},$$

or

$$-0.955 < \mu_1 - \mu_2 < 1.437,$$

where 2.038 is the critical value for a Student t-variate having 34 degrees of freedom (linearly interpolated between $v = 30$ and $v = 60$ in Table A3 of the appendix) corresponding to an upper-tail probability of $\alpha/2 = 0.025$. Because this confidence interval includes zero, the means would not be judged significantly different by a statistical test of $H_0: \mu_1 = \mu_2$ vs. $H_a: \mu_1 \neq \mu_2$.

Suppose now that $\sigma_1 \neq \sigma_2$ as one might infer from the estimated standard deviations and the confidence intervals for the Die 1 and Die 2 standard deviations that were calculated in Section 3.2.1. In this situation one should not use the pooled estimate of the variance, s_p^2. Rather than using the t-statistic (3.14), we use the following statistic which approximately follows a Student's t-distribution for normally distributed response variables:

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{(s_1^2/n_1 + s_2^2/n_2)^{1/2}}. \tag{3.16}$$

The number of degrees of freedom for this statistic is taken to be the (rounded) value of

$$v = \frac{(w_1 + w_2)^2}{w_1^2/(n_1 - 1) + w_2^2/(n_2 - 1)}, \tag{3.17}$$

where $w_1 = s_1^2/n_1$ and $w_2 = s_2^2/n_2$. Tables 3.6 and 3.7 summarize confidence interval and hypothesis-testing procedures for the difference of two means when independent (unpaired) observations from both populations or processes are available.

The analysis of the standard deviations from two independent samples often precedes a two-sample t-test on the equality of two means. An inference that the two population or process standard deviations are not significantly different from each other would lead one to use the two-sample t-statistic (3.14). A

TABLE 3.6 Confidence Intervals for the Difference in Means of Two Normal Populations: Independent (Unpaired) Samples

(a) *Two-Sided*

$$(\bar{y}_1 - \bar{y}_2) - t_{\alpha/2}\text{SE} < \mu_1 - \mu_2 < (\bar{y}_1 - \bar{y}_2) + t_{\alpha/2}\text{SE}$$

(b) *One-Sided Upper*

$$\mu_1 - \mu_2 < (\bar{y}_1 - \bar{y}_2) + t_{\alpha}\text{SE}$$

(c) *One-Sided Lower*

$$(\bar{y}_1 - \bar{y}_2) - t_{\alpha}\text{SE} < \mu_1 - \mu_2$$

$$\sigma_1 = \sigma_2$$

$$v = n_1 + n_2 - 2, \qquad \text{SE} = s_p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}, \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{v}$$

$$\sigma_1 \neq \sigma_2$$

$$v = \frac{(w_1 + w_2)^2}{w_1^2/(n_1 - 1) + w_2^2/(n_2 - 1)}, \quad \text{SE} = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^{1/2}, \quad w_j = \frac{s_j^2}{n_j}$$

TABLE 3.7 Decision Rules and *p*-Value Calculations for Tests on the Difference of Means of Two Normal Populations: Independent (Unpaired) Samples*

(a) $H_0: \mu_d = \mu_0$ vs $H_a: \mu_d \neq \mu_0$

Reject H_0 if:	$\lvert t_c \rvert > t_{\alpha/2}$
p-Value:	$p = 2\Pr\{t > \lvert t_c \rvert\}$

(b) $H_0: \mu_d \geqslant \mu_0$ vs $H_a: \mu_d < \mu_0$

Reject H_0 if:	$t_c < -t_{\alpha}(v)$
p-Value:	$p = \Pr\{t < t_c\}$

(c) $H_0: \mu_d \leqslant \mu_0$ vs $H_a: \mu_d > \mu_0$

Reject H_0 if:	$t_c > t_{\alpha}$
p-Value:	$p = \Pr\{t > t_c\}$

*$\mu_d = \mu_1 - \mu_2$, $t = [(\bar{y}_1 - \bar{y}_2) - \mu_0]/SE$, t_c = calculated value of t; v and SE are defined in Table 3.6.

conclusion that the standard deviations are significantly different would lead one to use the two-sample approximate t-statistic (3.16). A comparison of two standard deviations is also often performed when one is interested in comparing the variability between two processes. For example, one may desire to compare the variability of the results received from two different laboratories, or from two different types of machinery.

When testing the equality of two standard deviations from normal populations the following F-variate is used:

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}, \tag{3.18}$$

where s_1^2 is the sample variance of a random sample of size n_1 from a normal distribution with variance σ_1^2, and s_2^2 is the sample variance of a random sample of size n_2 from a normal distribution with variance σ_2^2. This F-statistic follows an F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom. Both the assumption of independent observations between and within the samples and the assumption of normal distributions are critical assumptions for the validity of the F-distribution for this variate. It is important, therefore, that the normality assumption be examined using the procedures recommended in Section 18.2 before this variate is used to compare standard deviations.

Comparisons of standard deviations are often made through the use of the corresponding variances, for the same reasons as cited in the previous section. Using the F-variate (3.18) and its corresponding probability distribution, the following two-sided confidence interval for the ratio of population variances is obtained:

$$\frac{s_1^2}{s_2^2 F_{\alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2 F_{1-\alpha/2}}, \tag{3.19}$$

where $F_{1-\alpha/2}$ and $F_{\alpha/2}$ are lower and upper critical points from an F-distribution having $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom and a probability of $\alpha/2$ in each tail of the distribution. The upper critical value $F_{\alpha/2}$ can be read directly from Table A5 in the appendix. The lower critical value can also be obtained from this table through the following calculation:

$$F_{1-\alpha/2}\{v_1, v_2\} = 1/F_{\alpha/2}\{v_2, v_1\}, \tag{3.20}$$

where, for example, $F_{1-\alpha/2}\{v_1, v_2\}$ is an F critical value with v_1 and v_2 degrees of freedom and upper-tail probability $1 - \alpha/2$.

The sample variances for Dies 1 and 2 in the tensile-strength data (see Table 2.2) are, respectively, $s_1^2 = 1.982$ and $s_2^2 = 4.219$. Thus, the sample standard deviations are $s_1 = 1.408$ and $s_2 = 2.054$. A 95% confidence interval for

the ratio of the population variances is

$$\frac{1.982}{(4.219)(2.68)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1.982}{(4.219)(0.37)},$$

or

$$0.175 < \frac{\sigma_1^2}{\sigma_2^2} < 1.259,$$

where $F_{0.025}\{17, 17\} = 2.68$ and $F_{0.975}\{17, 17\} = 1/F_{0.025}\{17, 17\} = 0.37$. On the basis of this procedure, we can state with 95% confidence that the true tensile-strength variance ratio for the two dies is between 0.175 and 1.259. This direct comparison of the two standard deviations is preferable to attempting to infer whether they are equal from a comparison of their individual confidence intervals such as those calculated in Section 3.2.1.

One-sided confidence intervals for the ratio of two variances can be obtained from the two-sided limits in equation (3.19) by using only the upper or the lower bound, as appropriate, and replacing either $\alpha/2$ with α or $1 - \alpha/2$ with $1 - \alpha$. Confidence intervals for the ratio of the standard deviations can be obtained by taking the square roots of the limits for the variance ratio. Table 3.8 displays one- and two-sided confidence intervals for the ratio of two standard deviations.

An important hypothesis that often is of interest in an experiment is

$$H_0: \sigma_1^2 = c^2\sigma_2^2 \quad \text{vs} \quad H_a: \sigma_1^2 \neq c^2\sigma_2^2,$$

TABLE 3.8 Confidence Intervals for the Ratio of Standard Deviations of Two Normal Populations*

(a) *Two-Sided*

$$\left(\frac{s_1^2}{s_2^2 F_{\alpha/2}}\right)^{1/2} < \frac{\sigma_1}{\sigma_2} < \left(\frac{s_1^2}{s_2^2 F_{1-\alpha/2}}\right)^{1/2}$$

(b) *One-Sided Upper*

$$\frac{\sigma_1}{\sigma_2} < \left(\frac{s_1^2}{s_2^2 F_{1-\alpha}}\right)^{1/2}$$

(c) *One-Sided Lower*

$$\left(\frac{s_1^2}{s_2^2 F_{\alpha}}\right)^{1/2} < \frac{\sigma_1}{\sigma_2}$$

*F_a = critical value of F; $v_1 = n_1 - 1$, $v_2 = n_2 - 1$.
Subscripts denote upper-tail probabilities.

where c is the constant specified by the experimenter. Frequently, $c = 1$ and the test is on the equality of the two variances. Note that the equivalent test for standard deviations is

$$H_0: \sigma_1 = c\sigma_2 \quad \text{vs} \quad H_a: \sigma_1 \neq c\sigma_2.$$

The test statistic for use in testing either of these hypotheses is

$$F = s_1^2/c^2 s_2^2. \tag{3.21}$$

If one is testing the equality of the population variances, the above F-statistic is simply the ratio of the sample variances. The decision rule for this test is

$$\begin{aligned} \text{Decision: Reject } H_0 \text{ if} \quad & F > F_{\alpha/2} \\ \text{or if} \quad & F < F_{1-\alpha/2}. \end{aligned}$$

Table 3.9 contains various one-sided and two-sided tests for the population variances. Included are both decision rules and methods for calculating the appropriate significance probabilities. Note again that these tests are identical

TABLE 3.9 Decision Rules and p-value Calculations for Tests on the Standard Deviations of Two Normal Distributions*

(a) $H_0: \sigma_1 = c\sigma_2$ vs $H_a: \sigma_1 \neq c\sigma_2$

Reject H_0 if:	$F_c < F_{1-\alpha/2}$ or $F_c > F_{\alpha/2}$
p-Value:	$p = 2\min(p_l, p_u)$
	$p_l = \Pr\{F < F_c\}$
	$p_u = \Pr\{F > F_c\}$

(b) $H_0: \sigma_1 \geq c\sigma_2$ vs $H_a: \sigma_1 < c\sigma_2$

Reject H_0 if:	$F_c < F_{1-\alpha}$
p-Value:	$p = \Pr\{F < F_c\}$

(c) $H_0: \sigma_1 \leq c\sigma_2$ vs $H_a: \sigma_1 > c\sigma_2$

Reject H_0 if:	$F_c > F_\alpha$
p-Value:	$p = \Pr\{F > F_c\}$

*$F = s_1^2/cs_2^2$, $v_1 = n_1 - 1$, $v_2 = n_2 - 1$; F_c = calculated F-value, F_α = critical value of F.

to those that would be used to test the corresponding hypotheses on the standard deviations.

Apart from the technical details of comparing two means or two standard deviations using independent samples, a very important principle was stated above and should be emphasized. Whenever possible, statistical procedures that account for all sources of variation should be included when making comparisons among two or more parameters. When comparing two means, for example, there are several alternative procedures available, including (a) an examination of the two estimated means (completely ad hoc; ignores the sampling error of the estimators), (b) comparing the two individual confidence intervals (ignores whether the samples are independent or paired; ignores the fact that the joint confidence that both intervals contain the respective population means is not the same as the confidence that each individual interval contains its respective mean), and (c) calculation of the confidence interval on the difference of the two population means, using paired or independent sample methods as appropriate. Clearly the last procedure is the most appropriate of the three. This is a preview of a general class of comparisons referred to as multiple comparison methods that is discussed in Chapter 6.

Another principle that was alluded to in the discussions in this section is the importance of knowing the assumptions needed for various analyses. Both an exact t statistic (3.14) and an approximate t statistic (3.16) were introduced in this chapter to compare two means. The exact t statistic requires an assumption of equal standard deviations for the two populations, whereas the approximate t statistic does not require that assumption. One might choose to first compare the two standard deviations using the F statistic (3.18), but that statistic is very sensitive to whether the data for each population are normally distributed. It has been found that, when in doubt as to whether the two standard deviations are equal, use of the approximate t statistic is generally preferable to performing a preliminary F test followed by use of either the exact or the approximate t statistic depending on the outcome of the F test. As stressed periodically throughout this text, some procedures such as those based on the approximate t statistic for comparing means are less sensitive to critical assumptions than other alternative procedures such as a preliminary F test preceding the use of the exact t statistic. Note, however, that the exact t statistic can be justified on the basis of other information such as previous analyses that show that standard deviations for the two populations can be considered to be equal.

3.5 COMPARING STANDARD DEVIATIONS FROM SEVERAL POPULATIONS

Analysis of variance (ANOVA) methods are presented throughout this text, beginning in Chapter 6, for comparing means from several populations or

processes. While similar methods are occasionally used for comparing several standard deviations, often using the natural logarithm of sample variances as the response variable, they are not a main focal point of this text. There is also a number of alternative procedures that are not based on ANOVA methods that can be used to compare standard deviations. Two of these are described below. Both are highly sensitive to departures from the assumption of normality; consequently, they should be used only after verification that the assumption of normally distributed errors is reasonable.

When using ANOVA models with data from designed experiments, a valuable assessment of the assumption of constant standard deviations across k factor-level combinations is given by the F-max test (see Exhibit 3.3). The F-max test is used to test the hypotheses

$$H_0: \sigma_1 = \sigma_2 = \cdots = \sigma_k \quad \text{vs} \quad H_a: \text{ at least two } \sigma_i \text{ differ.}$$

The F-max test requires that repeat observations be available so that sample standard deviations can be calculated for the observations at each factor-level combination. This test can also be used when the variabilities of k populations or processes are to be compared and random samples are available from each.

EXHIBIT 3.3 F-MAX TEST

1. Calculate the sample standard deviation of the responses from each population or process, or each factor-level combination, as appropriate. Denote these standard deviations by s_i and the corresponding sample sizes by n_i.
2. Calculate the ratio

$$F_{\max} = \left(\frac{\max(s_i)}{\min(s_i)} \right)^2. \tag{3.22}$$

3. If $n_1 = n_2 = \cdots = n_k = n$, use the critical values in Table A.8 of the appendix with $v = n - 1$ to determine whether to reject the hypothesis of equal standard deviations. If the n_i are unequal but not too different, use the harmonic mean of the n_i in place of n; i.e., use $v = n - 1$ with

$$n = k \left(\sum n_i^{-1} \right)^{-1}.$$

The F-max test can be conducted for any set of k samples of observations. Small p-values imply that the assumption of equal error standard deviations is not reasonable. Otherwise, the assumption is accepted as tenable, indicating that there is no strong evidence that one or more standard deviations are much larger than the others.

To illustrate the F-max test, we return to the tensile-strength example. In a manufacturing process, attention is often focused on variability. The variability of the tensile strengths may be of just as great importance as the average strengths of wires produced by the dies. Using the summary statistics given in Table 2.2, $F_{\max} = (2.054/0.586)^2 = 12.29$. This large a value of the $F_{\max}$ statistic causes rejection ($p < 0.001$) of the hypothesis of equal error standard deviations.

An examination of the point plots in Figure 2.2 or the box plots in Figure 3.3 reveals the reason for the rejection of this statistical test. Wire produced from Die 3 is much less variable than wire from the other two dies. Even elimination of the lowest observation for Die 2 (an apparent outlier in Figure 2.2) does not sufficiently reduce the variability of the remaining observations for Die 2 to result in a nonsignificant F-max statistic ($F_{\max} = 8.10$, $p < 0.001$).

A second commonly used test for equal error standard deviations is Bartlett's test (see Exhibit 3.4). Bartlett's test does not require equal sample sizes and is generally regarded as the best test available for the hypothesis of equal standard deviations. Its main drawbacks are its sensitivity to nonnormal errors and the somewhat more difficult calculations than the F-max test.

For the tensile-strength data, $B = 32.40$ and $v = 2$, resulting in a significance probability of $p < 0.001$. As with the F-max test, the hypothesis of equal standard deviations is rejected.

EXHIBIT 3.4 BARTLETT'S TEST

1. For each of the k samples, denote the standard deviations by s_i and the corresponding sample sizes by n_i.
2. Calculate the pooled sample variance

$$s_p^2 = \frac{\sum(n_i - 1)s_i^2}{\sum(n_i - 1)}$$

3. Compute the test statistic

$$\begin{aligned} B &= \ln(s_p^2)\sum(n_i - 1) - \sum(n_i - 1)\ln(s_i^2) \\ &= 2.3026\left\{\log_{10}(s_p^2)\sum(n_i - 1) - \sum(n_i - 1)\log_{10}(s_i^2)\right\} \end{aligned} \quad (3.23)$$

4. Reject the hypothesis of equal standard deviations if B exceeds an upper-tail chi-square critical value with significance level α and degrees of freedom $v = k - 1$.

REFERENCES

Text References

Some useful texts for engineers and scientists that cover much of the material in this chapter were listed at the end of Chapter 2. A useful reference for paired data is the following article.

Bradley, E. L. and Blackwood, L. G. (1989). "Comparing Paired Data: A Simultaneous Test for Means and Variances," *The American Statistician*, **43**: 234–235.

EXERCISES

1 Halon 1301 is an extinguishing agent for fires resulting from flammable vapors and liquids. A chemical company producing fire-extinguishing mixtures involving Halon 1301 sampled 20 batches and measured the concentrations by volume of Halon 1301 to be:

6.11	5.47	5.76	5.31	5.37	5.74	5.54	5.43	6.00	6.03
5.70	5.34	5.98	5.22	5.57	5.35	5.82	5.68	6.12	6.09

Find a 95% confidence interval for the mean concentration, and interpret the results in the context of this exercise. Do the same for the standard deviation of the concentrations.

2 Using the data in Exercise 1, test the following hypotheses:

$$H_0: \mu = 5.7,$$
$$H_a: \mu \neq 5.7.$$

Use a significance level of $\alpha = 0.10$. Calculate the significance probability for the observed value of the test statistic. Interpret the test results in the context of this exercise.

3 Use the Halon 1301 concentrations in Exercise 1 to test the hypotheses

$$H_0: \sigma = 0.5 \text{ vs } H_a: \sigma \neq 0.5.$$

Choose an appropriate significance level, calculate the p-value of the test, and interpret the results in the context of the exercise.

4 A brand of electrical wire is being studied to assess its resistivity characteristics. It is known from information furnished by the manufacturer that resistivity measurements can be considered normally distributed. Construct 95% confidence intervals on the mean and the standard deviation of the resistivity measurements using the data provided below. If one were testing hypotheses about the respective parameters, what values of the parameters would lead to nonrejection of a two-sided test for each?

0.141 0.138 0.144 0.142 0.139 0.146 0.143 0.142

5 Suppose the electrical wire in the previous exercise is to be compared with samples of wire from a second manufacturer using the data provided below. In particular, the means and standard deviations are to be compared to determine whether the two manufacturers' processes result in wire that have similar resistivity properties. Conduct appropriate analyses and draw conclusions about the resistivity characteristics of the two types of wire.

0.133 0.142 0.151 0.148 0.140 0.141 0.150 0.148 0.135

6 Two research laboratories were asked to shoot a 0.05-in. steel cube into a 0.125-in.-thick aluminum target and measure the resulting hole area in the target. Assuming the standard deviations of the measurements of hole areas are equal for both labs, compute a 95% confidence interval on the difference of the two mean hole area measurements. Using the calculated confidence interval, test the hypotheses

$$H_0: \mu_1 = \mu_2,$$
$$H_a: \mu_1 \neq \mu_2.$$

The data on the hole areas (in.^{2}) are as follows:

Laboratory 1:

0.304 0.305 0.302 0.310 0.294 0.293 0.300 0.296
0.300 0.298 0.316 0.304 0.303 0.309 0.305 0.298
0.292 0.294 0.301 0.307

Laboratory 2:

0.315 0.342 0.323 0.229 0.410 0.334 0.247 0.299
0.227 0.322 0.259 0.278 0.361 0.349 0.250 0.321
0.298 0.329 0.315 0.294

7 Assuming the standard deviations of the measurements of hole size for the two research laboratories in Exercise 6 are not equal, compute an approximate 95% confidence interval on the difference of hole-area means. Interpret the confidence interval in the context of this exercise. Use the confidence interval to test the hypothesis

$$H_0: \mu_1 = \mu_2,$$
$$H_a: \mu_1 \neq \mu_2.$$

8 A new alloy is proposed for use in protecting inner walls of freight containers from rupture. The engineering design specification requires a mean penetration of no greater than 2.250 mm when a beam of specified weight and size is rammed into the wall at a specified velocity. Sixty test runs resulted in an average penetration of 2.147 mm and a standard deviation of 0.041 mm. Does it appear from these data that the design specifications are being met? Why (not)? Cite any assumptions needed to perform your analysis.

9 An investigation of technician differences was conducted in the hematology laboratory of a large medical facility. In one portion of the investigation, blood specimens from seven donors were given to two laboratory technicians for analysis. Are any substantive differences detectable from the following analyses of the seven blood samples? Cite any assumptions needed to perform your analyses.

	Measurement	
Specimen	**Technician 1**	**Technician 2**
1	1.27	1.33
2	1.36	1.82
3	1.45	1.77
4	1.21	1.41
5	1.19	1.48
6	1.41	1.52
7	1.38	1.66

10 In the previous exercise, an examination of the laboratory techniques of two technicians was described. Investigate the variability of the blood-sample measurements taken by the technicians. Conduct an appropriate statistical test of the equality of the standard deviations. State all assumptions needed for the test and interpret the results.

11 In a study to determine the effectiveness of a fortified feed in producing weight gain in laboratory animals, one group of animals was administered a standard feed and the other a fortified feed over a six-week period. The weight gain (g) of each animal in the two groups is shown below. Do these data indicate that the fortified feed produces a greater mean weight gain than the standard feed? Is the variability in weight gain the same for the two feeds? What assumptions are you making to perform your analyses?

Standard: 8.9 3.0 8.2 5.0 3.9 2.2 5.7 3.2 9.6 3.1 8.8

Fortified: 5.7 12.0 10.1 13.7 6.8 11.9 11.7 10.4 7.3 5.3 11.8

12 The minor traffic accidents at the fifteen most dangerous locations in a metropolitan area were counted for ten randomly selected days during a six-month period. During the next six months, an intensive program of education and enforcement was instituted in the metropolitan area. Then the accidents were again counted for ten randomly selected days during the next six-month period. Do the data below indicate a change in the mean number of accidents? Justify any assumptions needed to perform your analyses.

Prior to Program	**Following Program**
18, 7, 24, 11, 28, 9, 15, 20, 16, 15	11, 10, 3, 6, 15, 9, 12, 8, 6, 13

13 Compare the variability of the fuel-economy measurements for the four vehicles in Table 2.1 of Chapter 2. Interpret the results in the context of that experiment.

14 Four different processes can be used to manufacture circular aluminum tops for food containers. A critically important measure in the determination of whether the tops can be adequately sealed on the food containers is a "sphericity index." The data below represent sphericity index measurements for several tops from one of the four processes. The target values for the process mean and the process standard deviation are, respectively, 0.75 and 0.02. Assess whether the sphericity index measurements from this process are consistent with the target values.

Process A Sphericity Index Measurements

0.709	0.731	0.706	0.722	0.713
0.734	0.720	0.711	0.729	0.722

15 Construct a box plot from the solar-energy data in Exercise 3 of Chapter 2. Are there any extreme observations depicted? What percentage of the observations lie within the box? Does the sample mean or sample median (or both) better depict a typical energy rate?

16 For the alternator lifetime data from the two manufacturers in Exercise 7 of Chapter 2:

(a) Construct a 95% confidence interval for the mean difference in lifetimes for the two makes of alternators. Interpret the results.

(b) Test the hypothesis of equal standard deviations for the two makes at the 10% significance level. Interpret the results.

(c) Test the hypothesis of equal means for the two makes at the 5% significance level. Interpret the results and reconcile with (a).

(d) Comment on the power of the hypothesis test in (c).

17 The following data were taken from a study of red-blood-cell counts before and after major surgery. Counts were taken on twenty-three patients, all of whom were of the same sex (female) and who had the same blood type (O+).

	Count			Count	
Patient	Pre-op	Post-op	Patient	Pre-op	Post-op
1	14	0	13	5	6
2	13	26	14	4	0
3	4	2	15	15	3
4	5	4	16	4	2
5	18	8	17	0	3
6	3	1	18	7	0
7	6	0	19	2	0
8	11	3	20	8	13
9	33	23	21	4	24
10	11	2	22	4	6
11	3	2	23	5	0
12	3	2			

(a) Make box plots of the pre-op and post-op blood counts. What distinguishing features, if any, are there in the distribution of the blood counts?

(b) Make a scatter diagram of the two sets of counts. Is there an apparent relationship between the two sets of counts?

(c) Make a scatter diagram of the difference of the counts and the sum of the counts. What can you conclude from this plot?

(d) Construct a 95% confidence interval for the mean difference in scores. Interpret the results.

18 Two different test methods are used for determining the viscosity of synthetic rubber. One method is used in the manufacturer's laboratory and the other is used in the customer's lab. Ten test samples were split with one part of the sample sent to each lab. The following data were obtained:

Sample	Manufacturer	Customer
A	91.70	92.26
B	91.87	96.59
C	93.00	95.05
D	93.94	93.66
E	92.61	94.61
F	92.89	95.60
G	92.63	93.44
H	92.35	93.84
I	92.21	93.76
J	93.44	96.17

Use the inferential and graphical techniques discussed in this chapter to assess whether the two test methods are equivalent. Based on your analysis what type of customer–supplier interactions can you anticipate?

19 A research program was funded by an oil company to evaluate a drag-reducing polymer additive for a concentrated brine solution. The polymer was tested at concentrations of 20 and 40 parts per million (by weight) in a 95% saturated salt water solution. Ten tests were performed for each polymer concentration. The system flow rates (gal/min) were recorded for each test. Construct box plots to graphically compare the different levels of polymer concentration. Test the hypotheses that the system flow rate means are equal and that the system flow rate standard deviations are equal. Discuss how these results support the interpretation of the box plots.

Concentration	**System Flow Rates**
20	348, 381, 335, 372, 355, 377, 361, 382, 335, 354
40	371, 387, 376, 390, 369, 364, 380, 385, 369, 383

20 Two petroleum-product refineries produce batches of fuel using the same refining process. A random sample of batches at each refinery resulted in the following values of cetane number measured on the fuel samples.

Use appropriate inferential statistics and graphics to compare the cetane measurements from the two refineries.

Refinery 1:

50.3580	47.5414	47.6311	47.6657	47.7793	47.2890	47.5472
48.2131	46.9531	47.9489	47.3514	48.3738	49.2652	47.2276
48.6901	47.5654	49.1038	49.8832	48.7042	47.9148	

Refinery 2:

45.8270	45.8957	45.2980	45.4504	46.1336	46.6862	45.6281
46.1460	46.3159	45.2225	46.1988	45.5000	45.7478	45.5658

PART II

Design and Analysis with Factorial Structure

CHAPTER 4

Statistical Principles in Experimental Design

This chapter motivates the use of statistical principles in the design of experiments. Several important facts are stressed:

- *statistically designed experiments are economical,*
- *they allow one to measure the influence of one or several factors on a response,*
- *they allow the estimation of the magnitude of experimental error, and*
- *experiments designed without adhering to statistical principles usually violate one or more of these desirable design goals.*

Test procedures in scientific and engineering experiments are often primarily guided by established laboratory protocol and subjective considerations of practicality. While such experimental procedures may be viewed as economical in terms of the number of test runs that must be conducted, the economy of effort can be deceiving for two reasons. First, economy is often achieved by severely limiting the number of factors whose effects are studied. Second, the sequence of tests may require that only one of the factors of interest be varied at a time, thereby preventing the evaluation of any joint effects of the experimental factors. The effective use of statistical principles in the design of experiments ensures that experiments are designed economically, that they are efficient, and that individual and joint factor effects can be evaluated.

In the next several chapters a variety of statistical experimental designs are presented. In this chapter we discuss general concepts that arise in virtually all experimental settings. The chapter begins with an introduction to the common terminology used in discussing statistical design procedures. Statistical

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

experimental design is then motivated by an examination of problems that frequently arise when statistical principles are not used in the design and conduct of a test program. Special emphasis is placed on the investigation of the joint effects of two or more experimental factors on a response. Finally, a summary of important design considerations is presented.

4.1 EXPERIMENTAL-DESIGN TERMINOLOGY

The terminology of experimental design is not uniform across disciplines or even, in some instances, across textbooks within a discipline. For this reason we begin our discussion of statistical experimental design with a brief definition of terms. Table 4.1 contains definitions of many terms which are

TABLE 4.1 Experimental-Design Terminology

Block. Group of homogeneous experimental units.
Confounding. One or more effects that cannot unambiguously be attributed to a single factor or interaction.
Covariate. An uncontrollable variable that influences the response but is unaffected by any other experimental factors.
Design (layout). Complete specification of experimental test runs, including blocking, randomization, repeat tests, replication, and the assignment of factor–level combinations to experimental units.
Effect. Change in the average response between two factor–level combinations or between two experimental conditions.
Experimental region (factor space). All possible factor–level combinations for which experimentation is possible.
Factor. A controllable experimental variable that is thought to influence the response.
Homogeneous experimental units. Units that are as uniform as possible on all characteristics that could affect the response.
Interaction. Existence of joint factor effects in which the effect of each factor depends on the levels of the other factors.
Level. Specific value of a factor.
Repeat tests. Two or more observations that have the same levels for all the factors.
Replication. Repetition of an entire experiment or a portion of an experiment under two or more sets of conditions.
Response. Outcome or result of an experiment.
Test run. Single combination of factor levels that yields an observation on the response.
Unit (item). Entity on which a measurement or an observation is made; sometimes refers to the actual measurement or observation.

in common use. A few of these terms have already been used in previous chapters but are included here for completeness.

The terms *response* and *factor* were defined in Chapter 1 (Section 1.2). A response variable is an outcome of an experiment. It may be a quantitative measurement such as the percentage by volume of mercury in a sample of river water, or it may be a qualitative result such as whether an aircraft engine mounting bolt can withstand a required shearing force. A factor is an experimental variable that is being investigated to determine its effect on a response. It is important to realize that a factor is considered controllable by the experimenter; that is, the values, or *levels*, of the factor can be determined prior to the beginning of the test program and can be executed as stipulated in the experimental design. While the term *version* is sometimes used to designate categorical or qualitative levels of a factor, we use *level* to refer to the values of both qualitative and quantitative factors. An *experimental region*, or *factor space*, consists of all possible levels of the factors that are candidates for inclusion in the design. For quantitative factors, the factor space often is defined by lower and upper limits for the levels of each factor.

Additional variables that may affect the response but cannot be controlled in an experiment are called *covariates*. Covariates are not additional responses; that is, their values are not affected by the factors in the experiment. Rather, covariates and the experimental factors jointly influence the response. For example, in many experiments both temperature and humidity affect a response, but the laboratory equipment can only control temperature; humidity can be measured but not controlled. In such experiments temperature would be regarded as an experimental factor and humidity as a covariate.

A *test run* is a single factor–level combination for which an observation (response) is obtained. *Repeat tests* are two or more observations that are obtained for a specified combination of levels of the factors. Repeat tests are conducted under as identical experimental conditions as possible, but they need not be obtained in back-to-back test runs. Repeat tests should not be two or more analytical determinations of the same response; they must be two or more identical but distinct test runs. *Replications* are repetitions of a portion of the experiment (or the entire experiment) under two or more different conditions, for example, on two or more different days.

Experimental responses are only comparable when they result from observations taken on *homogeneous* experimental units. An experimental unit was described in Chapter 1 (Section 1.2) as either a measurement or material on which a measurement is made. (Note: In keeping with the above discussion, *measurement* as used here can be either quantitative or qualitative.) Unless explicitly stated otherwise, we shall use the term to refer to a physical entity on which a measurement is made. *Homogeneous* experimental units do not differ from one another in any systematic fashion and are as alike as possible on all characteristics that might effect the response. While there is inherent random

variation in all experimental units, the ability to detect important factor effects and to estimate these effects with satisfactory precision depends on the degree of homogeneity among the experimental units.

If all the responses for one level of a factor are taken from experimental units that are produced by one manufacturer and all the responses for another level of the factor are taken from experimental units produced by a second manufacturer, any differences noted in the responses could be due to the different levels of the factor, to the different manufacturers, or to both. In this situation the effect of the factor is said to be *confounded* with the effect due to the manufacturers.

When a satisfactory number of homogeneous experimental units cannot be obtained, statistically designed experiments are often *blocked* so that homogeneous experimental units receive each level of the factor(s). Blocking divides the total number of experimental units into two or more groups or blocks (e.g., manufacturers) of homogeneous experimental units so that the units in each block are more homogeneous than the units in different blocks. Factor levels are then assigned to the experimental units in each block. If more than two blocks of homogeneous experimental units can be obtained from each manufacturer, both repeat tests (two or more identical factor–level combinations on units within a block) and replication (repetition of the design for one or more of the blocks from each manufacturer) can be included in the experiment.

The terms *design* and *layout* often are used interchangeably when referring to experimental designs. The layout or design of the experiment includes the choice of the factor–level combinations to be examined, the number of repeat tests or replications (if any), blocking (if any), the assignment of the factor–level combinations to the experimental units, and the sequencing of the test runs.

An *effect* of the design factors on the response is measured by a change in the average response under two or more factor–level combinations. In its simplest form, the effect of a single two-level factor on a response is measured as the difference in the average response for the two levels of the factor; that is

$$\text{factor effect} = \text{average response at one level} - \text{average response at a second level.}$$

Factor effects thus measure the influence of different levels of a factor on the value of the response. Individual and joint factor effects are discussed in Section 5.2.

To illustrate the usage of these experimental-design terms, two examples are now presented. The design shown in Table 4.2 is for an experiment that is to be conducted to study the flow of suspended particles in two types of coolants used with industrial equipment. The coolants are to be forced through a slit aperture in the middle of a fixed length of pipe (see Figure 4.1).

TABLE 4.2 Design for Suspended-Particulate Study*

Run No.	Test Fluid	Pipe Angle (degrees from horizontal)	Flow Rate (ft/sec)
1	1	60	60
2	2	30	60
3	1	60	90
4	2	45	60
5	2	15	90
6	1	15	60
7	2	15	60
8	1	45	60
9	2	45	90
10	1	15	90
11	1	45	90
12	1	30	60
13	2	60	60
14	1	30	90
15	2	30	90
16	2	60	90

*Covariate: temperature (°C).

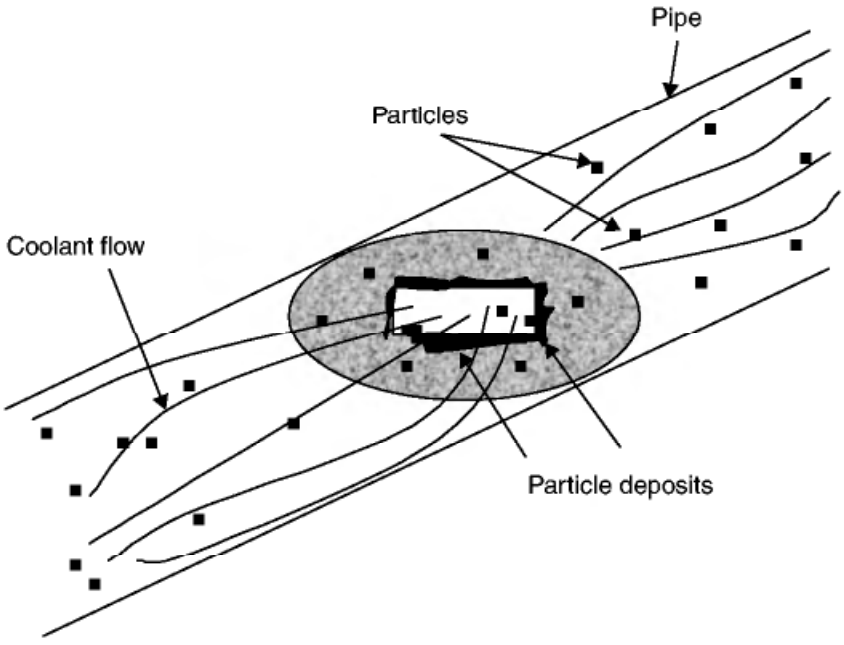


Figure 4.1 Suspended-particulate experiment.

Using two different flow rates and four different angles of inclination of the pipe, the experimenters wish to study the buildup of particles on the edge of the aperture.

There are three experimental factors in the study: coolant, a qualitative factor at two levels (1, 2); pipe angle, a quantitative factor at four levels (15, 30, 45, 60 degrees from horizontal); and flow rate, a quantitative factor at two levels (60, 90 ft/sec). All sixteen combinations (two coolants $\times$ four angles $\times$ two rates) of the factor levels are to be included in the design. The sequencing of the tests was determined by randomly assigning a run number to the factor–level combinations. All the test runs are to be conducted during a single day to eliminate day-to-day variation. It is believed, however, that the expected 20-degree change in temperature from early morning to late afternoon may have an effect on the test results. For this reason temperature will be measured as a covariate.

The second example, shown in Figure 4.2, represents an agricultural experiment in which five soil treatments (e.g., different types of fertilizer, different methods of plowing) are to be investigated to determine their effects on the yield of soybean plants. The experiment must be conducted on three different fields to obtain a sufficient number of homogeneous plots of ground. Each field contains five such homogeneous plots; however, the soil conditions on each field are known to be different from those on the other two fields.

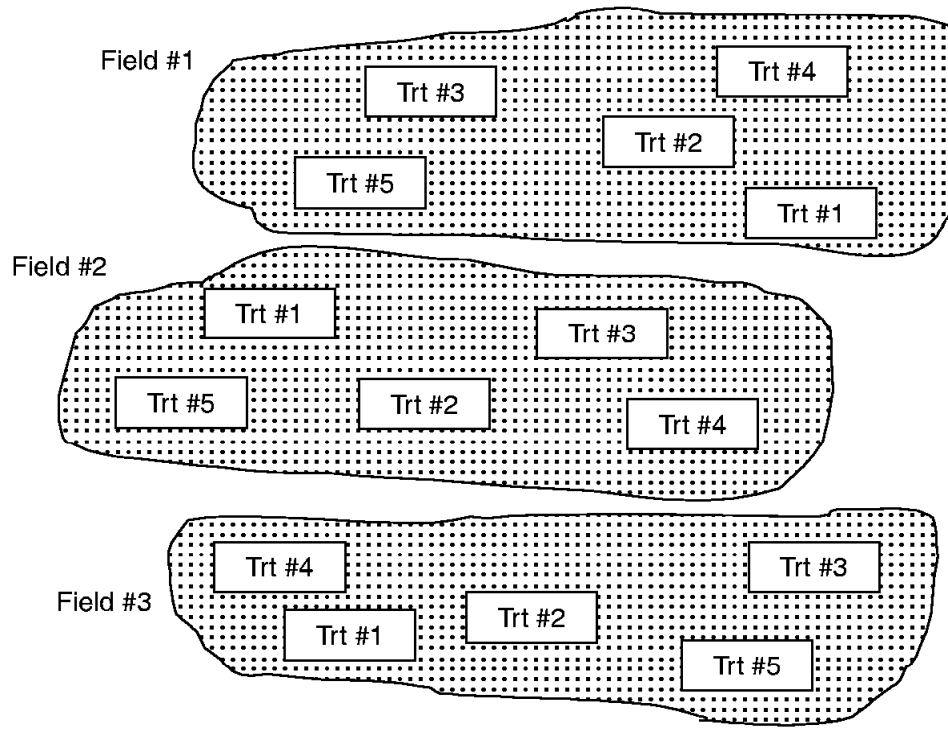


Figure 4.2 Field layout for soil treatment experiment.

In this experiment the fields are blocks and the plots are the experimental units. One qualitative factor, soil treatment, having five levels is under investigation. The soil treatments are randomly assigned to the plots in each field (block). The response variable is the yield in kilograms per plot.

In Table 4.2 the factor–level combinations are listed in the order in which the test runs will be conducted. There are no physical experimental units per se. Each test run takes the place of an experimental unit. In Figure 4.2 an experimental unit is a plot of ground. Unlike the previous example, there is no test sequence; all the tests are conducted simultaneously. The two examples illustrate two different ways to specify the statistical design of an experiment: in run order or as a physical layout.

4.2 COMMON DESIGN PROBLEMS

When statistical considerations are not incorporated in the design of an experiment, statistical analyses of the results are often inconclusive or, worse yet, misleading. Table 4.3 lists a few of many potential problems that can occur when statistical methodology is not used to design scientific or engineering experiments:

4.2.1 Masking Factor Effects

Researchers often invest substantial project funds and a great amount of time and effort only to find that the research hypotheses are not supported by the experimental results. Many times the lack of statistical confirmation is the result of the inherent variability of the test results.

Consider for example the data listed in Table 4.4. These are test results measuring cylinder pressure in a single-cylinder engine under strictly controlled laboratory conditions. The test results are for thirty-two consecutive firing cycles of the engine. Note the variation in the test results even for this highly controlled experiment. If an experiment is conducted using this engine

TABLE 4.3 Common Experimental-Design Problems

- Experimental variation masks factor effects.
- Uncontrolled factors compromise experimental conclusions.
- Erroneous principles of efficiency lead to unnecessary waste or inconclusive results.
- Scientific objectives for many-factor experiments may not be achieved with one-factor-at-a-time designs.

TABLE 4.4 Cylinder-Pressure Measurements under Controlled Laboratory Conditions

229	191	238	231	253	189	224	224
191	201	200	201	197	206	220	214
193	226	237	209	161	187	237	245
181	213	231	217	212	207	242	186

Average = 212.28, Standard Deviation = 21.63

TABLE 4.5 Skin Color Measurements

	Color Measurement		
Participant	Week 1	Week 2	Week 3
A	12.1	14.2	13.9
B	19.1	17.6	16.2
C	33.8	34.7	33.2
D	33.0	31.7	30.3
E	35.8	37.7	35.6
F	42.0	38.4	41.5
G	36.8	35.2	35.7

and the factor effects are of the same order of magnitude as the variation evident in Table 4.4, the effects may go undetected.

Table 4.5 illustrates another kind of variation in test results. In this study of skin color measurements not only is there variation among the participants, but there is also variation for each participant over the three weeks of the study. Experiments that are intended to study factor effects (e.g., suntan products) on skin color must be designed so that the variation in subjects and across time does not mask the effects of the experimental factors.

Figure 4.3 is a schematic representation of the relationship between the detectability of factor effects and the variability of responses. In this figure, test results from two levels of a factor are indicated by squares and circles, respectively. In both cases shown, the average response for each factor level remains constant; consequently, the factor effect (the difference in the averages of the two levels) does not change. Only the variability of the response changes from case to case.

In case 1, the variability of the test results is so great that one would question whether the factor effect is (a) measuring a true difference in the population or process means corresponding to the two factor levels or (b) simply due to the variation of the responses about a common mean. The data may

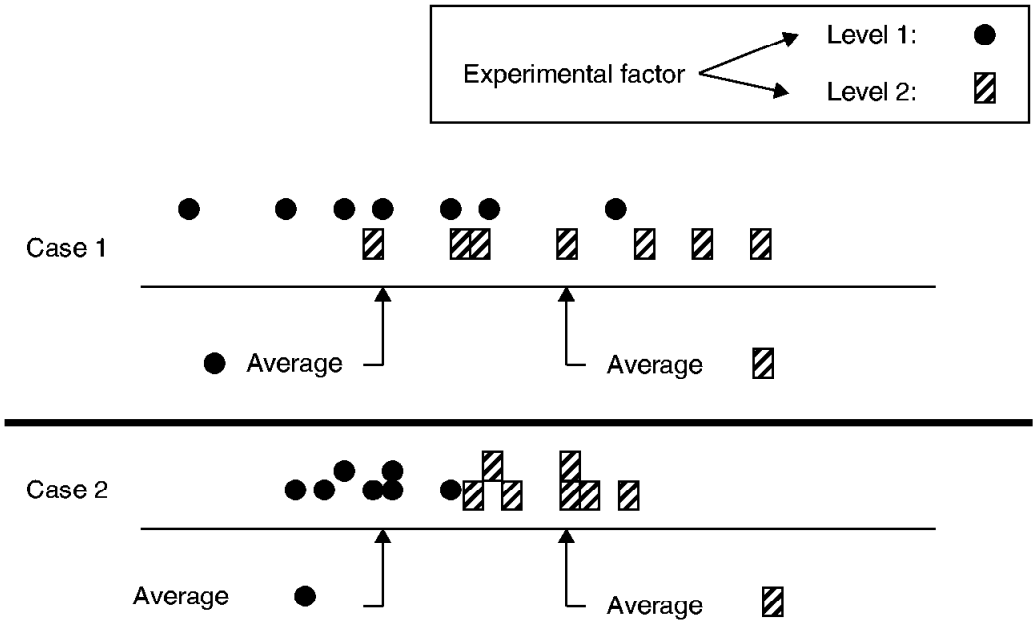


Figure 4.3 Experimental variability and factor effects.

not provide sufficiently convincing evidence that the two population means are different because of the variability in the responses.

In case 2, the variation of the responses is much less than that in case 1. There is strong evidence, due to the small variation in the responses relative to the large differences in the averages, that the factor effect is measuring a substantial difference between the population or process means corresponding to the two factor levels. Thus, the difference in the means due to the factor levels would be masked by the variability of the responses in case 1 but not in case 2. The implication of this example for the statistical design of experiments is that the variation in case 1 must be compensated for (e.g., by blocking or a large experiment size) to ensure that the difference in the means is detectable.

The importance of this discussion is that experimental error variation must be considered in the statistical design of an experiment. Failure to do so could result in true factor effects being hidden by the variation in the observed responses. Blocking and sample size are two key considerations in controlling or compensating for response variability.

4.2.2 Uncontrolled Factors

A second problem listed in Table 4.3 that frequently occurs when statistical considerations are not included in the design phase of a project is the effect of uncontrolled factors on the response. While few researchers would intentionally ignore factors that are known to exert important influences on a response,

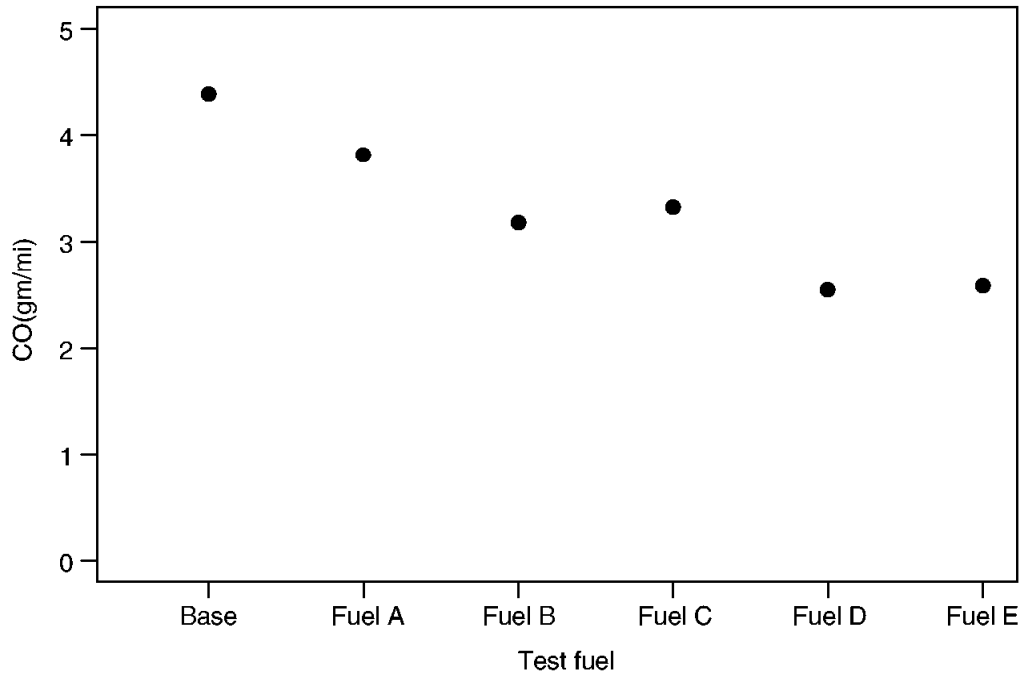


Figure 4.4 Average carbon monoxide emissions. (Data from *Performance Evaluation of Alcohol-Gasoline Blends in 1980 Model Automobiles, Phase II-Methanol-Gasoline Blend* Atlanta, GA: Coordinating Research Council, Table G-8 [1984].)

there are many subtle ways in which failure to carefully consider all factors of importance can compromise the conclusions drawn from experimental results.

Consider, for example, an experiment conducted to compare average carbon monoxide (CO) emissions for a commercial gasoline (base fuel) and five different methanol fuel blends. Suppose further that an adequate experimental design has been selected, including a sufficiently large number of test runs. Figure 4.4 exhibits actual results of such a test program. One of the important conclusions that can be drawn from an analysis of the experimental data is that the last two fuels have significantly lower average CO emissions than the other fuels tested.

Subsequent to a conclusion such as this, researchers often wish to determine which of several fuel properties (e.g., distillation temperatures, specific gravity, oxygen content) contribute to the reduction in the average emission levels. The difficulty with such a determination is that the fuel properties of interest cannot all be specifically controlled in the selection of fuel blends. Because of this, many fuel properties that might subsequently be of interest simultaneously vary across the six fuels, resulting in a confounding of their effects.

This experiment was specifically designed only to investigate the effects of six fuel blends on CO emissions. Studies could be specifically designed to study some of the fuel properties. In such studies the fuel properties would be varied in a systematic fashion and confounding among the properties could

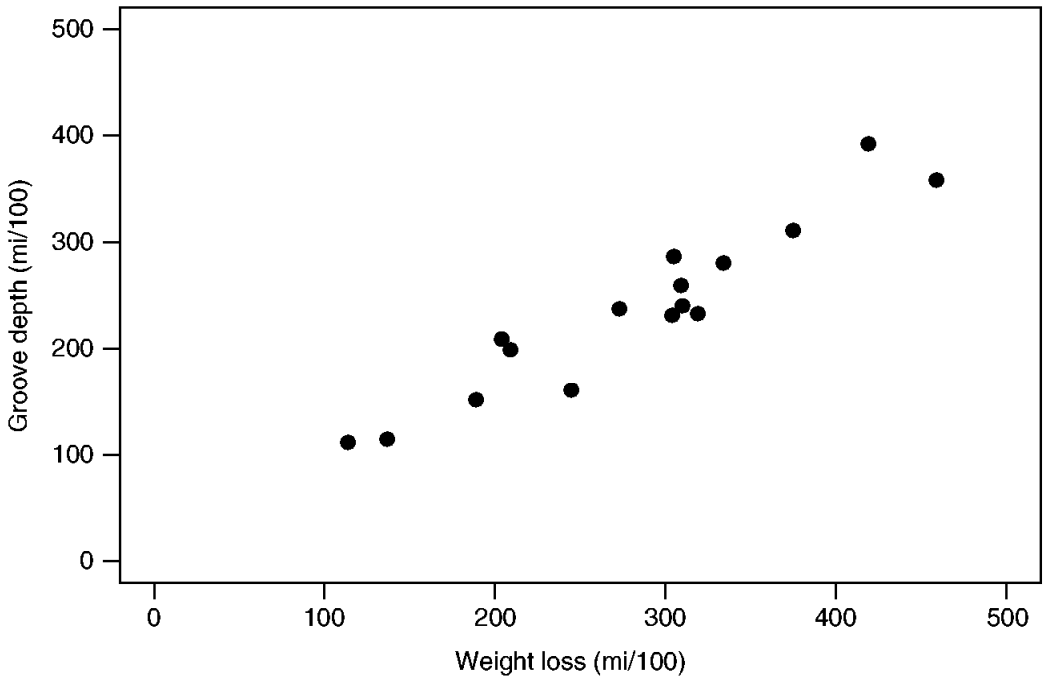


Figure 4.5 Estimates of tread life using two methods. (Data from Natrella, 1963, pp. 5–33.)

be eliminated by the choice of the design. Note that in this example it is the uncontrolled variation of the fuel properties that leads to the confounding of their effects on the response.

Another example in which uncontrolled factors influence a response is depicted in Figure 4.5. In this graph, results of two methods of determining tire wear are plotted: the *groove-depth* and the *weight-loss* method. In studies of this type one sometimes seeks to determine an empirical relationship between the two measurements, perhaps to calibrate the two methods of measurement.

The major source of the relationship between the two methods of estimating tread life is not the close agreement between the respective measurements. Some plotted points in Figure 4.5 that represent similar measurements on one method differ by as much as 5000 miles on the other method. The major source of the association among the plotted points is the large variation among the observations due to uncontrolled factors such as road conditions, weather, vehicles, and drivers. Thus, it is not so much a close relationship between the methods of estimating tread life as it is the large differences in uncontrolled test conditions (and consequent large differences in tire wear) that contributes to the linear trend observable in Figure 4.5.

4.2.3 Erroneous Principles of Efficiency

The preceding examples demonstrate the need to construct designs in which factors of interest are systematically varied and to consider the likely magnitude

of the inherent variation in the test results when planning the number of test runs. The third problem listed in Table 4.3 suggests that the desire to run economical experiments can lead to strategies that may in fact be wasteful or even prevent the attainment of the project's goals. The latter difficulty is elaborated on in the next section, where one-factor-at-a-time testing is discussed.

Time and cost efficiencies are always important objectives in experimental work. Occasionally efficiency becomes an overriding consideration and the project goals become secondary. If time or budgetary considerations lead to undue restrictions on the factors and levels that can be investigated, the project goals should be reevaluated relative to the available resources. This may lead to a decision to forgo the experimentation.

The problem of experiment efficiency is most acute when several factors must be investigated in an experiment. When guided only by intuition, many different types of designs could be proposed, each of which might lead to flawed conclusions. Some would choose to hold factors constant that could have important influences on the response. Others would allow many unnecessary changes of factors that are inexpensive to vary and few changes of critical factors that are costly to vary.

Efficiency is achieved in statistically designed experiments because each observation generally provides information on all the factors of interest in the experiment. Table 4.6 shows the number of test runs from Table 4.2 that provide information on each level of the test factors. If each of the test factors had been investigated separately using the same number of test runs shown in Table 4.6 then 48 test runs would have been needed.

Information on individual and joint factor effects can be obtained from a highly efficient experiment such as the one displayed in Table 4.2. It is

TABLE 4.6 Number of Test Runs for Each Factor Level: Suspended-Particulate Study

Factor	Level	Number of Test Runs
Test fluid	1	8
	2	8
Flow rate	60	8
	90	8
Pipe angle	15	4
	30	4
	45	4
	60	4
Equivalent single-factor experiment size		48

neither necessary nor desirable to investigate a single factor at a time in order to economically conduct experiments. Because there is a prevalent view that one-factor-at-a-time testing is appropriate when there are several factors to be investigated, we now focus on this type of experimentation.

4.2.4 One-Factor-at-a-Time Testing

Consider an experimental setting in which one wishes to determine which combinations of the levels of several factors optimize a response. The optimization might be to minimize the amount of impurities in the manufacture of a silicon wafer. It might be to maximize the amount of a compound produced from a reaction of two or more chemicals. It might be to minimize the number of defective welds made by a robot on an assembly line.

In each of these examples the optimization is a function of several factors, which can be experimentally investigated. Because of the feared complexity of simultaneously investigating the influence of several factors on a response, it is common practice to vary one factor at a time in the search for an optimum combination of levels of the factors. Table 4.7 lists a sequence of steps in the conduct of such an experiment.

The perceived advantages of one-factor-at-a-time testing are primarily two:

(i) the number of test runs is believed to be close to the minimum that can be devised to investigate several factors simultaneously, and

(ii) one can readily assess the factor effects as the experiment progresses, because only a single factor is being studied at any stage.

TABLE 4.7 One-Factor-at-a-Time Test Sequence

Stage 1:	Fix levels of factors $2, 3, \ldots, k$; determine the optimal level of factor 1.
Stage 2:	Use the optimal level of factor 1 from stage 1; fix levels of factors $3, 4, \ldots, k$; determine the optimal level of factor 2.
Stage 3:	Use the optimal levels of factors 1, 2 from stages 1, 2; fix levels of factors $4, 5, \ldots, k$; determine optimal level of factor 3.
⋮	⋮
Stage k:	Use the optimal levels of factors $1, 2, \ldots, k-1$ from stages $1, 2, \ldots, k-1$; determine the optimal level of factor k.

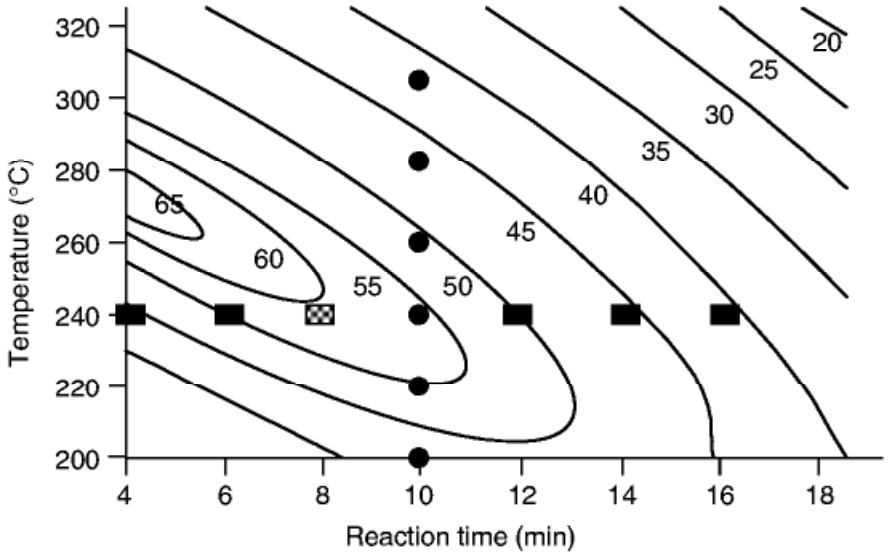


Figure 4.6 Contours of constant yield (%) of a chemical reaction.

These attractive features of one-factor-at-a-time testing are more than offset by the potential for failure to achieve the optimization sought in the design of the study.

Figure 4.6 is a typical contour plot of the yield of a chemical reaction. The curves in the figure are curves of constant yield as a function of the two factors of interest, temperature and reaction time. The curve labeled 45, for example, identifies combinations of temperature and reaction time for which the yield of the chemical reaction is 45%.

Many industrial experiments are conducted because contours such as those depicted in Figure 4.6 are unknown. Suppose one does not know the contours in Figure 4.6 and one wishes to conduct an experiment to determine combinations of temperature and reaction time that will maximize the yield. If the range of interest for the temperature variable is 200–300°C and that of the reaction time is 4–16 min, the region graphed in Figure 4.6 will include the entire region of interest. If one conducts a one-factor-at-a-time experiment by fixing reaction time and varying temperature from 200 to 300°C in increments of 20°, it would be reasonable to set the reaction time to 10 min, the middle of the range of interest.

The points identified by circles in Figure 4.6 represent the six observations that would be taken in this first stage of the experiment. Note that one might stop this stage of the testing after the test run at 260°C, because the yield would be declining noticeably thereafter. The optimal level for temperature based on these test runs would be 240°C.

The second stage of the experiment, identified by the squares in Figure 4.6, is a sequence of test runs for different reaction times with temperature set at its "optimal" level of 240°C. The optimal level of reaction time would be

determined to be 8 min if observations were taken in 2-min increments as indicated in the figure. Note that the optimal levels of temperature (240) and reaction time (8) determined in this manner produce a yield of 60%.

The true maximum yield for this chemical process exceeds 70%. It is obtainable at a temperature of approximately 270°C and a reaction time of 4 min. Only by experimenting in a suitable grid of points that includes the corners of the experimental region can one locate optimum yields such as the one depicted in Figure 4.6.

This one-factor-at-a-time experiment also does not allow the fitting of a model from which the contours shown in Figure 4.6 can be drawn. This is because the observations along the two paths indicated by the circles and squares cannot characterize the curvature of the region. This is another drawback to one-factor-at-a-time testing, because many times the goal of a study is not only to optimize the response but also to model its behavior over the experimental region (factor space).

A second example of an investigation in which one-factor-at-a-time experimentation may fail to achieve an optimum response is shown in Table 4.8. This investigation is intended to identify optimal combinations of three catalysts for reducing the reaction time of a chemical process. The desired ranges on the catalysts are 3–6% for catalysts A and B, 4–8% for catalyst C. Suppose that in an initial investigation of this process, only combinations of the two extreme levels of these factors are to be examined; that is catalyst $A = (3, 6)$, catalyst $B = (3, 6)$, and catalyst $C = (4, 8)$.

Table 4.8 lists the results of two possible one-factor-at-a-time experiments. In experiment 1 the first test run has all three catalysts at their low levels. The reaction time was longer than one minute, an unacceptable duration. The

TABLE 4.8 Reaction Times of a Chemical Process

Run No.	Level of Catalyst			Reaction Time (sec)
	A	*B*	*C*	
Experiment 1				
1	3	3	4	>60
2	6	3	4	>60
3	3	6	4	>60
4	3	3	8	>60
Experiment 2				
5	6	6	4	>60
6	3	6	8	>60
7	6	3	8	54.3
8	6	6	8	27.6

second test run holds catalysts B and C fixed and changes A to its upper level. Again the reaction time is unacceptable. Because the higher amount of catalyst A produced no better results than its lower level, the third test run again sets catalyst A to its lower level and changes catalyst B to its upper level. The reaction time is unacceptable. Finally, the last run of this experiment has catalyst C at its upper level and the other two catalysts at their lower levels. Again the results are unacceptable.

This application of one-factor-at-a-time testing might lead to the conclusion that no combination of catalysts would reduce the reaction time to an acceptable amount. Note, however, that the sequence of test runs for experiment 2 does lead to acceptable reaction times. Experiment 2 systematically tests pairs of factors at their upper levels. This experiment would lead to the optimal choice in four test runs. If experiment 2 were conducted as a continuation of experiment 1, the optimal combination would not be identified until all eight possible combinations of the factor levels have been tested. In this case one-factor-at-a-time testing achieves no economy of effort relative to testing all possible combinations of the factor level.

The drawbacks to this type of testing should now be apparent. Optimal factor combinations may not be obtained when only one factor is varied at a time. Also, the combinations of levels that are tested do not necessarily allow appropriate models to be fitted to the response variable. Additional test runs may have to be added if an estimate of experimental error is to be obtained.

One-factor-at-a-time experimentation is not only used to determine an optimum combination of factors. Often this type of testing is used merely to assess the importance of the factors in influencing the response. This can be an impossible task with one-factor-at-a-time designs if the factors jointly, not just individually, influence the response.

One-factor-at-a-time experimentation does not always lead to incorrect or suboptimal results. The examples used in this section are intended to illustrate the dangers that this type of experimentation pose. As mentioned in the introduction to this chapter, there are economically efficient statistical experimental designs that do permit the fitting of curved response contours, the investigation of joint factor effects, and estimation of experimental-error variation. Many of these designs are discussed in subsequent chapters.

4.3 SELECTING A STATISTICAL DESIGN

To avoid many of the potential pitfalls of experimentation that were mentioned in the previous sections of this chapter, several key criteria should be considered in the design of an experiment. Among the more important design considerations are those listed in Table 4.9.

TABLE 4.9 Statistical Design Criteria

Consideration of Objectives
• Nature of anticipated conclusions
• Definition of concepts
• Determination of observable variables
Factor Effects
• Elimination of systematic error
• Measurement of covariates
• Identification of relationships
• Exploration of entire experimental region
Precision
• Estimation of variability (uncertainty)
• Blocking
• Repeat tests, replication
• Adjustment for covariates
Efficiency
• Multiple factors
• Screening experiments
• Fractional factorials
Randomization

4.3.1 Consideration of Objectives

The first criterion listed in Table 4.9 is the most obvious and necessary for any experiment. Indeed, one always sets goals and determines what variables to measure in any research project. Yet each of the three subheadings in this criterion has special relevance to the statistical design of an experiment.

Consideration of the nature of the anticipated conclusions can prevent unexpected complications when the experiment is finished and the research report is being written. The fuel study discussed in Section 4.2.2 is a good example of the need to clearly resolve the specific objectives of an experiment. If one's goal is to study the effects of various fuel properties on a response, the experimental design might be altogether different than if one's goal is just to compare six fuels.

Concept definition and the determination of observable variables influence both the experimental design and the collection of information on uncontrollable factors. For example, suppose one wishes to study the effects of radiation on human morbidity and mortality. One cannot subject groups of humans to varying levels of radiation, as one would desire to do in a designed experiment, if it is believed that such exposure would lead to increased illness or

death. Alternatives include studies with laboratory animals and the collection of historical data on humans. The latter type of study is fraught with the problem of uncontrollable factors similar to that of the tire-wear example in Section 4.2.2. Laboratory studies allow many important factors to be controlled, but the problem of drawing inferences on human populations from animal studies arises.

Apart from these difficulties, statistical issues relating to the measurement of responses and covariates arise. For example, what measure of mortality should be used? Should one use raw death rates (e.g., number of deaths per 100,000 population) or should the death rates be age-, sex-, or race-adjusted? Should covariates such as atmospheric radiation be measured? If so, how will the analysis of the experimental data incorporate such covariate measurements?

4.3.2 Factor Effects

A second criterion that must be considered in the selection of an appropriate statistical design is the likely effects of the factors. Inclusion of all relevant factors, when experimentally feasible, is necessary to ensure that uncontrolled systematic variation of these factors will not bias the experimental results. An accounting for uncontrollable systematic variation through the measurement of covariates is necessary for the same reason.

Anticipated factor effects also influence the choice of a statistical design through their expected relationships with one another. If each factor is believed to affect the response independently of any other factor or if joint effects of the factors are of secondary interest (as in certain pilot studies), screening experiments (Chapter 7) can be used to assess the effects of the factors. If the effect of one factor on the response depends on the levels of the other factors, a larger design is needed.

In general, an experimental design should allow the fitting of a general enough model so that the salient features of the response and its relationships with the factors can be identified. For example, the design should permit polynomial terms in the quantitative factors to be included in the fitted model so that curvature in the response function can be assessed. The design should permit an assessment of the adequacy of the fitted model. If the fitted model is judged to be an inadequate representation of the response function, the design should form the basis for an expanded design from which more elaborate models (e.g., higher-order polynomials) can be fitted.

When assessing factor effects it is important to explore the entire experimental region of interest. The combinations of factor levels used in a statistical design should be selected to fill out the experimental region. If a factor is only studied over a narrow portion of the experimental region, important effects may go undetected.

4.3.3 Precision and Efficiency

The next two categories of design criteria listed in Table 4.9, precision and efficiency, will be amply discussed in each of the next several chapters as they relate to specific statistical designs. Because the term *precision* is used frequently throughout this book, we comment briefly on its meaning in this section.

Precision refers to the variability of individual responses or to the variability of effects that measure the influence of the experimental factors (see Exhibit 4.1). Precision is a property of the random variables or statistics and not of the observed values of those variables or statistics. For example, an (observed) effect is said to be sufficiently precise if the standard deviation of the statistic that measures the effect is suitably small. In its simplest form, an effect is simply the difference between two averages. The corresponding statistic is

$$\bar{y}_1 - \bar{y}_2.$$

An observed effect is then said to be sufficiently precise if the standard deviation (or, equivalently, the variance) of this statistic is sufficiently small. In practice, one does not know the value of the standard deviation, but it can be estimated from the data.

EXHIBIT 4.1

Precision. The precision of individual responses refers to the variability inherent in independent observations taken on a response under identical conditions. The precision of a factor effect refers to the variability of the statistic that is used to measure the effect. Satisfactory precision is usually defined in terms of a sufficiently small standard deviation of the response variable or the statistic measuring the effect, respectively.

Blocking, repeat tests, replication, and adjustment for covariates can all increase precision in the estimation of factor effects. Blocking increases precision (decreases variability) by controlling the systematic variation attributable to nonhomogeneous experimental units or test conditions. Adjustment for covariates increases precision by eliminating the effects of uncontrolled factors from the variability that would otherwise be attributed to random error.

Repeat tests and replication increase precision by reducing the standard deviations of the statistics used to estimate effects. For example, as mentioned in Section 1.3 (see also Section 2.3), the standard deviation of a sample mean based on n independent observations from a single population or process is $\sigma/n^{1/2}$, where σ is the population or process standard deviation. Increasing the number of repeat tests or replications increases the sample size n, thereby decreasing the standard deviation.

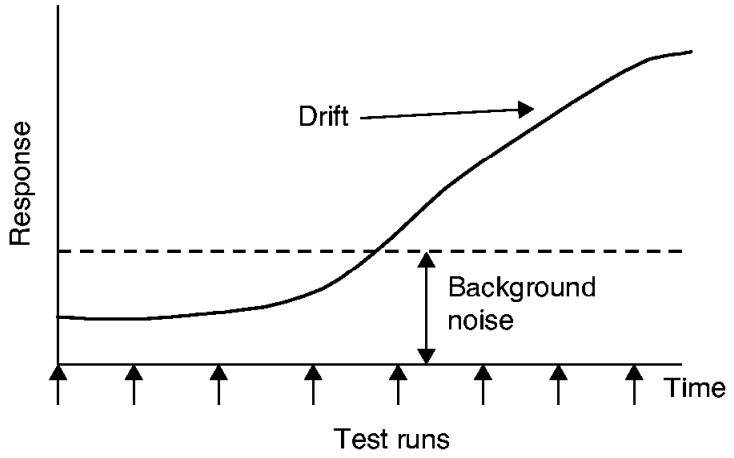


Figure 4.7 Influence of machine drift; test runs indicated by arrows.

4.3.4 Randomization

Randomization of the sequence of test runs or the assignment of factor–level combinations to experimental units protects against unknown or unmeasured sources of possible bias. Randomization also helps validate the assumptions needed to apply certain statistical techniques.

The protection that randomization affords against unknown bias is easily appreciated by considering the common problem of instrument drift. If during an experiment instrument drift builds over time as illustrated in Figure 4.7, later tests will be biased because of the drift. If all tests involving one level of a factor are run first and all tests involving the second level of a factor are run last, comparisons of the factor levels will be biased by the instrument drift and will not provide a true measure of the effect of the factor.

Randomization of the test runs cannot prevent instrument drift. Randomization can help ensure that all levels of a factor have an equal chance of being affected by the drift. If so, differences in the responses for pairs of factor levels will likely reflect the effects of the factor levels and not the effect of the drift.

The design criteria listed in Table 4.9 are not intended to be comprehensive. They are presented as a guide to some of the more important considerations that must be addressed in the planning stages of most experiments.

4.4 DESIGNING FOR QUALITY IMPROVEMENT

The contributions of statistical methodology to strategies for quality improvement can be divided into two broad categories: on-line and off-line statistical procedures. Control-chart techniques are an example of on-line procedures.

Such procedures assure that the process or system is in statistical control and that it maintains whatever consistency it is capable of achieving.

While in the past it was generally believed that on-line statistical quality-control techniques offer satisfactory assurances of customer satisfaction, today it is recognized that off-line investigations using engineering design techniques and statistical design of experiments provide the greatest opportunity for quality improvement and increased productivity. Off-line experiments are performed in laboratories, pilot plants, and preliminary production runs, ordinarily prior to the complete implementation of production or process operations. As suggested by these considerations, statistical design of experiments is an integral component of off-line quality-improvement studies.

A high-quality product results when key product properties have both the desired average value (target or aim) and small variation around the target (consistency or uniformity). Once a target has been determined, based on customer needs and manufacturing capabilities, quality improvement centers on achieving the target value (on the average) and on reducing variability about the target value.

One widely used Japanese quality-improvement philosophy, the *Taguchi approach* (see Chapter 12), has statistical design of experiments as its core. This off-line approach to product quality improvement integrates engineering insight with statistically designed experiments. Together these experimental strategies are used to determine process conditions for achieving the target value and to identify variables that can be controlled to reduce variation in key performance characteristics.

It is important to note that these "new" philosophies and experimental strategies have been highly promoted primarily because of the renewed emphasis on product and service quality improvement as a means to attain competitive advantage. Most of the basics of these philosophies have been advocated for many years by quality-control specialists. The fundamentals of the experimental design strategies are also well known. The reason for the new popularity of these philosophies and experimental strategies is the competitive environment of today's marketplace in many manufacturing and service industries.

The statistical design techniques discussed in this book can be used to determine desired factor settings so that a process average or a quality characteristic of key product properties are close to the target (on aim) and the variability (standard deviation plus offset from target) is as small as possible. All the procedures recommended, including screening experiments, factorial experiments, and specialized designs such as split-plot designs and response-surface designs, can be used in suitable quality-improvement settings.

The responses of interest in quality-improvement studies include both measures of product or process location and measures of dispersion. The emphasis in the discussions in this book is on the analysis of location measures.

Measures of dispersion can also be used as response variables. Note, however, the distinction between a response that is a measure of dispersion for a specific set of factor levels and a random factor effect (see Section 10.1). In the former case, several observations are made at the same levels of the factors, and a measure of dispersion (e.g., the standard deviation) is calculated and used as the response for that factor–level combination.

When dispersion measures are of interest to an investigator, $\ln s$ is often used as the response, where the standard deviation s is calculated for each of the sets of repeat observations corresponding to each of the factor–level combinations in the design. Ordinarily an equal number r of repeat test runs is required at each factor–level combination, with $r > 4$.

If both location and dispersion measures are of interest to an investigator, the two response functions (one for level and one for dispersion) can be overlaid graphically with contour plots to explore the tradeoffs in achieving optimal levels of the design factors. Many times the goal will not be to find optimal points (maxima or minima) in the response surfaces that coincide, but rather to locate flat regions (mesas or plains) that give stability of the responses. This is particularly true in product design (robustness) and process design (process control).

Some of the major benefits associated with off-line quality improvement procedures can be extended to production or process facilities in full operation. Evolutionary operation (EVOP) is used as an experimental strategy in such environments when only two or three factors can be varied at a time, and only small changes in the factor levels can be tolerated. As such, EVOP is a hybrid of on-line and off-line quality improvement techniques.

EVOP implements statistical designs on operating production or process facilities as part of the normal operations of these facilities. In this manner information for process improvement is collected by operating personnel, and normal production can continue undisturbed. Two-level factorial experiments (Chapter 5) around a center point are typically used. As operating conditions that lead to improved process characteristics are identified, the experimental region is moved to explore around this new set of conditions. This procedure is repeated until no further improvement is obtained.

The EVOP strategy has many of the desirable statistical design criteria listed in Table 4.9. However, certain risks are inherent due to exploring a limited experimental region and a small number of factors.

A consequence of the EVOP approach for process improvement is that many repeat test runs are needed at each set of factor–level combinations. This large number of repeat tests is needed because factor levels can only be changed by small amounts so that existing quality levels will not be seriously degraded at some of the factor–level settings. Because of this requirement, there is a weak "signal" (change in the response) relative to the "noise" (experimental error or process variation). This usually results in the need to collect

many observations so that the standard deviations of the statistics used to measure the effects are sufficiently small and statistically significant effects can be detected.

From this discussion, it is apparent that implementation of EVOP procedures requires a commitment by management to an extended experiment before major improvements in quality can be realized. The extended investigation is necessitated by the large number of observations that are required and the limitation to studying only two or three factors at a time. The dangers of having confounded effects and of not detecting important joint factor effects adds to the necessity for a commitment by management to an ongoing, long-lasting (preferably permanent) program.

Thus, experimentation as part of off-line quality improvement investigations or on-line ones using EVOP methodology, as well as on-line quality monitoring using control charts, can utilize many of the design and analysis techniques discussed in this book. The benefits in terms of customer satisfaction and retention, reduction of scrap or rework, and increased productivity for these quality-improvement programs well outweigh the costs and other requirements for implementation of any of these on-line or off-line procedures.

REFERENCES

Text References

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experiments*, New York: John Wiley and Sons, Inc.

Montgomery, D. C. (2000). *Design and Analysis of Experiments, Fifth Edition*, New York: John Wiley & Sons, Inc.
Chapter 1 of the previous two texts contain an introduction to the philosophy of scientific experimentation, including the role and strategy of statistical experimental design in such experimentation.

Cox, D. R. (1958). *Planning of Experiments*, New York: John Wiley and Sons, Inc.

Davies, O. L. (Ed.) (1971). *Design and Analysis of Industrial Experiments*, New York: John Wiley and Sons, Inc.
The previous two texts contain detailed discussions on the principles of statistical experimental design. Comparative experiments, randomization, sample-size considerations, blocking, the use of covariates, and factorial experiments all receive extensive treatment. Much of the discussion centers around examples.

Diamond, W. J. (2001). *Practical Experiment Designs for Engineers and Scientists, Third Edition*, New York: John Wiley and Sons. Inc.

Hocking, R. R. (1996). *Methods and Applications of Linear Models: Regression and the Analysis of Variance*, New York: John Wiley and Sons, Inc.

Natrella, M. G. (1963). *Experimental Statistics*, National Bureau of Standards Handbook 91, Washington, D.C.: U.S. Government Printing Office. (Reprinted by John Wiley and Sons, Inc.)

Chapter 11 contains short discussions on statistical considerations in the design of industrial experiments. An update of this popular text in electronic format can be found at www.itl.nist.gov/div898/handbook. The statistical software package, Dataplot, is also available at this web site.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied Linear Regression Models, Third Edition*, New York: Richard D. Irwin, Inc.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied Linear Statistical Models, Fourth Edition*, New York: Richard D. Irwin, Inc.

Ostle, B. and Malone, L. C. (1988). *Statistics in Research*, Ames, IA: The Iowa State University Press.

Schmidt, S. R. and Launsby, R. G. (2000). *Understanding Industrial Designed Experiments, 4th Edition*, Colorado Springs: Air Academy Press & Associates.

Designing for quality is specifically discussed in the following references. The above reference by Montgomery also addresses issues relating to this topic.

Box, G. E. P. and Draper, N. R. (1969). *Evolutionary Operation*, New York: John Wiley and Sons, Inc.

Gunter, B. (1987). "A Perspective on the Taguchi Method," *Quality Progress*, **20**, 44–52.

Hahn, G. J. (1976). "Process Improvement Using Evolutionary Operation," *Chemtech*, **6**, 204–206.

Hunter, J. S. (1985) "Statistical Design Applied to Product Design," *Journal of Quality Technology*, **17**, 210–221.

Myers, R. H. and Montgomery, D. C. (1995). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, New York: John Wiley & Sons, Inc.

Ross, P. J. (1996). *Taguchi Techniques for Quality Engineering, Second Edition*, Milwaukee: Quality Press.

Rhyder, R. F. (1997). *Manufacturing Process Design and Optimization*, New York: Marcel Dekker, Inc.

Scherkenbach, W. W. (1986). *The Deming Route to Quality and Productivity*, Rockville, MD: Mercury Press.

EXERCISES

1 A test program was conducted to evaluate the quality of epoxy–glass-fiber pipes. The test program required sixteen pipes, half of which were

manufactured at each of two manufacturing plants. Each pipe was produced under one of two operating conditions at one of two water temperatures. The following test conditions constituted the experimental protocol:

Run No.	Plant	Operating Conditions	Water Temp. (°F)
1	1	Normal	175
2	1	Normal	150
3	1	Severe	150
4	2	Severe	175
5	1	Normal	175
6	1	Normal	150
7	2	Normal	150
8	1	Severe	175
9	2	Normal	175
10	2	Severe	150
11	2	Normal	150
12	1	Severe	175
13	2	Severe	175
14	2	Severe	150
15	1	Severe	150
16	2	Normal	175

Identify which of the following statistical design features are included in the test program:

(a) Factor(s)
(b) Factor levels
(c) Blocks
(d) Experimental units
(e) Repeat tests
(f) Replications
(g) Test runs
(h) Covariate(s)

2 An accelerator for the hydration and hardening of cement is being developed to aid in the structural construction and integrity of bridge trusses. Two mixes of fast-setting cement were tested by measuring the compressive strength (psi) of five 2-inch cubes of mortar, 24 hours after setting. This experiment was conducted twice and the following measurements of compressive strength were determined:

	Compressive Strength (psi)	
Experiment	**Mix A**	**Mix B**
1	2345	2285
	2119	2143
	2150	2362
	2386	2419
	2297	2320
2	2329	2387
	2310	2345
	2297	2352
	2311	2360
	2314	2349

Identify which of the statistical design features (a) to (h) of Exercise 1 occur in this experiment.

3 An experiment is to be designed to study the effects of several factors on the quality of electronic signals transmitted over telephone lines. The factors of interest are the baud rate (speed of transmission), signal intensity, and type of transmission device. The project leader elects to investigate baud rates from 300 to 9600 bps in increments of 100 bps, three different signal intensities, and two of three transmission devices. The use of only two transmission devices is necessitated by the cost of the project. The projected costs for the experimentation are estimated to be $500 for each baud rate, $2000 for each signal intensity, and $10,000 for each transmission device. These costs are primarily for equipment, personnel, and data processing and are largely unaffected by the number of test runs in the experiment. Identify the advantages and disadvantages of the proposed experiment. Suggest remedies for the disadvantages.

4 Automobile emission-control systems are known to deteriorate as a vehicle is used. A concern in many experiments is that emission results may be affected by how far test vehicles have previously been driven prior to being tested. Below are carbon monoxide (CO) emissions for several test runs (under identical conditions) for each of three automobiles. Also recorded are the odometer mileages (MILES) at the start of each test run. Use a scatterplot with different symbols for each vehicle to assess whether initial mileage should be used as a covariate in a test program involving these three vehicles.

Vehicle 1		Vehicle 2		Vehicle 3	
MILES	CO	MILES	CO	MILES	CO
6814	1.36	7155	2.12	6842	1.33
6843	1.39	7184	2.01	6879	1.45
6942	1.68	7061	1.83	6952	1.54
6970	1.61	7091	1.81	6991	1.70
6885	1.52			7027	1.98
6913	1.29			7056	1.92
7059	1.80				

5 Water flows through tubes in a fluidized-bed combustor, where it is converted to steam. The bed of the combustor consists of sorbent particles, which are used to absorb pollutants given off during combustion. The particles are suspended in the combustor by a gas flowing up from the bottom of the particle bed. The problem under investigation concerns the corrosion effects of the exterior surface of the tubes used to carry the water through the pipes. The following factors may influence the corrosion effects:

Factor	Levels
Bed temperature (°C)	700, 1000
Tube temperature below the bed temperature (°C)	0, 200, 400
Particle size (μm)	1000, 3000, 4000
Environment	Oxidizing, sulfidizing
Particle material	Limestone, dolomite
Tube material	Stainless steel, iron

Calculate the number of test runs needed for all possible combinations of the factor levels to be included in the experiment. How many of these test runs are used in calculating the average corrosion measurements for each of the above factor levels? Suppose each factor was investigated in a separate experiment using the same number of test runs just computed for each level. What would the total experiment size be for the separate investigations of each factor?

6 For the experiment in Exercise 2, calculate the averages and the sample standard deviations of the compressive strengths for each mix in each of the

two experiments. From the averages just calculated, calculate the effect of the two mixes, separately for each experiment. What tentative conclusions can be drawn from the calculated effect of the mixes for each of the experiments? Do the sample standard deviations affect the conclusions? Why (not)?

7 Spectral properties of a compound mixed in a chemical solution are such that measurements of the refraction of light passing through the solution are about 5% greater than for the solution without the presence of the compound. Below are 10 measurements of light refraction (ratio of the angle of incidence to the angle of refraction) for the chemical solution with and without the compound. What do these data suggest about the ability of this procedure to detect the compound?

Solution without Compound	Solution with Compound
1.41	1.55
1.39	1.62
1.50	1.88
1.47	1.92
1.42	1.59
1.48	1.91
1.43	1.30
1.45	1.71
1.45	1.26
1.48	1.41

8 Sketch two geometrically different types of contours plots for which one-factor-at-a-time testing will lead to a result near the vicinity of the optimal response value within an appropriate experimental region. How could these contours be modified so that the same test sequence leads to a value that is not optimal in the experimental region?

9 A new prototype diesel engine is being tested by its manufacturer to determine at what level speed (rpm) and load (lb-ft) the brake-specific fuel consumption (BFSC, lb/bph-hr) is minimized. Suppose that the BSFC response surface for this engine can be modeled as follows:

$$\begin{aligned}\text{BSFC} = {} & 0.9208 - 0.0016 \times \text{load} - 0.0003 \times \text{speed} \\ & + 3.1164 \times 10^{-6} \times (\text{load})^2 + 8.3849 \times 10^{-8} \times (\text{speed})^2 \\ & - 2.0324 \times 10^{-7} \times \text{load} \times \text{speed}.\end{aligned}$$

In this model, the engine operates at speeds from 1500 to 2800 rpm and loads from 100 to 500 lb-ft. Using this model, calculate responses for a

one-factor-at-a-time experiment to find the levels of load and speed that minimize BSFC. Plot rough contours of this surface (e.g., plot response values for a grid of speed and load values, and interpolate between the plotted points). Does the one-factor-at-a-time experiment lead to the vicinity of the optimal speed–load combination?

10 Refer to the engine experiment of Exercise 9. Another response of interest in this experiment is the "knock" produced by pressure changes in the engine. Suppose the response surface for this response is

$$\begin{aligned}\text{knock} = &-525.5132 + 0.2642 \times \text{speed} + 3.6783 \times \text{load} \\ &- 0.0015 \times 10^{-2} \times (\text{speed})^2 - 0.2278 \times 10^{-2} \times (\text{load})^2 \\ &- 0.0975 \times 10^{-2} \times \text{load} \times \text{speed}.\end{aligned}$$

Repeat the procedures requested in Exercise 9 for the minimization of knock. Note from the contour plots that the minimization of BSFC and the minimization of knock do not lead to the same optimal speed–load combination. Overlay the two sets of contour plots to determine a reasonable compromise for speed and load that will come close to minimizing both responses.

11 Intake valve seat data were collected on an automobile engine to evaluate the effects of load condition and the valve temperature build condition. A high and low value of load condition was to be run with both a high and low value of the valve temperature. For each of these four test combinations, the vehicle was run on an outdoor dynamometer through two complete driving cycles. The data from each of the four valves in the engine were analyzed independently. Although four temperature positions were measured on each valve, the average valve seat temperature at the four valve positions was used in the analysis. Because ambient temperature may influence the valve temperature response, it was also measured in the experiment. Identify which of the statistical design features (a) to (h) of Exercise 1 occur in this experiment.

12 A study was run to obtain estimates of the internal corrosion rates in inaccessible offshore pipeline locations. The test matrix consisted of 11 exposures of carbon steel specimens in slowly flowing aqueous solutions saturated with gas mixtures containing H_2S, CO_2, and oxygen gases at various pressures. Chloride also was utilized because it was known to accelerate the pitting corrosion in stagnant solutions. The test matrix is shown below.

Test Number	CO_2 (psi)	H_2S (psi)	Oxygen (ppmv)	Chloride (wt.%)
1	10.0	0.0	0	1.0
2	0.0	0.5	0	1.0
3	10.0	0.5	0	1.0
4	0.0	0.0	100	1.0
5	10.0	0.0	100	1.0
6	0.0	0.5	100	1.0
7	10.0	0.5	100	1.0
8	0.0	0.0	1000	0.0
9	10.0	0.0	1000	0.0
10	0.0	0.5	1000	0.0
11	10.0	0.5	1000	0.0

Test conditions were selected to be generally representative of wet conditions within pipelines where slowly flowing liquids are present. Review the above experiment and list any flaws you see in this experimental design. If any are found, determine how you might correct them.

13 Fuel consumption data were collected on a lab-based engine in which three fuel-economy device types and three engine speeds were used to obtain nine test conditions. The three devices included a baseline type and two prototypes labeled A and B. The three engine speeds were set at 1500, 3000, and 4500 rpm. Each of the test combinations was run in triplicate. Identify which of the statistical design features (a) to (h) of Exercise 1 occur in this experiment. Also, determine the number of independent test runs needed for this experiment.

14 In Section 4.3.4, randomization was discussed as a way to protect against unknown sources of possible bias. Explain how you would use randomization in the design in Exercise 13. List at least one possible source of variation that might effect the fuel consumption measurements.

15 Back-to-back repeats occur in an experiment when the repeat test runs for a given set of factor combinations are made immediately after one another. Random repeats occur when the repeat test runs are assigned in a random sequence just as are the remainder of the other test runs. Discuss the pros and cons of using back-to-back repeat test runs versus using random repeat test runs in a given experiment.

16 In some experimental factor settings for the various factor combinations cannot be exactly obtained. For example, in blending gasoline to use in studying the effects of fuel properties on automobile emissions, target fuel properties are generally established for each test fuel. However, when the fuels are actually blended, the target values may be slightly missed,

and, thus, all that can be used is the average value obtained from several laboratory measurements of each fuel property. An example set of target and actual values for one such fuel property, cetane number, is given below.

Fuel	Target Cetane Number	Actual Cetane Number
1	42	42.8
2	47	48.0
3	52	53.2
4	42	42.4
5	47	47.7
6	52	53.0

A question arises as to whether to use the target values in the data analysis or the actual values for the controlled factors. Explain what you would do in this situation, and why you would choose to do it.

17 Uncontrolled factors often occur in experiments. In the fuel study in Exercise 16, suppose the test fuels were evaluated by running them in the same engine in a laboratory. Name some possible uncontrolled factors in this experiment. What should the experimenter do with these factors to minimize their effects on the automobile emission measurements?

18 Suppose the time and resources are not available to take the two repeat test runs for each of the 8 factor–level combinations for the experiment described in Exercise 2. Instead, it is only possible to conduct 4 additional tests for a total of 12 rather than 16 test runs. Discuss how you might assign these four extra test runs to the various factor–level combinations.

19 For the experiment in Exercise 18 of Chapter 3, calculate the averages and standard deviations for the viscosity of synthetic rubber for the manufacturer's lab and the customer's lab. Compute the effects of the two labs, separately. What tentative conclusions can be drawn from the calculated effect of the two labs? Do the standard deviations affect the conclusions? Explain your answer.

20 An experiment was run to determine if a set of randomly chosen freshmen college students could distinguish between five different brands of cola. The five brands of cola included Coca-Cola, Pepsi, RC Cola, a generic cola from a local grocery store, and Sam's Brand cola. The response variable was a score (from 1 to 10) assigned by each student to the sweetness, flavor, and fizz attributes of each cola. Discuss the design implications (that is, the effects on the analysis or possible conclusions for each design) of choosing 25 students for the experiment and assigning five to taste and evaluate each cola versus having each student taste and evaluate all five colas.

CHAPTER 5

Factorial Experiments in Completely Randomized Designs

The most straightforward statistical designs to implement are those for which the sequencing of test runs or the assignment of factor combinations to experimental units can be entirely randomized. In this chapter we introduce completely randomized designs for factorial experiments. Included in this discussion are the following topics:

- *completely randomized designs, including the steps needed to randomize a test sequence or assign factor–level combinations to experimental units;*
- *factorial experiments, the inclusion of all possible factor–level combinations in a design; and*
- *calculation of factor effects as measures of the individual and joint influences of factor levels on a response.*

Experiments are conducted to investigate the effects of one or more factors on a response. When an experiment consists of two or more factors, the factors can influence the response individually or jointly. Often, as in the case of one-factor-at-a-time experimentation, an experimental design does not allow one to properly assess the joint effects of the factors.

Factorial experiments conducted in completely randomized designs are especially useful for evaluating joint factor effects. Factorial experiments include all possible factor–level combinations in the experimental design.

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

Completely randomized designs are appropriate when there are no restrictions on the order of testing, and/or when all the experimental units to be used in the experiment can be regarded as homogeneous.

In the first two sections of this chapter factorial experiments, completely randomized designs, and joint factor effects are characterized. The construction of completely randomized designs, including the randomization procedures, are detailed. The last two sections of this chapter detail the calculation of factor effects. The calculation of factor effects is shown to be a valuable aid in interpreting the influence of factor levels on a response.

5.1 FACTORIAL EXPERIMENTS

A (complete) factorial experiment includes all possible factor–level combinations in the experimental design. Factorial experiments can be conducted in a wide variety of experimental designs. One of the most straightforward designs to implement is the *completely randomized design.*

In a completely randomized design all the factor–level combinations in the experiment are randomly assigned to experimental units, if appropriate, or to the sequence of test runs. Randomization is important in any experimental design because an experimenter cannot always be certain that every major influence on a response has been included in the experiment. Even if one can identify and control all the major influences on a response, unplanned complications are common. Instrument drift, power surges, equipment malfunctions, technician or operator errors, or a myriad of other undetectable influences can bias the results of an experiment.

Randomization does not prevent any of the above experimental complications from occurring. As mentioned in the last chapter, randomization affords protection from bias by tending to average the bias effects over all levels of the factors in the experiment. When comparisons are made among levels of a factor, the bias effects will tend to cancel and the true factor effects will remain. Randomization is not a guarantee of bias-free comparisons, but it is certainly inexpensive insurance.

There are numerous ways to achieve randomization in an experimental design. Any acceptable randomization procedure must, however, adhere to procedures that satisfy the definition given in Exhibit 5.1. With this definition of randomization, the steps used to construct a completely randomized design are given in Exhibit 5.2. Note that the randomization procedure described in Exhibit 5.2 is equivalent to obtaining a simple random sample without replacement (Section 1.2) of the integers 1 to N.

EXHIBIT 5.1

Randomization. Randomization is a procedure whereby factor–level combinations are (a) assigned to experimental units or (b) assigned to a test sequence in such a way that every factor–level combination has an equal chance of being assigned to any experimental unit or position in the test sequence.

EXHIBIT 5.2 CONSTRUCTION OF FACTORIAL EXPERIMENTS IN COMPLETELY RANDOMIZED DESIGNS

1. Enumerate all factor–level combinations. Include repeat tests.
2. Number the factor–level combinations, including any repeat tests, sequentially from 1 to N.
3. From a random-number table (e.g., Table A1 of the Appendix) or from a computer-generated sequence of random numbers, obtain a random sequence of the integers 1 to N.
4. Assign the factor–level combinations to the experimental units or conduct the test runs in the order specified by the random number sequence. If both a randomized assignment to experimental units and a randomized test sequence are to be included in the design, use two random number sequences.

Figure 5.1 lists factors that are of interest in a laboratory investigation of torque forces experienced by rotating shafts. In this experiment a rotating shaft is to be braced by a stationary cylindrical sleeve, and lubricants are to

Factor	*Levels*
Shaft alloy	Steel, aluminum
Sleeve metal	Porous, nonporous
Lubricant type	Lub 1, Lub 2, Lub 3, Lub 4

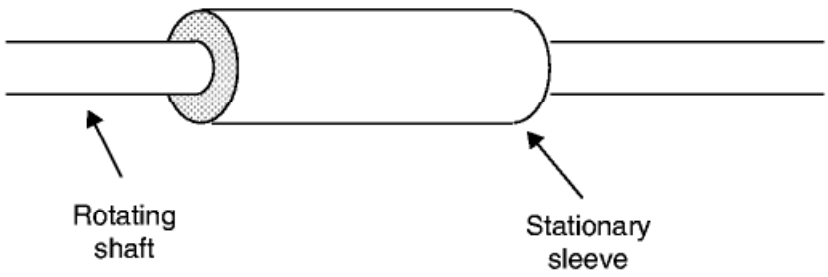


Figure 5.1 Factors for torque experiment.

be applied to the inner wall of the sleeve to reduce the amount of friction between the shaft and the sleeve. The investigators wish to study shafts made of steel and aluminum alloys. The sleeves are to be made of two types of metal, one porous and one nonporous, to determine whether the lubricants are more effective when they adhere to the sleeve. Four lubricants of widely differing properties are to be selected from among many commercial brands available.

This experiment is to be conducted in a laboratory using a small-scale version of the shaft–sleeve apparatus found in large industrial equipment. It is feasible to consider experimental designs that include all possible combinations of the factors. Complete randomization of the run order is also feasible for this laboratory experiment. Thus, a completely randomized factorial experiment is appropriate.

Table 5.1 lists all of the sixteen possible factor–level combinations for the factors from Figure 5.1. Suppose that row 20 and column 3 is selected as a starting point in the random-number table, Table A1 in the Appendix. Numbers between 1 and 16 can then be selected as consecutive two-digit numbers (proceeding left to right across each row), ignoring all numbers greater than 16 (this is only one of many ways to use the table). The following random test sequence is then obtained:

8, 13, 4, 7, 5, 1, 11, 15, 9, 3, 12, 10, 6, 14, 16, 2.

TABLE 5.1 Factor–Level Combinations for Torque Study

Combination Number	Shaft Alloy	Sleeve Metal	Lubricant Type
1	Steel	Porous	Lub 1
2	Steel	Porous	Lub 2
3	Steel	Porous	Lub 3
4	Steel	Porous	Lub 4
5	Steel	Nonporous	Lub 1
6	Steel	Nonporous	Lub 2
7	Steel	Nonporous	Lub 3
8	Steel	Nonporous	Lub 4
9	Aluminum	Porous	Lub 1
10	Aluminum	Porous	Lub 2
11	Aluminum	Porous	Lub 3
12	Aluminum	Porous	Lub 4
13	Aluminum	Nonporous	Lub 1
14	Aluminum	Nonporous	Lub 2
15	Aluminum	Nonporous	Lub 3
16	Aluminum	Nonporous	Lub 4

TABLE 5.2 Randomized Test Sequence for Torque Study

Run Number	Combination Number	Shaft Alloy	Sleeve Metal	Lubricant Type
1	8	Steel	Nonporous	Lub 4
2	13	Aluminum	Nonporous	Lub 1
3	4	Steel	Porous	Lub 4
4	7	Steel	Nonporous	Lub 3
5	5	Steel	Nonporous	Lub 1
6	1	Steel	Porous	Lub 1
7	11	Aluminum	Porous	Lub 3
8	15	Aluminum	Nonporous	Lub 3
9	9	Aluminum	Porous	Lub 1
10	3	Steel	Porous	Lub 3
11	12	Aluminum	Porous	Lub 4
12	10	Aluminum	Porous	Lub 2
13	6	Steel	Nonporous	Lub 2
14	14	Aluminum	Nonporous	Lub 2
15	16	Aluminum	Nonporous	Lub 4
16	2	Steel	Porous	Lub 2

Assigning this test order to the factor–level combinations in Table 5.1 results in the completely randomized design shown in Table 5.2.

Experimental designs should, whenever possible, include repeat test runs. Repeat test runs, as discussed in Section 4.1, are two or more experimental tests or observations in which the factor–level combinations are identical. Responses from repeat tests exhibit variability only from experimental sources of variation such as variation due to changes in test conditions and to measurement error. The sample standard deviations from repeat tests can be used to estimate the uncontrolled experimental variability.

An alternative to the inclusion of repeat tests in an experiment occurs when some factor interactions (Section 5.2) can be assumed to be zero. It is often true that experiments involving large numbers of factors have interactions that can be assumed to be zero or negligible relative to the uncontrolled error variation. In such circumstances the statistics that would ordinarily measure the interactions of these factors can be used to measure experimental variability. Fractional factorial and screening experiments (Chapter 7) also exploit this assumption to reduce the number of test runs needed to investigate all the design factors.

As mentioned above, when possible, repeat tests should be included in the design of an experiment. Even if fractional factorial or screening experiments are used, the inclusion of several repeat tests allows one to (a) estimate

uncontrolled experimental-error variation and (b) investigate the adequacy of the fitted model (Section 15.2) without having to make the assumption that some of the interactions are zero. It is especially important to include repeat tests if one is unsure about the validity of an assumption that interactions are zero.

The inclusion of repeat tests in an experimental design does not necessarily require that each factor–level combination be repeated or that each be repeated an equal number of times. When possible, *balance* — the inclusion of all factor–level combinations an equal number of times — should be the goal in a complete factorial experiment. The only requirement is that a sufficient number of repeat tests be included so that a satisfactory estimate of experimental error can be obtained. In general, it is unwise to select a single factor–level combination and repeat it several times. A better approach when all combinations cannot be repeated in an experiment is to randomly select, without replacement, the factor–level combinations to be repeated. As a rough rule of thumb, a minimum of six repeat tests should be included to estimate the experimental error standard deviation.

If one wishes to have two repeat tests for each factor level in the torque study, one will list each factor–level combination twice. Then random numbers will be obtained until two of each combination are included. This random number sequence then dictates the run order. One such sequence produced the run order shown in Table 5.3.

Several benefits accompany the factorial experiments shown in Tables 5.2 and 5.3. First, the inclusion of each factor level with a variety of levels of other factors means that the effects of each factor on the response are investigated under a variety of different conditions. This allows more general conclusions to be drawn about the factor effects than if each factor effect were studied for a fixed set of levels of the other factors.

A second benefit of these designs is that the randomization protects against unknown biases, including any unanticipated or unobservable "break-in" effects due to greater or lesser care in conducting the experiment as it progresses. Note too that the randomization of the repeat tests in Table 5.3 ensures that responses from repeat tests give a valid estimate of the experimental error of the test runs. If "back-to-back" repeat tests are conducted, the estimate of experimental error can be too small because any variability associated with setting up and tearing down the equipment would not be present.

A third major benefit of factorial experiments conducted in completely randomized designs is the ability to investigate joint factor effects. There are joint factor effects among the factors in the torque study. The importance of planning experiments so that joint factor effects can be measured is the topic of discussion in the next section.

TABLE 5.3 Randomized Test Sequence for Torque Study Including Repeat Tests

Run Number	Combination Number	Shaft Alloy	Sleeve Metal	Lubricant Type
1	4	Steel	Porous	Lub 4
2	2	Steel	Porous	Lub 2
3	7	Steel	Nonporous	Lub 3
4	16	Aluminum	Nonporous	Lub 4
5	10	Aluminum	Porous	Lub 2
6	4	Steel	Porous	Lub 4
7	11	Aluminum	Porous	Lub 3
8	12	Aluminum	Porous	Lub 4
9	8	Steel	Nonporous	Lub 4
10	16	Aluminum	Nonporous	Lub 4
11	7	Steel	Nonporous	Lub 3
12	8	Steel	Nonporous	Lub 4
13	12	Aluminum	Porous	Lub 4
14	3	Steel	Porous	Lub 3
15	5	Steel	Nonporous	Lub 1
16	11	Aluminum	Porous	Lub 3
17	5	Steel	Nonporous	Lub 1
18	1	Steel	Porous	Lub 1
19	10	Aluminum	Porous	Lub 2
20	9	Aluminum	Porous	Lub 1
21	1	Steel	Porous	Lub 1
22	9	Aluminum	Porous	Lub 1
23	15	Aluminum	Nonporous	Lub 3
24	15	Aluminum	Nonporous	Lub 3
25	6	Steel	Nonporous	Lub 2
26	13	Aluminum	Nonporous	Lub 1
27	13	Aluminum	Nonporous	Lub 1
28	6	Steel	Nonporous	Lub 2
29	3	Steel	Porous	Lub 3
30	14	Aluminum	Nonporous	Lub 2
31	2	Steel	Porous	Lub 2
32	14	Aluminum	Nonporous	Lub 2

5.2 INTERACTIONS

Interaction means the presence of joint factor effects (see Exhibit 5.3). The definition in Exhibit 5.3 implies that the presence of interactions precludes an assessment of the effects of one factor without simultaneously assessing the effects of other factors. This is the essence of an interaction effect: when

interactions occur, the factors involved cannot be evaluated individually. The use of the term *interaction* parallels its use in other contexts. For example, a drug interaction occurs when one pharmaceutic administered in combination with or shortly after another drug alters the effect of one or both drugs.

EXHIBIT 5.3

Interaction. An interaction exists among two or more factors if the effect of one factor on a response depends on the levels of other factors.

To better understand the implications of interaction effects, three examples will now be presented. The first example is based on the data shown in Table 5.4. This table presents data on the life (in hours) of certain cutting

TABLE 5.4 Cutting-Tool Life Data*

Tool Life (hr)	Lathe Speed (rpm)	Tool Type
18.73	610	A
14.52	950	A
17.43	720	A
14.54	840	A
13.44	980	A
24.39	530	A
13.34	680	A
22.71	540	A
12.68	890	A
19.32	730	A
30.16	670	B
27.09	770	B
25.40	880	B
26.05	1000	B
33.49	760	B
35.62	590	B
26.07	910	B
36.78	650	B
34.95	810	B
43.67	500	B

*Data from Montgomery, D. C., Peck, E. A., and Vining, G. G. (2001). *Introduction to Linear Regression Analysis*, Third Edition, New York: John Wiley & Sons, Inc. Copyright 2001 by John Wiley & Sons, Inc. Used by permission.

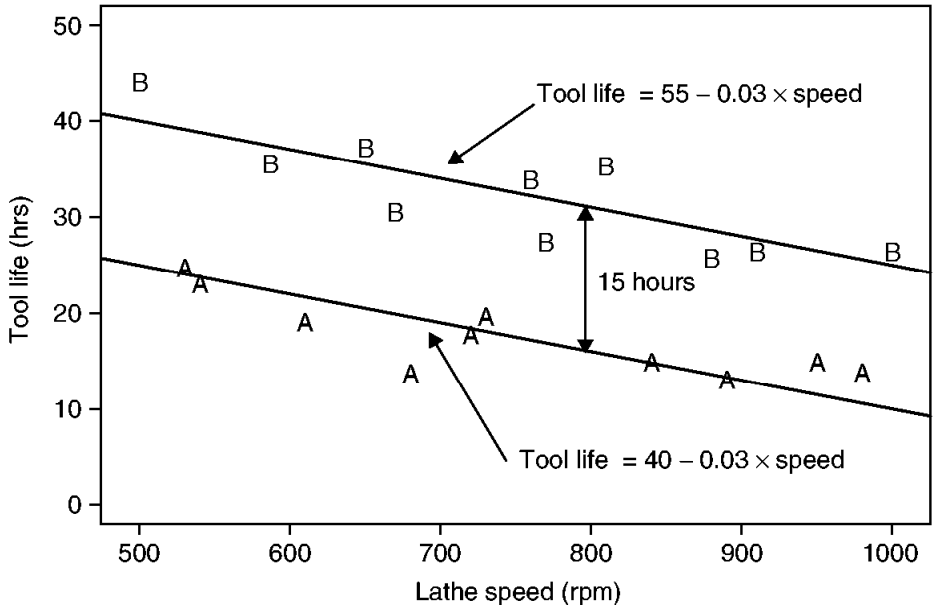


Figure 5.2 Effects of tool type and lathe speed on tool life. (Plotting symbol is the tool type.)

tools used with lathes. There are two types of cutting tools of interest in this example, labeled type A and type B. Data on the lifetimes of these cutting tools are collected at many different operating speeds of the lathe. Figure 5.2 is a plot of tool life versus lathe speed for each tool type. The plotting symbol denotes the tool type (A = Type A, B = Type B). The labeling of the plotted points is an informative variant of scatterplots (Section 1.1). *Labeled scatterplots* allow the inclusion of information on an additional variable (here, tool type) without increasing the dimensionality of the graph.

A statistical analysis of these data using covariance analysis (Section 16.2) reveals that the data for each tool type can be well fitted using straight lines with equal slope. The intercepts for the two lines differ by about 15 hours. These fits are superimposed on the plotted points in Figure 5.2.

For these data, the two variables lathe speed and tool type do not interact. The effect of tool type is the same at all lathe speeds: Tool B lasts 15 hours longer, on the average, than tool A at all lathe speeds. Likewise, the effect of lathe speed is the same for both tool types: increasing lathe speed decreases the tool life by approximately 0.03 hours per rpm for both tool types. The lack of interaction is graphically indicated by the parallel lines in Figure 5.2.

Now consider the data shown in Figure 5.3. These data are average chemical yields of a process in an investigation conducted on a pilot plant. The factors of interest in this study are the operating temperature of the plant, the type of catalyst used in the process, and the concentration of the catalyst.

Figure 5.4 contains alternative plots of the chemical-yield averages. In these plots the averages for each concentration and each temperature are plotted separately for each catalyst. We join the averages for each catalyst in Figure 5.4

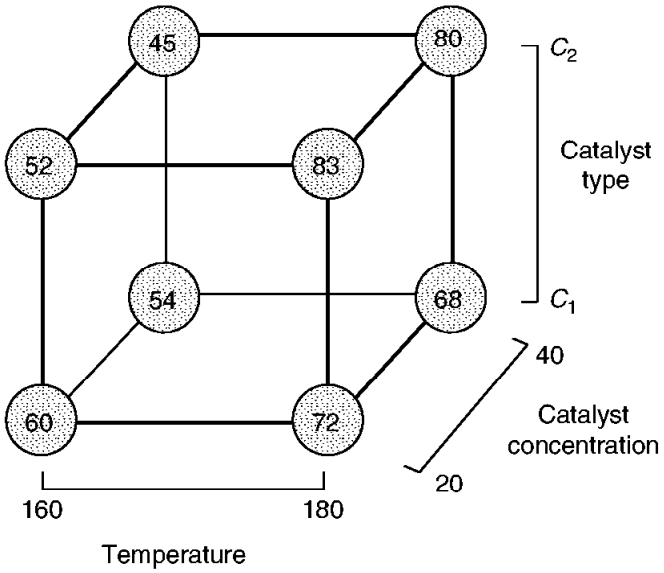


Figure 5.3 Average chemical yields for pilot-plant experiment. Data from Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*, New York: John Wiley & Sons, Inc. Used by permission.

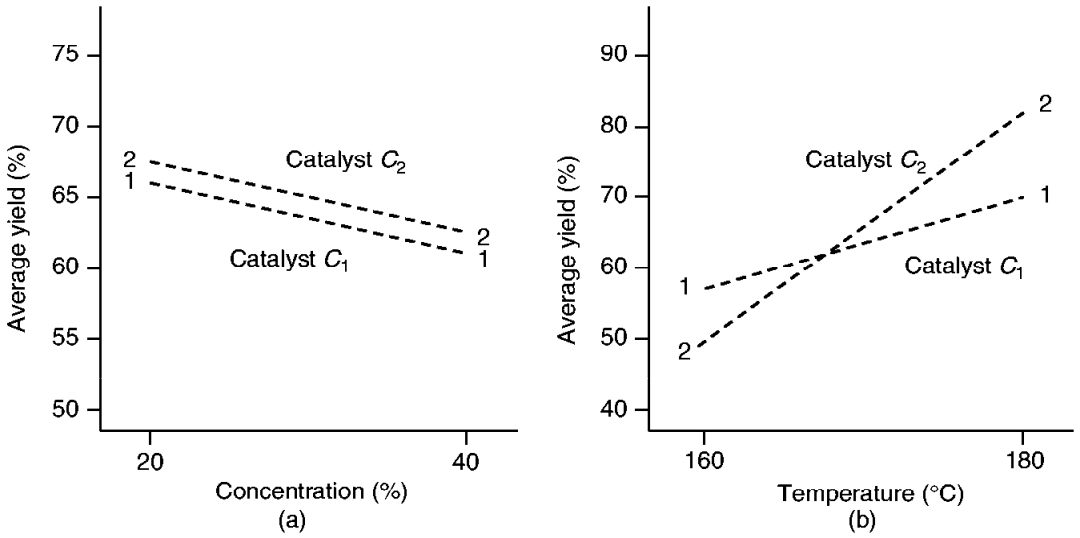


Figure 5.4 Interaction plots for chemical yield data. (a) Catalyst and concentration. (b) Catalyst and temperature.

by dashed lines for visual emphasis. The reader should not infer that there is a straight-line relationship between yield and concentration (Figure 5.4a) or temperature (Figure 5.4b) for each catalyst. The lines simply highlight the change in average yield for each catalyst as a function of concentration or temperature levels. Dashed rather than solid lines are used as a reminder that the relationships are not necessarily linear.

The visual impression left by Figure 5.4a is similar to that of Figure 5.2 and is the reason for connecting the averages by dashed lines. The effect of concentration on average yield, as suggested by the parallel lines, is the same for each catalyst. An increase from 20 to 40% in concentration produces a decrease of approximately 5 grams, regardless of the catalyst used. Thus, concentration has approximately the same effect on yield for each catalyst, and vice versa. There is no two-factor interaction between catalyst and concentration.

Contrast these results with the trends indicated in Figure 5.4b. The average chemical yield increases 13 grams for catalyst C_1 as the temperature is increased from 160 to 180°C; however, the chemical yield increases much more dramatically for catalyst C_2 as temperature is increased. In fact, the higher yield at 160°C occurs with catalyst C_1, but at 180°C the higher yield occurs with catalyst C_2.

Figure 5.4b is a graphical representation of an interaction. As in Figure 5.4a, the dashed lines connecting the plotted points are used for visual emphasis only, not to suggest a linear effect of temperature. With this understanding, the visual impression of Figure 5.4b is that the effect on yield of changing temperature from 160 to 180°C depends on which catalyst is used. In addition, the optimum (higher) yield depends on both the temperature and the catalyst. If the plant is operated at 160°C, catalyst C_1 produces a higher average yield. If the plant is operated at 180°C, catalyst C_2 produces a higher average yield. Consequently, one cannot draw conclusions about either of these factors without specifying the level of the other factor. Even though catalyst C_2 has a higher average yield (65) than catalyst C_1 (63.5), it would be misleading to conclude that catalyst C_2 is always preferable to catalyst C_1. The preference for one of the catalysts depends on the operating temperature of the plant.

In general, interactions are indicated graphically by nonparallel trends in plots of average responses. The dashed line segments connecting the average responses for each level of the factors need not intersect as they do in Figure 5.4b. Any nonparallel changes in the average responses is an indication of an interaction. As with all statistical analyses, one confirms the existence (or absence) of interaction effects implied by graphical displays with formal statistical inference procedures that account for experimental variation. Thus, the lack of parallelism in the changes of the average responses must be sufficiently large that they represent a real factor effect, not just uncontrolled experimental variability.

The final example of interaction effects is taken from the data generated from the torque study of the last section. Torque measurements were obtained using the test sequence of the completely randomized design in Table 5.3. The torque measurements were averaged over the repeat runs, and the averages are plotted in Figure 5.5.

In Figure 5.5a the average torque measurements for the aluminum shaft are plotted as a function of the sleeve and lubricant used. The dashed line segments

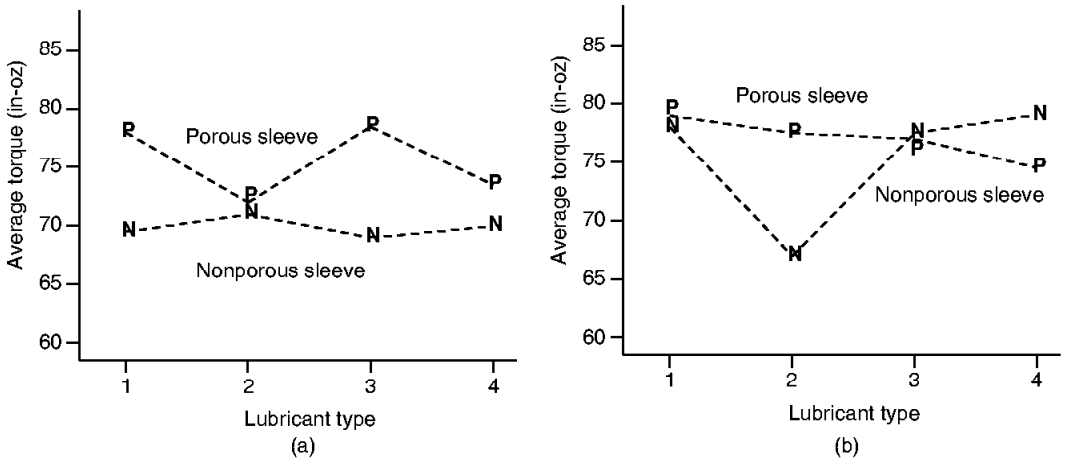


Figure 5.5 Interaction plot for torque study. (a) Aluminum shaft. (b) Steel shaft.

connecting the points are not parallel, suggesting the presence of an interaction between sleeve and lubricant on the average torque. Both the porous and the nonporous sleeves produce average torque measurements of approximately the same magnitude when lubricant 2 is used. The average torque measurements differ greatly between sleeve types for each of the other lubricants.

In Figure 5.5b the average torque measurements for the steel shaft are plotted. There is a strong indication of an interaction between the sleeve and lubricant factors; however, the trends in this plot are not similar to those in Figure 5.5a. As can be confirmed by a statistical analysis of the averages, there is a three-factor interaction among the three design factors. With the aluminum shaft, low average torque measurements are achieved with all the lubricants when the nonporous sleeve is used. Contrast this result with the substantially lower average torque measurement for lubricant 2 and the nonporous sleeve than for the other lubricant–sleeve combinations with the steel shaft. Moreover, if lubricant 4 is used, the porous sleeve yields lower average torque measurements than does the nonporous sleeve with the steel shaft.

From these examples it should be clear that the presence of interactions requires that factors be evaluated jointly rather than individually. It should also be clear that one must design experiments to measure interactions. Failure to do so can lead to misleading, even incorrect, conclusions. Factorial experiments enable all joint factor effects to be estimated. If one does not know that interaction effects are absent, factorial experiments should be seriously considered.

The graphical procedures illustrated in this section allow a visual inspection of factor effects. Statistical analyses, especially interval estimation and hypothesis testing, utilize numerical estimates of the factor effects. In the next section, calculations of factor effects are detailed. These calculations supplement a graphical assessment of interactions and are useful in quantifying the influence that changing factor levels have on average response values.

Because of the possible existence of interactions, complete factorial experiments should be conducted whenever possible. In experiments with a large number of factors, however, complete factorial experiments may not be economically feasible even if it is believed that some interactions among the factors may exist. In such settings, fractional factorial and screening experiments are important alternatives. Designs for these experiments are discussed in Chapter 7.

5.3 CALCULATION OF FACTOR EFFECTS

An effect was defined in Table 4.1 as a change in the average response corresponding to a change in factor–level combinations or to a change in experimental conditions. We now wish to specify more completely several types of effects for factors that have only two levels (see Exhibit 5.4). In the appendix to this chapter we generalize these effects to factors whose numbers of levels are greater than two, with special emphasis on factors whose numbers of levels are powers of two. The calculation of effects is facilitated by the introduction of algebraic notation for individual responses and averages of responses. This notation is also helpful in clarifying the concepts of confounding and design resolution that will be presented in Chapter 7.

We denote factors in designed experiments by uppercase Latin letters: $A, B, C, \ldots, K$. An individual response is represented by a lowercase Latin letter, usually y, having one or more subscripts. The subscripts, one for each factor, designate the specific factor–level combination from which the response is obtained. There may also be one or more subscripts designating repeat observations for fixed factor levels. For example, in the torque study described in Figure 5.1 the factors and their levels could be designated as follows:

Factor	Factor Symbol	Level	Subscript Symbol
Alloy	A	Steel	$i = 1$
		Aluminum	$i = 2$
Sleeve	B	Porous	$j = 1$
		Nonporous	$j = 2$
Lubricant	C	Lub 1	$k = 1$
		Lub 2	$k = 2$
		Lub 3	$k = 3$
		Lub 4	$k = 4$

A fourth subscript l could be used to denote the repeat-test number, provided there is at least one repeat test; otherwise, the fourth subscript is superfluous and not included.

With this convention in notation, responses from a three-factor experiment with r repeat tests per factor–level combination would be denoted y_{ijkl} with $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, c$, and $l = 1, \ldots, r$. The upper limits a, b, c, and r on the subscripts denote the numbers of levels of the three factors A, B, C and the number of repeat observations, respectively.

EXHIBIT 5.4 EFFECTS FOR TWO-LEVEL FACTORS

Main effect. The difference between the average responses at the two levels of a factor.

Two-factor interaction. Half the difference between the main effects of one factor at the two levels of a second factor.

Three-factor interaction. Half the difference between the two-factor interaction effects at the two levels of a third factor.

Average responses are represented by replacing one or more of the subscripts by a dot and placing a bar over the symbol for the response. Using dot notation,

$$\bar{y}_{\bullet\bullet\bullet\bullet} = n^{-1} \sum_{ijkl} y_{ijkl}$$

denotes the overall average response for a three-factor experiment with repeat tests, where $n = abcr$ and the summation is over all the observations (all possible values of the subscripts). Similarly,

$$\bar{y}_{i\bullet\bullet\bullet} = (bcr)^{-1} \sum_{jkl} y_{ijkl}$$

is the average response for all responses having the ith level of factor A. The symbol $\bar{y}_{2\bullet1\bullet}$ denotes the average response for all test runs having the second level of the first factor and the first level of the third factor. We denote by $\bar{y}_{ijk\bullet}$ a typical average response, across repeat observations, for one of the factor–level combinations.

Consider now a factorial experiment with each factor having two levels and r repeat observations for each combination of levels of the factors. The factor–level combinations in the experiment can be represented in any of several equivalent ways. An especially useful way to represent the factor–level combinations is using the *effects representation*, which we now describe.

Let one of the levels of a factor be coded -1 and the other level be coded $+1$. It is arbitrary which level receives each code, although it is customary for quantitative factors to let the lower level be -1. A straightforward way to list all the unique combinations of the factor levels (ignoring repeats) using this coding is given in Exhibit 5.5.

EXHIBIT 5.5 EFFECTS CODING OF FACTOR LEVELS

1. Designate one level of each factor as -1 and the other level as $+1$.
2. Lay out a table with column headings for each of the factors $A, B, C, \ldots, K$.
3. Let $n = 2^k$, where k is the number of factors.
4. Set the first $n/2$ of the levels for factor A equal to -1 and the last $n/2$ equal to $+1$. Set the first $n/4$ levels of factor B equal to -1, the next $n/4$ equal to $+1$, the next $n/4$ equal to -1, and the last $n/4$ equal to $+1$. Set the first $n/8$ of the levels for factor C equal to -1, the next $n/8$ equal to $+1$, etc. Continue in this fashion until the last column (for factor K) has alternating -1 and $+1$ signs.

Table 5.5 shows the effects representation for a two-factor factorial experiment. Observe that each time one of the factor levels is coded -1 its actual value is its "lower" level, while each time it is coded a $+1$ its actual value is the factor's "upper" level. The designation of lower and upper is completely arbitrary for a qualitative variable. The effects representation in Table 5.5 can be related to the calculation of main effects. Before doing so, we discuss the effects coding for interactions.

The effects coding for interactions is similar to that of individual factors and is derivable directly from the effects coding for the individual factors. Table 5.6 shows the effects coding for a two-factor, two-level factorial experiment, including the interaction column. The elements in the interaction column AB are the products of the individual elements in the columns labeled A and B. For example, the second element in the AB column is the product of the second elements in the A and B columns: $(-1)(+1) = -1$.

Also listed in Table 5.6 are the symbolic average responses (these would be individual responses if $r = 1$) for each of the factor–level combinations designated by the effects coding for the factors. A subscript 1 denotes the lower

TABLE 5.5 Equivalent Representations for Factor Levels in a Two-Factor Factorial Experiment

Factor Levels*		Effects Representation	
Factor A	Factor B	A	B
Lower	Lower	-1	-1
Lower	Upper	-1	$+1$
Upper	Lower	$+1$	-1
Upper	Upper	$+1$	$+1$

*Either factor level can be designated "lower" or "upper" if the levels are qualitative.

TABLE 5.6 Effects Representation for Main Effects and Interaction in a Two-Factor Factorial Experiment

Effects			Average
A	B	AB	Response
−1	−1	+1	$\bar{y}_{11\bullet}$
−1	+1	−1	$\bar{y}_{12\bullet}$
+1	−1	−1	$\bar{y}_{21\bullet}$
+1	+1	+1	$\bar{y}_{22\bullet}$

level of a factor, and a subscript 2 denotes the upper level. The first average response shown is $\bar{y}_{11\bullet}$, since $A = -1$ and $B = -1$ denotes the combination with both factors at their lower levels. The other average responses shown in the table are identified in a similar manner.

Now consider the main effect for factor A, designated $M(A)$, for a two-factor, two-level experiment with repeat tests. From the definition of a main effect as the difference in average responses at each level of a factor, the main effect for A can be calculated as

$$\begin{aligned} M(A) &= \bar{y}_{2\bullet\bullet} - \bar{y}_{1\bullet\bullet} \\ &= \frac{-\bar{y}_{11\bullet} - \bar{y}_{12\bullet} + \bar{y}_{21\bullet} + \bar{y}_{22\bullet}}{2}. \end{aligned}$$

The latter equality holds because, for example, $\bar{y}_{1\bullet\bullet} = (\bar{y}_{11\bullet} + \bar{y}_{12\bullet})/2$. Observe that the signs on the response averages for this main effect are the same as those in the effects representation for A in Tables 5.5 and 5.6. In the same way one can readily verify that

$$M(B) = \frac{-\bar{y}_{11\bullet} + \bar{y}_{12\bullet} - \bar{y}_{21\bullet} + \bar{y}_{22\bullet}}{2}.$$

Again, the signs on the response averages in the main effect for B are the same as those in the B columns of Tables 5.5 and 5.6.

The definition of a two-factor interaction effect is that it is half the difference between the main effects for one of the factors at the two levels of the other factor. The main effect for factor A at each level of factor B is calculated as follows for the lower level of factor B:

$$M(A)_{j=1} = \bar{y}_{21\bullet} - \bar{y}_{11\bullet}$$

and for the upper level of factor B:

$$M(A)_{j=2} = \bar{y}_{22\bullet} - \bar{y}_{12\bullet}.$$

Note that the main effect for factor A at the first (lower) level of B is again the difference of two averages, those for all responses at each of the levels of A using only those observations that have B fixed at its lower level ($j = 1$). Similarly, the main effect for A at the second (upper) level of B is the difference of the average responses for each level of A using only those observations that have the second level of B.

From these effects the interaction between factors A and B, denoted $I(AB)$, can be calculated as

$$\begin{aligned} I(AB) &= \frac{M(A)_{j=2} - M(A)_{j=1}}{2} \\ &= \frac{(\bar{y}_{22\bullet} - \bar{y}_{12\bullet}) - (\bar{y}_{21\bullet} - \bar{y}_{11\bullet})}{2} \\ &= \frac{+\bar{y}_{11\bullet} - \bar{y}_{12\bullet} - \bar{y}_{21\bullet} + \bar{y}_{22\bullet}}{2}. \end{aligned}$$

The signs on the average responses for this interaction effect are the same as those in the AB column of Table 5.6.

This example illustrates a general procedure (see Exhibit 5.6) for the calculation of main effects and interaction effects for two-level factorial experiments. This procedure can be used with any number of factors so long as there are an equal number of repeat observations for each of the factor–level combinations included in the experiment (that is, the design is balanced).

EXHIBIT 5.6 CALCULATION OF EFFECTS FOR TWO-LEVEL FACTORS

1. Construct the effects representation for each main effect and each interaction.
2. Calculate linear combinations of the average responses (or individual responses if $r = 1$), using the signs in the effects column for each main effect and for each interaction.
3. Divide the respective linear combinations of average responses by 2^{k-1}, where k is the number of factors in the experiment.

Table 5.7 shows the effects representation of the main effects and interactions for the chemical-yield study. The averages listed are obtained from Figure 5.3. Also shown in the table are the calculated main effects and interactions.

TABLE 5.7 Effects Representation and Calculated Effects for the Pilot-Plant Chemical-Yield Study

Factors

Symbol	Designation
A	Temperature
B	Concentration
C	Catalyst

Results

Effect Representation							Average Yield
A	*B*	*C*	*AB*	*AC*	*BC*	*ABC*	($r = 2$)
−1	−1	−1	+1	+1	+1	−1	60
−1	−1	+1	+1	−1	−1	+1	52
−1	+1	−1	−1	+1	−1	+1	54
−1	+1	+1	−1	−1	+1	−1	45
+1	−1	−1	−1	−1	+1	+1	72
+1	−1	+1	−1	+1	−1	−1	83
+1	+1	−1	+1	−1	−1	−1	68
+1	+1	+1	+1	+1	+1	+1	80
Calculated effects							
23.0	−5.0	1.5	1.5	10.0	0.0	0.5	

$s_e = 2.828$

The estimated standard error (see Section 6.2) of any one of these effects is $2s_e/n^{1/2}$, where $n = 16$ is the total number of test runs ($k = 3$ and $r = 2$ for this experiment) and s_e is the estimated standard deviation of the uncontrolled experimental error. For this experiment, $s_e = 2.828$. Hence, the estimated standard error of any one of the effects shown in Table 5.7 is 1.414. Comparing the calculated effects with this estimated standard error, it is apparent that the only interaction that is substantially larger than the standard error is that between temperature and catalyst. These calculations confirm the graphical conclusions drawn from Figure 5.4. In addition, although none of the interactions involving concentration are substantially larger than the standard error, the main effect for concentration is over 3.5 times larger than the effects standard error.

The effects representation of main effects and interactions is used in Chapter 7 to show how the intentional confusing or confounding of effects known to be zero or small can be used to control for extraneous variability in

the designing of experiments and to reduce the number of test runs required for a factorial experiment. Statistical analyses of effects for factorial experiments are detailed in Chapters 6, 10, and 13, depending on the design used and the effects being analyzed.

5.4 GRAPHICAL ASSESSMENT OF FACTOR EFFECTS

A visual examination of calculated effects provides important information about the influence of factors on the response variable. Plotting the effects is especially important when the experimental design does not permit satisfactory estimation of the uncontrolled error variation.

Highly fractionated designs used for screening experiments (see Chapter 7) often are used to obtain information about only the major factor effects on a response. These designs are generally conducted so that only main effects are estimated and few degrees of freedom are available for estimating the uncontrolled error standard deviation. With such designs, the imprecise estimate of the error standard deviation may limit the usefulness of the interval estimation and the testing procedures discussed in Chapter 6.

Factor effects can be calculated for any complete factorial experiment using the techniques discussed in the previous section. These techniques can also be used to calculate effects for fractional factorial experiments designed using the procedures described in Chapter 7. Once they are calculated, normal quantile-quantile plots of the effects highlight those effects that have unusually large magnitudes.

Comparisons of raw data or residuals from model fits with the normal probability distribution (Section 2.3) using normal quantile-quantile plots are detailed in Section 18.2. If, following such a comparison, raw data or residuals can be considered normally distributed, then factor effects can also be considered normally distributed. Even if data or residuals are not normally distributed, factor effects can often be considered normally distributed because they are calculated from averages for which the Central Limit Property (Exhibit 2.11) justifies the normal probability distribution. For the remainder of this section, factor effects are assumed to be normally distributed, the justification coming from either normal quantile–quantile plots of the raw data or residuals or from the Central Limit Property. It is also assumed that the uncontrolled error variability is also assumed to have a constant standard deviation for all combinations of the factor levels.

If factor effects are normally distributed and inert (that is, the factor levels do not affect the mean response), then the calculated factor effects are a random sample from a normal probability distribution with a zero mean and a constant standard deviation. The ordered (from largest negative to largest positive) factor effects then behave like ordered quantiles (see Exhibit 5.7)

from a normal probability distribution. This result is important because a plot of ordered factor effects versus theoretical quantiles from a normal probability distribution should approximate a straight line. Effects that depart substantially from a straight line formed by the smaller (in magnitude) effects indicate main effects or interactions that are larger than one would expect if all the effects were inert. Exhibit 5.8 lists the steps needed to compare graphically factor effects to theoretical quantiles from a normal probability distribution.

EXHIBIT 5.7 QUANTILES

A quantile, denoted $Q(f)$, is a quantity that divides a data set or a population into two groups so that a specified fraction or proportion f of the sample or population have values less than or equal to the value of the quantile. Ordered data values $y_{(1)} \leqslant y_{(2)} \leqslant \cdots \leqslant y_{(n)}$ are quantiles corresponding to the data fractions $1/n, 2/n, \ldots, 1$.

EXHIBIT 5.8 NORMAL QUANTILE–QUANTILE PLOTS OF FACTOR EFFECTS

1. Calculate the factor effects. Designate the ordered effects data quantiles $Q(f_i)$, for $i = 1, 2, \ldots, k$.
2. Calculate standard normal quantiles $Q_{SN}(f_i)$ for $f_i = (i - 3/8)/(k + 1/4)$ for $i = 1, 2, \ldots, k$ using the following formula

$$Q_{SN}(f_i) = 4.91\{f_i^{0.14} - (1 - f_i)^{0.14}\}. \tag{5.1}$$

3. Plot $Q(f_i)$ on the vertical axis versus the corresponding $Q_{SN}(f_i)$ on the horizontal axis.

To illustrate the information obtainable from quantile plotting of effects, we include Figure 5.6, a normal quantile–quantile plot from the pilot plant chemical yield study. The calculated effects are shown in Table 5.7. As is characteristic of this type of normal quantile–quantile plot, most of the effects are small and appear to approximate a straight line; however, three of them deviate markedly from the others.

A straight line has been superimposed on Figure 5.6 over the four small effects. The main effects for temperature (A) and concentration (B), and the interaction between temperature and catalyst (AC) appear to deviate substantially from the straight line. As is confirmed in Chapter 6 (see Table 6.4), these three effects are statistically significant and are the dominant ones.

Plotting the effects is useful even if statistical tests can be performed on them. A visual comparison of the effects aids in the assessment of whether

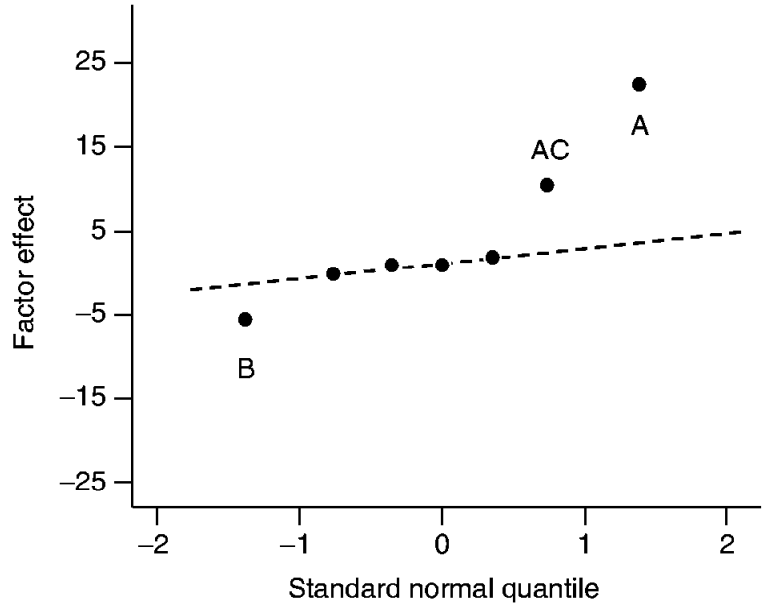


Figure 5.6 Normal quantile–quantile plot of chemical yield data factor effects.

they are statistically significant because they are measured with great precision, resulting in very high power (Section 2.7) for the corresponding statistical tests, or because the effects truly are the dominant ones.

As with all visual comparisons, one should whenever possible confirm conclusions drawn from an analysis of plotted effects with the calculation of confidence intervals or tests of appropriate statistical hypotheses. This confirmation is desirable because of the need to compare apparent large effects with an appropriate estimate of uncontrolled experimental error variation. Such comparisons reduce the chance of drawing inappropriate conclusions about the existence or dominance of effects. In the absence of estimates of experimental error variation, however, the plotting of effects is still a vital tool for the analysis of factor effects.

APPENDIX: CALCULATION OF EFFECTS FOR FACTORS WITH MORE THAN TWO LEVELS

The general procedures detailed in Section 5.3 for representing effects of two-level factors as differences of averages can be extended to any number of factor levels, but the representation is not unique. In this appendix we briefly outline an extension to factors whose numbers of levels are greater than two, with special emphasis on factors whose numbers of levels are powers of two.

A main effect for a two-level factor can be uniquely defined as the difference between the average responses at the two levels of the factor. When a factor has more than two levels, say k, there is more than one main effect

and there are many ways of defining the main effects. One way is to calculate differences between averages for levels 1 to $k-1$ from the average for level k:

$$\bar{y}_{1\bullet} - \bar{y}_{k\bullet}, \bar{y}_{2\bullet} - \bar{y}_{k\bullet}, \ldots, \bar{y}_{k-1\bullet} - \bar{y}_{k\bullet}.$$

An alternative way is to calculate the differences of the averages at the various levels from the overall average:

$$\bar{y}_{1\bullet} - \bar{y}_{\bullet\bullet}, \bar{y}_{2\bullet} - \bar{y}_{\bullet\bullet}, \ldots, \bar{y}_{k\bullet} - \bar{y}_{\bullet\bullet}.$$

Note that main effects calculated from one of these definitions can be determined from the other by taking linear combinations of the effects (see the exercises).

In the remainder of this appendix, attention is directed to factors whose numbers of levels is a power of two. The generalization of the effects representation of two-level factors to this situation allows many of the design features of fractional factorial experiments (Chapter 7) to be extended to factors having more than two levels.

One way to specify main effects for a factor whose number of levels, say k, is a power of two, $k = 2^m$, is to define m two-level factors to represent the k-level factor. One can show that any main effect for the original factor that is defined as a difference in averages can be expressed as a linear combination of the columns of main effects and interactions for the m two-level factors.

For example, suppose A is a four-level factor ($k = 4$). Define two two-level factors A_1 and $A_2 (m = 2)$ as follows:

Level of Factor A	**Level of A_1**	**Level of A_2**
1	−1	−1
2	−1	+1
3	+1	−1
4	+1	+1

Represent the response averages in the equivalent forms $\bar{y}_{1\bullet}, \bar{y}_{2\bullet}, \bar{y}_{3\bullet}, \bar{y}_{4\bullet}$ and $\bar{y}_{11\bullet}, \bar{y}_{12\bullet}, \bar{y}_{21\bullet}, \bar{y}_{22\bullet}$ corresponding to the two representations of the four factor levels. One can show that any definition of a main effect for A that can be written in the form

$$c_1\bar{y}_{1\bullet} + c_2\bar{y}_{2\bullet} + c_3\bar{y}_{3\bullet} + c_4\bar{y}_{4\bullet},$$

where the sum of the coefficients $c_1, \ldots, c_4$ is zero (*contrasts* of the averages), can be written as a linear combination of the main effects A_1, A_2 and the "interaction" A_1A_2, calculated from the $\bar{y}_{ij}$, where, for example, the "main effect A_1" is $-\bar{y}_{1\bullet} - \bar{y}_{2\bullet} + \bar{y}_{3\bullet} + \bar{y}_{4\bullet}$.

It is important to note that all the "main effects" and "interactions" of the m constructed two-level factors are needed to represent just the main effects of the original factors. Continuing the above example, the effects represented by A_1, A_2, and A_1A_2 each represent a main effect for the original factor A. This is because one cannot express all the differences in the factor–level averages

$$\bar{y}_{1\bullet} - \bar{y}_{k\bullet}, \quad i = 1, 2, \ldots, k-1,$$

or

$$\bar{y}_{1\bullet} - \bar{y}_{\bullet\bullet}, \quad i = 1, 2, \ldots, k,$$

just in terms of the linear combinations of the averages indicated by the signs in A_1 and A_2. One can, however, express these differences as linear combinations of A_1, A_2, and A_1A_2. Similarly, A_1B, A_2B, and A_1A_2B each represent a two-factor interaction of the four-level factor A with a second factor B. In particular, the interaction A_1A_2B does not represent a three-factor interaction, because A_1A_2 is a main effect for factor A.

This approach can be extended to factors having eight ($m = 3$), sixteen ($m = 4$), or any number of levels that is a power of two. In each case, the complete effects representation of the m constructed two-level factors, including all $2^m - 1$ main effects and interactions, is needed to completely specify the main effects of the original factors.

REFERENCES

Text References

Extensive coverage of complete factorial experiments appears in the texts listed at the end of Chapter 4. Each of the texts below discusses the construction and use of factorial experiments.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*, New York: John Wiley & Sons, Inc.

Davies, O. L. (Ed.) (1971). *The Design and Analysis of Industrial Experiments*, New York: Macmillan Co.

Diamond, W. J. (1981). *Practical Experimental Designs*, Belmont, CA: Wadsworth, Inc.

The first, third, and fourth references below contain detailed information on the construction and use of quantile plots. The first and second references detail the comparison of sample distributions with many common reference distributions, including the normal probability distribution.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*, Boston, MA: Duxbury Press.

Nelson, Wayne B. (1979). "How to Analyze Data with Simple Plots," ASQC Technical Conference Transactions, 89–94. Milwaukee, WI: American Society for Quality Control.

Shapiro, S. S. (1980). *How to Test Normality and Other Distributional Assumptions.* Milwaukee, WI: American Society for Quality Control.

Wilk, M. B. and Gnanadesikan, R. (1968). "Probability Plotting Methods for Data Analysis," *Biometrika*, **55**, 1–17.

Data References

The average chemical yields for the pilot-plant experiment in Figure 5.3 are taken from

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*, New York: John Wiley & Sons, Inc., Table 10.1, p. 308.

The cutting tool life data are taken from

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2001). *Introduction to Linear Regression Analysis*, Third Edition, New York: John Wiley & Sons, Inc.

EXERCISES

1 A test program is to be conducted to evaluate the performance of artillery projectiles against a variety of targets. The six target types to be included in the experiment are aluminum, steel, titanium, topsoil, sand, and simulated fuel tanks. The projectiles will be fired from two different angles (measured from horizontal to line of flight): 60 and 45°. Construct a completely randomized design for this factorial experiment.

2 The operating characteristics of a truck diesel engine that operates on unleaded gasoline are to be investigated. The time (sec) needed to reach stable operation under varying conditions is the response of primary interest. Tests are to be run at four engine speeds (1000, 1200, 1400, and 1600 rpm) and four temperatures (70, 40, 20, and 10°F). Because a large budget is available for this project, repeat tests can be conducted at each factor–level combination. Construct a completely randomized design for this experiment.

3 An experiment is to be conducted to study heat transfer in article molds used in the manufacture of drinking glasses. Three molten-glass temperatures are to be varied in the experiment: 1100, 1200 (the industry standard),

and 1300°C. Two cooling times are also to be varied in the experiment: 15 and 20 sec. In addition, two types of glass, type A and type B, are to be studied. Due to budget constraints, repeat tests can only be run for the 1200°C test runs. Construct a completely randomized design for this factorial experiment.

4 The fuel economy of two lubricating oils for locomotive engines is to be investigated. Fuel economy is measured by determining the brake-specific fuel consumption (BSFC) after the oil has operated in the engine for 10 minutes. Each oil is to be tested five times. Suppose that the run order and the resulting data are as follows:

Run	BSFC	Run	BSFC
1	0.536	6	0.550
2	0.535	7	0.552
3	0.538	8	0.559
4	0.537	9	0.563
5	0.542	10	0.571

Use graphical and numerical methods to discuss how conclusions about the effectiveness of the two oils might be compromised if the order that the oils were tested was

A, A, A, A, A, B, B, B, B, B.

Would the same difficulties occur if a randomization of the order of testing resulted in the following test sequence

A, B, A, A, B, B, A, B, B, A?

How would conclusions about the effectiveness of the two oils change, based on which test sequence was used?

5 Suppose that in the experiment described in Exercise 3 the variation in mold temperature (a measure of heat transfer) is the primary response of interest. Further, suppose budget constraints dictate that only the 1200° test runs (with repeats) can be conducted. The data below show test results for

two plants that manufacture drinking glasses. Use graphical techniques to assess whether an interaction effect exists between glass type and cooling time. Assess the presence of interactions separately for the two plants.

Plant 1

Glass Type	Cooling Time (sec)	Mold Temperature (°C)
A	15	467
B	15	462
A	20	469
B	20	467
A	15	470
B	15	463
A	20	472
B	20	469

Plant 2

Glass Type	Cooling Time (sec)	Mold Temperature (°C)
A	15	473
B	15	469
A	20	466
B	20	465
A	15	462
B	15	462
A	20	463
B	20	467

6 Calculate the main effects and the two-factor interaction between glass type and cooling time for each plant in Exercise 5. Do these calculated effects reinforce the conclusions drawn from the graphs?

7 Suppose that the experiment described in Exercise 3 can be conducted as stated except that no funds are available for repeat tests. The data below represent test results from such an experiment. The response of interest is, as in the previous exercise, the mold temperature. Use graphical techniques to assess whether interactions exist among the design factors.

Molten-Glass Temperature (°C)	Glass Type	Cooling Time	Mold Temperature
1300	A	15	482
1200	A	20	469
1300	B	15	471
1100	A	15	459
1200	B	20	470
1300	A	20	480
1200	B	15	475
1200	A	15	460
1100	B	20	479
1100	A	20	462
1300	B	20	469
1100	B	15	471

8 In Exercise 7, consider only the data for the 15-second cooling time. Calculate the main effects and two-factor interaction effect for glass type and molten-glass temperature, using only

(a) 1100 and 1200 °C,

(b) 1100 and 1300 °C as the factor levels.

Interpret these calculated effects along with suitable graphics displaying the effects.

9 In Exercise 7, suppose only two molten glass temperatures had been investigated, 1100 and 1300°C (that is, remove the data for 1200°C). Display the data on a cube plot as in Figure 5.3. Calculate the main effects, two-factor interaction effects, and the three-factor interaction effect.

10 In Exercise 9 do the effects approximate a straight line when plotted in a normal quantile–quantile graph? What conclusions do you draw about possibly significant effects? Display and interpret those effects that appear to be nonzero using suitable graphs.

11 Construct a table of the effects representation for the main effects and two-factor interaction of glass type and cooling time for

(a) only the portion of the experiment involving plant 1 described in Exercise 5, and

(b) the entire experiment.

Using the effects representation, calculate the numerical values of the main effects and interactions. Plot the effects in a normal quantile plot. Are these calculated effects consistent with the visual interpretations of the interaction plots?

12 An experiment is to be conducted in which the effects of the angle of incidence, incident light intensity, wavelength, concentration of a compound, acidity of a chemical solution, and opaqueness of a glass flask affect the intensity of light passing through a mixture of the solution and the compound. Each of the above factors is to be investigated at two levels. Construct a table of the effects representation for all main effects and interactions for a complete factorial experiment.

13 Normal quantile–quantile plots are used to compare raw data with the normal probability distribution. To construct such a plot follow a procedure similar to the one detailed in Exhibit 5.8 for factor effects. Determine $Q(f_i)$ for each of the data values and plot $Q(f_i)$ versus the corresponding value calculated in Equation 5.1. Construct a normal quantile–quantile plot for the Refinery 1 data introduced in Exercise 20 of Chapter 3. Do these data appear consistent with the normal probability distribution?

14 A chemical engineer ran a two-level factorial experiment to study the process yield as a function of time, temperature, and catalyst. The following effects were calculated from the study:

Term	Effect
Time	2.9594
Temp	2.7632
Catalyst	0.1618
Time × Temp	0.8624
Time × Catalyst	0.0744
Temp × Catalyst	−0.0867

Using a normal quantile–quantile plot, assess the effects for possible significance.

15 A scientist investigated the affects of annealing time and annealing temperature on the density of a polymer. A response–surface design was used and a quadratic model was fit to the data with the following results:

Term	Effect
Time	7.55
Temperature	10.05
Time × Temperature	3.69
$(\text{Time})^2$	−6.09
$(\text{Temperature})^2$	−7.88

Use a normal quantile–quantile plot to assess the possible significance of the effects.

16 An experiment consists of a single factor that has four levels. Suppose each level of the factor is repeated r times during the experiment. Consider the following two ways of defining the effects of the four levels of the factor:

(a) $\bar{y}_{1\bullet} - \bar{y}_{4\bullet}$, $\bar{y}_{2\bullet} - \bar{y}_{4\bullet}$, $\bar{y}_{3\bullet} - \bar{y}_{4\bullet}$:

(b) $\bar{y}_{1\bullet} - \bar{y}_{\bullet\bullet}$, $\bar{y}_{2\bullet} - \bar{y}_{\bullet\bullet}$, $\bar{y}_{3\bullet} - \bar{y}_{\bullet\bullet}$, $\bar{y}_{4\bullet} - \bar{y}_{\bullet\bullet}$.

Show that each of these sets of main effects can be determined from the other.

17 Construct two two-level factors, say A_1 and A_2, to represent the four-level factor in Exercise 16. Make a symbolic effects representation table similar to Table 5.6 for these new factors and their interaction. Show that each of the sets of main effects in Exercise 16 can be expressed as a linear function of the main effects and interaction of the two constructed variables. (Hint: Denote the three columns of effects by $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_{12}$. Express any of the main effects in (a) or (b) as $b_1\bar{y}_{1\bullet} + b_2\bar{y}_{2\bullet} + b_3\bar{y}_{3\bullet} + b_4\bar{y}_{4\bullet}$. Put the coefficients $b_1, \ldots, b_4$ in a column vector $\mathbf{b}$, and show that $\mathbf{b} = c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + c_3\mathbf{a}_{12}$ for suitably chosen constants c_1, and c_2, and c_3.)

18 Construct a main-effects table for a factor that has eight levels by defining three two-level factors to represent the eight levels of the original factor. Then write out seven main effects for the eight-level factor similar to one of the two definitions in Exercise 16. Show that one (or more) of these main effects can be written as a linear function of the seven effects for the three constructed factors.

19 An experiment is to have three factors each at two levels, and one factor at four levels. Create two two-level factors from the four-level factor. List the factor–level combinations for a complete factorial experiment in these five factors. Show by comparison that this is the same as a complete factorial experiment in the original four factors.

20 Repeat Exercise 19 for an experiment that is to have two two-level factors and one eight-level factor.

21 An experiment is conducted to study the effect of coating machine, operator, and scrim material on the density (grams/ft^2) of a coated fabric. Each of the four operators (Bob, Mary, Sam, and Ben) ran each of the two machines in the plant (machines 15 and 21) with both types of scrim material (A and B). Two repeat tests were obtained for each of the process conditions. Data from the study are shown in the table below. The standard deviation of an individual value after accounting for the three study factors, the uncontrolled experimental error standard deviation, is 7.1 grams/ft^2.

Machine 15			Machine 21		
Operator	Material	Density	Operator	Material	Density
Bob	A	190.1	Bob	A	209.4
Bob	B	182.0	Bob	B	194.2
Mary	A	200.8	Mary	A	224.6
Mary	B	196.0	Mary	B	189.1
Sam	A	205.6	Sam	A	202.1
Sam	B	191.8	Sam	B	183.1
Ben	A	183.5	Ben	A	191.8
Ben	B	170.5	Ben	B	186.7
Bob	A	179.4	Bob	A	216.2
Bob	B	177.8	Bob	B	188.3
Mary	A	191.0	Mary	A	210.1
Mary	B	180.7	Mary	B	186.1
Sam	A	198.3	Sam	A	207.2
Sam	B	206.9	Sam	B	198.0
Ben	A	182.1	Ben	A	184.4
Ben	B	187.5	Ben	B	187.5

(a) Make interaction plots of each pair of factors.

(b) How many observations are in each of the plotted averages?

(c) Describe in words what is being displayed for each of the interaction plots.

(d) Interpret the results in light of the standard deviation for an individual value.

CHAPTER 6

Analysis of Completely Randomized Designs

In this chapter techniques for the analysis of data obtained from balanced, completely randomized designs are presented. Analysis-of-variance procedures are introduced as a methodology for the simultaneous assessment of factor effects. The analysis of complete factorial experiments in which factor effects are fixed (constant) is stressed. Multiple comparison procedures that are used to understand the magnitude and direction of individual and joint factor effects are also discussed. Adaptations of these procedures to unbalanced designs, to designs with nested factors, to blocking designs, and to alternative models are detailed in subsequent chapters. The major topics included in this chapter are:

- *analysis-of-variance decomposition of the variation of the observed responses into components due to assignable causes and to uncontrolled experimental error,*
- *estimation of model parameters for both qualitative and quantitative factors,*
- *statistical tests of factor effects, and*
- *multiple comparison techniques for comparing means or groups of means.*

Chapter 3 of this book details inference procedures for means and standard deviations from one or more samples. The analysis of data from many diverse types of experiments can be placed in this context. Many others, notably the multifactor experiments described in Chapters 4 and 5, require alternative inferential procedures. In this chapter we introduce general procedures for analyzing multifactor experiments.

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

To more easily focus on the concepts involved in the analysis of multifactor experiments, we restrict attention in this chapter to balanced complete factorial experiments involving factors whose effects on a response are, apart from uncontrollable experimental error, constant. Analytic techniques for data obtained from completely randomized designs are stressed. Thus, the topics covered in this chapter are especially germane to the types of experiments discussed in Chapter 5.

Analysis-of-variance procedures are introduced in Section 6.1 for multifactor experiments. Estimation of analysis-of-variance model parameters in Section 6.2 includes interval estimation and the estimation of polynomial effects for quantitative factors. Statistical tests for factor effects are presented in Section 6.3. Multiple comparisons involving linear combinations of means is discussed in Section 6.4.

6.1 BALANCED MULTIFACTOR EXPERIMENTS

Single-factor experiments consist of a single controllable design factor whose levels are believed to affect the response. If the factor levels affect the mean of the response, the influences of the factor levels are referred to as *fixed* (constant) *effects*. If the factor levels affect the variability of the response, the factor effects are referred to as *random effects*.

Multifactor experiments consist of two or more controllable design factors whose *combinations* of levels are believed to affect the response. The number of levels can be different for each factor. Balanced complete factorial experiments have an equal number of repeat tests for all *combinations* of levels of the design factors. The torque study and the chemical pilot plant experiment introduced in Chapter 5 are examples of balanced multifactor experiments in which the factor effects are all fixed. In the following subsections, we introduce the analysis-of-variance procedures for multifactor experiments involving fixed effects. A single-factor experiment is an important special case that is illustrated with an example.

6.1.1 Fixed Factor Effects

Factor effects are called *fixed* if the levels of the factor are specifically selected because they are the only ones for which inferences are desired (see Exhibit 6.1). Because the factor levels can be specifically chosen, it is assumed that the factor levels exert constant (apart from experimental error) influences on the response. These effects are modeled by unknown parameters in statistical models that relate the mean of the response variable to the factor levels (see Section 6.1.2). Fixed factor levels are not randomly selected nor are they the result of some chance mechanisms; they are intentionally chosen

for investigation. This is the primary distinction between fixed and random factor levels (see Section 10.1).

EXHIBIT 6.1

Fixed Factor Effects. Factors have fixed effects if the levels chosen for inclusion in the experiment are the only ones for which inferences are desired.

The pilot-plant chemical-yield study introduced in Section 5.2 is an experiment in which all the factors have fixed effects. Table 6.1 lists the factor–level combinations and responses for this experiment. The two temperatures, the two concentrations, and the two catalysts were specifically chosen and are the only ones for which inferences can be made. One cannot infer effects on the chemical yield for factor levels that are not included in the experiment — for example, that the effect of a temperature of 170°C would be between those of 160 and 180°C — unless one is willing to make additional assumptions about the change in the mean response between the levels included in the experiment.

At times one is willing to assume, apart from experimental error, that there is a smooth functional relationship between the quantitative levels of a factor and the response. For example, one might be willing to assume that temperature has a linear effect on the chemical yield. If so, the fitting of a model using two levels of temperature would allow one to fit a straight line between the average responses for the two levels and infer the effect of a temperature of, say, 170°C. In this setting the levels of temperature are still fixed and inferences can only be drawn on the temperature effects at these two

TABLE 6.1 Test Results from Pilot-Plant Chemical-Yield Experiment*

Temperature (°C)	Concentration (%)	Catalyst	Yield (g)
160	20	C_1	59, 61
		C_2	50, 54
	40	C_1	50, 58
		C_2	46, 44
180	20	C_1	74, 70
		C_2	81, 85
	40	C_1	69, 67
		C_2	79, 81

*Two repeat test runs for each factor–level combination.

levels. It is the added assumption of a linear effect of temperature between 160 and 180°C that allows inferences to be made for temperatures between the two included in the experiment.

6.1.2 Analysis-of-Variance Models

Analysis-of-variance (ANOVA) procedures separate or partition the variation observable in a response variable into two basic components: variation due to assignable causes and to uncontrolled or random variation (see Exhibit 6.2). Assignable causes refer to known or suspected sources of variation from variates that are controlled (experimental factors) or measured (covariates) during the conduct of an experiment. Random variation includes the effects of all other sources not controlled or measured during the experiment. While random sources of response variability are usually described as including only chance variation or measurement error, any factor effects that have been neither controlled nor measured are also included in the measurement of the component labeled "random" variation.

EXHIBIT 6.2 SOURCES OF VARIATION

Assignable Causes. Due to changes in experimental factors or measured covariates.
Random Variation. Due to uncontrolled effects, including chance causes and measurement errors.

Assignable causes corresponding to controllable factors can be either fixed or random effects. Fixed effects were described in the last section. Random effects will be discussed in Part III of this text. A more complete and explicit description of factor effects requires the specification of a statistical model that includes both assignable causes and random variation. Analysis-of-variance models for experiments in which all factor effects are fixed adhere to the assumptions listed in Exhibit 6.3.

EXHIBIT 6.3 FIXED-EFFECTS MODEL ASSUMPTIONS

1. The levels of all factors in the experiment represent the only levels of which inferences are desired.
2. The analysis-of-variance model contains parameters (unknown constants) for all main effects and interactions of interest in the experiment.
3. The experimental errors are statistically independent.
4. The experimental errors are satisfactorily modeled by the normal probability distribution with mean zero and (unknown) constant standard deviation.

A statistical model for the pilot-plant chemical yield example is

$$y_{ijkl} = \mu_{ijk} + e_{ijkl}, \quad i = 1, 2, \quad j = 1, 2, \quad k = 1, 2, \quad l = 1, 2. \tag{6.1}$$

In this model, μ_{ijk} represents the effects of the assignable causes and e_{ijkl} represents the random error effects. The expression of the model as in Equation (6.1) stresses the connection between multifactor experiments with fixed effects and sampling from a number of populations or processes that differ only in location; that is, in their means. The assignable cause portion of the model can be further decomposed into terms representing the main effects and the interactions among the three model factors:

$$\mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}. \tag{6.2}$$

The subscript i in the model (6.1) refers to one of the levels of temperature, the subscript j to one of the levels of concentration, the subscript k to one of the catalysts, and the subscript l to one of the two repeat tests. The components of (6.2) permit the mean of the response to be affected by both the individual factor levels (main effects) and combinations of two or more factor levels (interactions).

The parameters in (6.2) having a single subscript represent main effects for the factor identified by the subscript. Two-factor interactions are modeled by the terms having two subscripts, and the three-factor interaction by terms having three subscripts. The first term in equation (6.2) represents the overall response mean, that is, a term representing an overall population or process mean from which the various individual and joint factor effects are measured.

The interaction components of (6.2) model joint effects that cannot be suitably accounted for by the main effects. Each interaction effect measures the incremental joint contribution of factors to the response variability over the contribution of the main effects and lower-order interactions. Thus, a two-factor interaction is present in a model only if the effects of two factors on the response cannot be adequately modeled by the main effects. Similarly, a three-factor interaction is included only if the three main effects and the three two-factor interactions do not satisfactorily model the effects of the factors on the response.

The parameters representing the assignable causes in (6.2) can be expressed in terms of the model means μ_{ijk}. Specification of the right side of (6.2) must be carefully considered both to allow meaningful interpretation and because the right side is not unique. To illustrate the lack of uniqueness, for any two values of μ and α_i, both $\mu + \alpha_i$ and $\mu^* + \alpha_i^*$ with $\mu^* = \mu + c$ and $\alpha_i^* = \alpha_i - c$ give the same numerical value for μ_{ijk} regardless of the value of the constant c. For this representation to be unique, we must impose constraints on the values

of the parameters. The commonly used constraints we impose are:

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0, \quad \sum_k \gamma_k = 0, \quad \sum_i (\alpha\beta)_{ij} = 0, \quad \sum_j (\alpha\beta)_{ij} = 0, \text{ etc.}$$

Thus, the constraints require that each of the parameters in (6.2) must sum to zero over any of its subscripts.

With these constraints, the effects parameters can be expressed in terms of averages of the model means. For example,

$$\begin{aligned} \mu &= \bar{\mu}_{\bullet\bullet\bullet} \quad \alpha_i = \bar{\mu}_{i\bullet\bullet} - \bar{\mu}_{\bullet\bullet\bullet} \quad \beta_j = \bar{\mu}_{\bullet j\bullet} - \bar{\mu}_{\bullet\bullet\bullet} \quad \gamma_k = \bar{\mu}_{\bullet\bullet k} - \bar{\mu}_{\bullet\bullet\bullet}, \\ (\alpha\beta)_{ij} &= \bar{\mu}_{ij\bullet} - \bar{\mu}_{i\bullet\bullet} - \bar{\mu}_{\bullet j\bullet} + \bar{\mu}_{\bullet\bullet\bullet}, \quad \text{etc.}, \\ (\alpha\beta\gamma)_{ijk} &= \mu_{ijk} - \bar{\mu}_{ij\bullet} - \bar{\mu}_{i\bullet k} - \bar{\mu}_{\bullet jk} \\ &\quad + \bar{\mu}_{i\bullet\bullet} + \bar{\mu}_{\bullet j\bullet} + \bar{\mu}_{\bullet\bullet k} - \bar{\mu}_{\bullet\bullet\bullet}. \end{aligned} \tag{6.3}$$

Note that main-effect parameters are simply the differences between the model means for one level of a factor and the overall model mean. The two-factor interaction parameters are the differences in the main effects for one level of one of the factors at a fixed level of a second and the main effect of the first factor:

$$\begin{aligned} (\alpha\beta)_{ij} &= (\bar{\mu}_{ij\bullet} - \bar{\mu}_{\bullet j\bullet}) - (\bar{\mu}_{i\bullet\bullet} - \bar{\mu}_{\bullet\bullet\bullet}) \\ &= (\text{main effect for } A \text{ at level } j \text{ of } B) \\ &\quad - (\text{main effect for } A). \end{aligned}$$

The three-factor interaction can be similarly viewed as the difference in the two-factor interaction between any two of the factors, say A and B, at a fixed level of the third factor, say C, and the two-factor interaction between A and B:

$$\begin{aligned} (\alpha\beta\gamma)_{ijk} &= (\mu_{ijk} - \bar{\mu}_{i\bullet k} - \bar{\mu}_{\bullet jk} + \bar{\mu}_{\bullet\bullet k}) \\ &\quad - (\bar{\mu}_{ij\bullet} - \bar{\mu}_{i\bullet\bullet} - \bar{\mu}_{\bullet j\bullet} + \bar{\mu}_{\bullet\bullet\bullet}) \\ &= (A \times B \text{ interaction at level } k \text{ of } C) \\ &\quad - (A \times B \text{ interaction}). \end{aligned}$$

It is important to note that while the means in the model (6.1) are uniquely defined, main-effect and interaction parameters can be introduced in many ways; that is they need not be defined as in (6.2). The form (6.2) is used in this book explicitly to facilitate an understanding of inference procedures for multifactor experiments.

Even though the main-effect and interaction parameters can be defined in many ways, certain functions of them are uniquely defined in terms of the

model means. The functions that are uniquely defined are precisely those that allow comparison among factor–level effects. We shall return to this topic in Section 6.2.

When considering whether to include interaction terms in a statistical model, one convention that is adopted is to use only hierarchical factor effects. A model is hierarchical (see Exhibit 6.4) if any interaction term involving k factors is included in the specification of the model only if the main effects and all lower-order interactions involving two, three,..., $k-1$ of the factors are also included in the model. In this way, a high-order interaction term is only added to the model if the main effects and lower-order interactions are not able to satisfactorily model the response.

EXHIBIT 6.4

Hierarchical Model. A statistical model is hierarchical if an interaction term involving k factors is included only when the main effects and all lower-order interaction terms involving the k factors are also included in the model.

We return to analysis-of-variance models in the next section, where the estimation of the model parameters is discussed. We now wish to examine the analysis-of-variance partitioning of the total sum of squares for multifactor experiments. The response variability can be partitioned into components for each of the main effects and interactions and for random variability.

6.1.3 Analysis-of-Variance Tables

The pilot-plant experiment shown in Table 6.1 is balanced, because each of the factor–level combinations appears twice in the experiment. We shall use this example and the three-factor model (6.1) to illustrate the construction of analysis-of-variance tables for complete factorial experiments from balanced designs. The three-factor model is sufficiently complex to demonstrate the general construction of main effects and interaction sums of squares.

Generalize the model (6.1) to three factors, of which factor A has a levels (subscript i), factor B has b levels (subscript j), and factor C has c levels (subscript k). Let there be r repeat tests for each combination of the factors. Thus, the experiment consists of a complete factorial experiment in three factors with r repeat tests per factor–level combination.

Throughout earlier chapters of this book the sample variance or its square root, the standard deviation, was used to measure variability. The first step in an analysis-of-variance procedure is to define a suitable measure of variation for which a partitioning into components due to assignable causes and to random variation can be accomplished. While there are many measures that could be used, the numerator of the sample variances is used for a variety of

computational and theoretical reasons. This measure of variability is referred to as the *total sum of squares* (TSS):

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

The total sum of squares adjusts for the overall average by subtracting it from each individual response; consequently, it is sometimes referred to as the total *adjusted* sum of squares, or TSS(adj). Because we shall always adjust for the sample mean, we drop the word "adjusted" in the name.

For a three-factor experiment, the total sum of squares consists of the squared differences of the observations y_{ijkl} from the overall average response $\bar{y}_{\bullet\bullet\bullet\bullet}$:

$$\text{TSS} = \sum_i \sum_j \sum_k \sum_l (y_{ijkl} - \bar{y}_{\bullet\bullet\bullet\bullet})^2. \tag{6.4}$$

To partition the above total sum of squares into components for the assignable causes (factors A, B, C) and for random (uncontrolled experimental) variation, add and subtract the factor–level combination averages $\bar{y}_{ijk\bullet}$ from each term in the summation for the total sum of squares. Then

$$\begin{aligned}
\text{TSS} &= \sum_i \sum_j \sum_k \sum_l (y_{ijkl} - \bar{y}_{ijk\bullet} + \bar{y}_{ijk\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet})^2 \\
&= r \sum_i \sum_j \sum_k (\bar{y}_{ijk\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet})^2 + \sum_i \sum_j \sum_k \sum_l (y_{ijkl} - \bar{y}_{ijk\bullet})^2 \qquad (6.5) \\
&= \text{MSS} + \text{SS}_\text{E}.
\end{aligned}$$

The first component of (6.5) is the contribution to the total sum of squares of all the variability attributable to assignable causes, often called the *model* sum of squares:

$$\text{MSS} = r \sum_i \sum_j \sum_k (\bar{y}_{ijk\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet})^2.$$

The model sum of squares can be partitioned into the following components:

$$\text{MSS} = \text{SS}_A + \text{SS}_B + \text{SS}_C + \text{SS}_{AB} + \text{SS}_{AC} + \text{SS}_{BC} + \text{SS}_{ABC}, \tag{6.6}$$

where SS_A, SS_B, and SS_C measure the individual factor contributions (main effects of the three experimental factors) to the total sum of squares, SS_{AB}, SS_{AC}, and SS_{BC} measure the contributions of the two-factor joint effects, and SS_{ABC} measures the contribution of the joint effects of all three experimental

factors above the contributions measured by the main effects and the two-factor interactions.

The sums of squares for the factor main effects are multiples of the sums of the squared differences between the averages for each level of the factor and the overall average; for example,

$$SS_A = bcr \sum_{i=1}^{a} (\bar{y}_{i\bullet\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet})^2. \tag{6.7}$$

The differences between the averages for each level of a factor and the overall average yields a direct measure of the individual factor–level effects, the main effects of the factor. The sum of squares (6.7) is a composite measure of the combined effects of the factor levels on the response. Note that the multiplier in front of the summation sign is the total number of observations that are used in the calculation of the averages $\bar{y}_{i\bullet\bullet\bullet}$ for each level of factor A. Sums of squares for each of the main effects are calculated as in (6.7). (Note: computational formulas are shown below in Table 6.2.)

Just as the main effect for a factor can be calculated by taking differences between individual averages and the overall average, interaction effects can be calculated by taking differences between (a) the main effects of one factor at a particular level of a second factor and (b) the overall main effect of the factor. For example, the interaction effects for factors A and B can be obtained by calculating:

$$\begin{aligned}&\text{(Main effect for } A \text{ at level } j \text{ of } B) - \text{(main effect for } A)\\ &= (\bar{y}_{ij\bullet\bullet} - \bar{y}_{\bullet j\bullet\bullet}) - (\bar{y}_{i\bullet\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet})\\ &= \bar{y}_{ij\bullet\bullet} - \bar{y}_{i\bullet\bullet\bullet} - \bar{y}_{\bullet j\bullet\bullet} + \bar{y}_{\bullet\bullet\bullet\bullet}.\end{aligned}$$

This is the same expression one would obtain by reversing the roles of the two factors. The interaction effects can be combined in an overall sum of squares for the interaction effects of factors A and B:

$$SS_{AB} = cr \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{ij\bullet\bullet} - \bar{y}_{i\bullet\bullet\bullet} - \bar{y}_{\bullet j\bullet\bullet} + \bar{y}_{\bullet\bullet\bullet\bullet})^2. \tag{6.8}$$

TABLE 6.2 Symbolic Analysis-of-Variance Table for Three Factors

ANOVA

Source of Variation	Degrees of Freedom (df)	Sum of Squares	Mean Square	F-Value
A	$a-1$	SS_A	$\mathrm{MS}_A = \mathrm{SS}_A/\mathrm{df}(A)$	$F_A = \mathrm{MS}_A/\mathrm{MS}_E$
B	$b-1$	SS_B	$\mathrm{MS}_B = \mathrm{SS}_B/\mathrm{df}(B)$	$F_B = \mathrm{MS}_B/\mathrm{MS}_E$
C	$c-1$	SS_C	$\mathrm{MS}_C = \mathrm{SS}_C/\mathrm{df}(C)$	$F_C = \mathrm{MS}_C/\mathrm{MS}_E$
AB	$(a-1)(b-1)$	SS_{AB}	$\mathrm{MS}_{AB} = \mathrm{SS}_{AB}/\mathrm{df}(AB)$	$F_{AB} = \mathrm{MS}_{AB}/\mathrm{MS}_E$
AC	$(a-1)(c-1)$	SS_{AC}	$\mathrm{MS}_{AC} = \mathrm{SS}_{AC}/\mathrm{df}(AC)$	$F_{AC} = \mathrm{MS}_{AC}/\mathrm{MS}_E$
BC	$(b-1)(c-1)$	SS_{BC}	$\mathrm{MS}_{BC} = \mathrm{SS}_{BC}/\mathrm{df}(BC)$	$F_{BC} = \mathrm{MS}_{BC}/\mathrm{MS}_E$
ABC	$(a-1)(b-1)(c-1)$	SS_{ABC}	$\mathrm{MS}_{ABC} = \mathrm{SS}_{ABC}/\mathrm{df}(ABC)$	$F_{ABC} = \mathrm{MS}_{ABC}/\mathrm{MS}_E$
Error	$abc(r-1)$	SS_E	$\mathrm{MS}_E = \mathrm{SS}_E/\mathrm{df}(\mathrm{Error})$	
Total	$abcr-1$	TSS		

Calculations

$$\mathrm{TSS} = \sum_i \sum_j \sum_k \sum_l y_{ijkl}^2 - \mathrm{SS}_M \qquad \mathrm{SS}_M = y_{\bullet\bullet\bullet\bullet}^2/n$$

$$\mathrm{SS}_A = \sum_i y_{i\bullet\bullet\bullet}^2/bcr - \mathrm{SS}_M \qquad \mathrm{SS}_{AB} = \sum_i \sum_j y_{ij\bullet\bullet}^2/cr - \mathrm{SS}_M - \mathrm{SS}_A - \mathrm{SS}_B$$

$$\mathrm{SS}_B = \sum_j y_{\bullet j\bullet\bullet}^2/acr - \mathrm{SS}_M \qquad \mathrm{SS}_{AC} = \sum_i \sum_k y_{i\bullet k\bullet}^2/br - \mathrm{SS}_M - \mathrm{SS}_A - \mathrm{SS}_C$$

$$\mathrm{SS}_C = \sum_k y_{\bullet\bullet k\bullet}^2/abr - \mathrm{SS}_M \qquad \mathrm{SS}_{BC} = \sum_j \sum_k y_{\bullet jk\bullet}^2/ar - \mathrm{SS}_M - \mathrm{SS}_B - \mathrm{SS}_C$$

$$\mathrm{SS}_{ABC} = \sum_i \sum_j \sum_k y_{ijk\bullet}^2/r - \mathrm{SS}_M - \mathrm{SS}_A - \mathrm{SS}_B - \mathrm{SS}_C - \mathrm{SS}_{AB} - \mathrm{SS}_{AC} - \mathrm{SS}_{BC}$$

$$\mathrm{SS}_E = \mathrm{TSS} - \mathrm{SS}_A - \mathrm{SS}_B - \mathrm{SS}_C - \mathrm{SS}_{AB} - \mathrm{SS}_{AC} - \mathrm{SS}_{BC} - \mathrm{SS}_{ABC}$$

$$y_{\bullet\bullet\bullet\bullet} = \sum_i \sum_j \sum_k \sum_l y_{ijkl} \qquad y_{i\bullet\bullet\bullet} = \sum_j \sum_k \sum_l y_{ijkl} \qquad n = abcr$$

$$y_{ij\bullet\bullet} = \sum_k \sum_l y_{ijkl} \qquad y_{ijk\bullet} = \sum_l y_{ijkl} \qquad \mathrm{df}(A) = a-1, \text{ etc.}$$

The sum of squares for the three-factor interaction, SS_{ABC}, can be obtained in a similar fashion by examining the differences between the interaction effects for two of the factors at fixed levels of the third factor and the overall interaction effect of the two factors. The result of combining these three-factor interaction effects is

$$SS_{ABC} = r\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(\bar{y}_{ijk\bullet} - \bar{y}_{ij\bullet\bullet} - \bar{y}_{i\bullet k\bullet} - \bar{y}_{\bullet jk\bullet} + \bar{y}_{i\bullet\bullet\bullet} + \bar{y}_{\bullet j\bullet\bullet} + \bar{y}_{\bullet\bullet k\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet})^2. \tag{6.9}$$

The final term needed to complete the partitioning of the total sum of squares is the second component of (6.5), the error sum of squares, SS_E. Algebraically, the error sum of squares is simply the sum of the squared differences between the individual repeat-test responses and the averages for the factor–level combinations:

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{r}(y_{ijkl} - \bar{y}_{ijk\bullet})^2. \tag{6.10}$$

The information about factor effects that is obtained from the partitioning of the total sum of squares is summarized in an analysis-of-variance table. A symbolic ANOVA table for the three-factor interaction model we have been considering is shown in Table 6.2. In this table, the first column, "source of variation," designates the partitioning of the response variability into the various components included in the statistical model.

The second column, "degrees of freedom," partitions the sample size into similar components that relate the amount of information obtained on each factor, the interaction, and the error term of the model. Note that each main effect has one less degree of freedom than the number of levels of the factor. Interaction degrees of freedom are conveniently calculated as products of the degrees of freedom of the respective main effects. The number of error degrees of freedom is the number of factor–level combinations multiplied by one less than the number of repeat tests for each combination. These degrees of freedom can all be found by identifying constraints on the terms in each of the sums of squares.

The degrees of freedom indicate how many statistically independent response variables or functions of the response variables comprise a sum of squares. For example, the $n = abcr$ response variables y_{ijlk} in a three-factor balanced complete factorial experiment are assumed to be statistically independent (Exhibit 6.3). Consequently the *unadjusted* total sum of squares

$$\sum_i \sum_j \sum_k \sum_l y_{ijkl}^2$$

has n degrees of freedom. The total *adjusted* sum of squares (6.3) does not have n degrees of freedom because there is a constraint on the n terms in the sum of squares;

$$\sum_i \sum_j \sum_k \sum_l (y_{ijkl} - \bar{y}_{\bullet\bullet\bullet\bullet}) = 0.$$

Hence, there are $abcr - 1$ degrees of freedom associated with the sum of squares TSS.

Similarly, the error sum of squares SS_E does not have n degrees of freedom, even though there are n terms in its sum of squares. For each combination of the factor levels,

$$\sum_l (y_{ijkl} - \bar{y}_{ijk\bullet}) = 0.$$

Thus the r deviations $y_{ijkl} - \bar{y}_{ijk\bullet}$ have $r - 1$ degrees of freedom. Because this is true for each of the factor–level combinations, the error sum of squares has $abc(r - 1)$ degrees of freedom. In a similar fashion, each of the main effect and interaction degrees of freedom can be calculated by determining the constraints on the terms in each sum of squares. The results are the degrees of freedom listed in Table 6.2.

The third column of Table 6.2 contains the sums of squares for various main effects and interactions, and the error components of the model. The fourth column, "mean squares," contains the respective sums of squares divided by their numbers of degrees of freedom. These statistics are used for forming the F-ratios in the next column, each main effect and interaction mean square being divided by the error mean square.

Table 6.3 lists summary statistics used in the calculation of the sums of squares for the pilot-plant chemical-yield experiment. The sums of squares can be calculated using these statistics and the computational formulas shown in Table 6.2. The complete ANOVA table is displayed in Table 6.4. Several of the F-statistics in the table are much larger than 1. The interpretation of these large F-values will be discussed in Section 6.3.

Figure 6.1 is a schematic diagram of an experiment that was conducted to study flow rates (cc/min) of a particle slurry that was pumped under constant pressure for a fixed time period through a cylindrical housing. The factor of interest in this experiment is the type of filter placed in the housing. The test sequence was randomized so that the experiment was conducted using a completely randomized design. The purpose of the experiment was to assess the effects, if any, of the filter types on the average flow rates.

TABLE 6.3 Summary Statistics for Pilot-Plant Chemical-Yield Experiment

	Factor	
Symbol	Subscript	Label
A	i	Temperature
B	j	Concentration
C	k	Catalyst

$$\sum_i \sum_j \sum_k \sum_l y_{ijkl}^2 = 68,748 \quad y_{\bullet\bullet\bullet\bullet}^2 = (1028)^2 = 1,056,784$$

$$\sum_i y_{i\bullet\bullet\bullet}^2 = (422)^2 + (606)^2 = 545,320$$

$$\sum_j y_{\bullet j\bullet\bullet}^2 = (534)^2 + (494)^2 = 529,192$$

$$\sum_k y_{\bullet\bullet k\bullet}^2 = (508)^2 + (520)^2 = 528,464$$

$$\sum_i \sum_j y_{ij\bullet\bullet}^2 = (224)^2 + (198)^2 + (310)^2 + (296)^2 = 273,096$$

$$\sum_i \sum_k y_{i\bullet k\bullet}^2 = (228)^2 + (194)^2 + (280)^2 + (326)^2 = 274,296$$

$$\sum_j \sum_k y_{\bullet jk\bullet}^2 = (264)^2 + (270)^2 + (244)^2 + (250)^2 = 264,632$$

$$\sum_i \sum_j \sum_k y_{ijk\bullet}^2 = (120)^2 + (104)^2 + (108)^2 + (90)^2$$
$$+(144)^2 + (166)^2 + (136)^2 + (160)^2 = 137,368$$

TABLE 6.4 ANOVA Table for Pilot-Plant Chemical-Yield Study

Source of Variation	df	Sum of Squares	Mean Square	F-Value
Temperature T	1	2116.00	2116.00	264.50
Concentration Co	1	100.00	100.00	12.50
Catalyst Ca	1	9.00	9.00	1.13
$T \times$ Co	1	9.00	9.00	1.13
$T \times$ Ca	1	400.00	400.00	50.00
Co $\times$ Ca	1	0.00	0.00	0.00
$T \times$ Co $\times$ Ca	1	1.00	1.00	0.13
Error	8	64.00	8.00	
Total	15	2699.00		

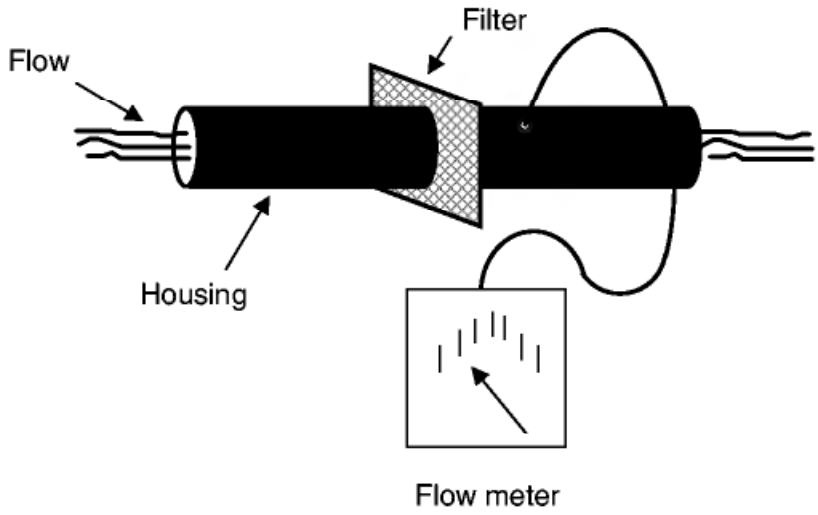


Figure 6.1 Flow-rate experiment.

The only factor in this experiment is the filter type chosen for each test run. This factor has four levels. Because inferences are desired on these specific filter types, the factor has fixed effects. If only two filter types were included in the experiment, the two-sample inference procedures detailed in Chapter 3 could be used to assess the filter effects on the flow rates. To assess the influence of all four filter types, the inferential techniques just discussed must be used to accommodate more than two factor levels (equivalently, more than two population or process means).

TABLE 6.5 Analysis of Variance Table for Flow-Rate Data

ANOVA

Source of Variation	Degrees of Freedom (df)	Sum of Squares	Mean Square	F-Value
Filter	3	0.00820	0.00273	1.86
Error	12	0.01765	0.00147	
Total	15	0.02585		

Data

Filter	Flow Rates				Total
A	0.233	0.197	0.259	0.244	0.933
B	0.259	0.258	0.343	0.305	1.165
C	0.183	0.284	0.264	0.258	0.989
D	0.233	0.328	0.267	0.269	1.097
Total					4.184

The partitioning of the total sum of squares is summarized in the ANOVA table, Table 6.5. This special case of ANOVA for single-factor experiments is often termed a *one-way* analysis of variance and the corresponding ANOVA model a *one-way classification model.* In this case the F-statistic generalizes the two sample t-statistic (Equation 3.14) to comparisons of more than two population or process means. The F-statistic for this example is not appreciably larger than 1. In Section 6.3 we shall see how to interpret this finding.

6.2 PARAMETER ESTIMATION

Inferences on specific factor effects requires the estimation of the parameters of ANOVA models. In Section 6.2.1 estimation of the error standard deviation is discussed. In Section 6.2.2 the estimation of both the means of fixed-effects models and parameters representing main effects and interactions are discussed. Relationships are established between estimates of the means of the fixed-effects models and estimates of the parameters representing the main effects and interactions. Alternative estimates of factor effects for quantitative factor levels are presented in Section 6.2.3.

6.2.1 Estimation of the Error Standard Deviation

A key assumption in the modeling of responses from designed experiments is that the experimental conditions are sufficiently stable that the model errors

can be considered to follow a common probability distribution. Ordinarily the distribution assumed for the errors is the normal probability distribution. There are several justifications for this assumption.

First, the model diagnostics discussed in Chapter 18 can be used to assess whether this assumption is reasonable. Quantile plots, similar to those discussed in Sections 5.4 and 18.2, provide an important visual assessment of the reasonableness of the assumption of normally distributed errors. Second, fixed-effects models are extensions of the one- or two-sample experiments discussed in Chapter 3. The central limit property (Section 2.3) thus provides a justification for the use of normal distributions for averages and functions of averages that constitute factor effects even if the errors are not normally distributed. Finally, randomization provides a third justification: the process of randomization induces a sampling distribution on factor effects that is closely approximated by the theoretical distributions of the t and F statistics used to make inferences on the parameters of ANOVA models.

Assuming that the errors follow a common distribution, the error mean square from the ANOVA table is a pooled estimator of the common error variance. If the model consists of all main effect and interaction parameters (a *saturated* model) and the experiment is a complete factorial with repeat tests, the error sum of squares is similar in form to equation (6.10) for three factors. With the balance in the design, the error mean square is simply the average of the sample variances calculated from the repeat tests for each factor–level combination:

$$s_e^2 = \mathrm{MS}_E = \frac{1}{abc} \sum_i \sum_j \sum_k s_{ijk}^2. \tag{6.11}$$

If there are no repeat tests and the model is saturated, there is no estimate of experimental error available. Equation (6.11) is not defined because none of the sample variances can be calculated. In this situation an estimate of experimental-error variation is only obtainable if some of the parameters in the saturated model can be assumed to be zero.

To illustrate why estimates of experimental-error variation are available when certain model parameters are assumed to be zero, consider model (6.1). When there is a single factor, $y_{ij} = \mu + \alpha_i + e_{ij}$. If the factor levels all have the same constant effect on the response, all the α_i in the model are zero. Then $\bar{y}_{i\bullet} = \mu + \bar{e}_{i\bullet}$ and $\bar{y}_{\bullet\bullet} = \mu + \bar{e}_{\bullet\bullet}$, so that the main effect for factor level i is

$$\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet} = \bar{e}_{i\bullet} - \bar{e}_{\bullet\bullet}.$$

Thus, SS_A [equation (6.7)] does not estimate any function of the main effect parameters. It measures variability in the averages of the errors. One can show that both MS_A and MS_E in this situation are estimators of the error variance σ^2.

As mentioned in Section 6.1, any effects not included in the specification of the ANOVA model are measured in the random error component of the partitioning of the total sum of squares. Thus, if one can assume that certain model parameters (equivalently, the corresponding factor effects, apart from random error) are zero, estimation of the error variance can be accomplished even if there is no replication in the design.

Ordinarily, it is reasonable to assume that high-order interactions are zero. This principle can also be used to justify the estimation of experimental-error variation in the absence of repeat tests. We stress here, as in Chapters 4 and 5, that repeat tests provide the best estimate of experimental-error variation; however, there are many experimental constraints that dictate that repeat tests cannot be included in the design of an experiment. When such constraints occur, an examination of proposed ANOVA models should be made to see whether experimental error can be estimated from high-order interactions that can reasonably be assumed to be zero.

The confidence-interval techniques presented in Section 3.2 can be used with analysis-of-variance models to estimate the error variance or the error standard deviation. Under an assumption of statistically independent, normally distributed errors, $\nu \cdot \mathrm{MS}_E/\sigma^2$ follows a chi-square sampling distribution with ν degrees of freedom, where ν is the number of error degrees of freedom from the ANOVA table. Confidence intervals for the error standard deviation can be constructed as shown in Table 3.3 with the replacement of s^2 by MS_E and the number of degrees of freedom $(n-1)$ by ν.

6.2.2 Estimation of Effects Parameters

The parameters associated with the main effects and interactions in analysis-of-variance models are estimated by functions of the factor–level response averages. For any fixed-effects ANOVA model, including the one- and two-sample models (3.1) and (3.11), the model means are estimated by the corresponding response averages. For example, in the three-factor model (6.1), the averages for the factor–level combinations, $\bar{y}_{ijk\bullet}$, estimate the model means $\mu_{ijk\bullet}$.

Confidence intervals can be placed on any of the model means using the procedures discussed in Section 3.1. A $100(1-\alpha)\%$ confidence interval for μ_{ijk} from a balanced three-factor factorial experiment is

$$\bar{y}_{ijk\bullet} - \frac{t_{\alpha/2} s_e}{r^{1/2}} < \mu_{ijk} < \bar{y}_{ijk\bullet} + \frac{t_{\alpha/2} s_e}{r^{1/2}}, \tag{6.12}$$

where $t_{\alpha/2}$ is a Student's t critical value with ν degrees of freedom corresponding to an upper-tail probability of $\alpha/2$, $s_e = (\mathrm{MS}_E)^{1/2}$, and the error mean square has $\nu = abc(r-1)$ degrees of freedom. Confidence intervals for other model means can be obtained in a similar fashion. For example, an

interval estimate of $\overline{\mu}_{i\bullet\bullet}$ can be obtained by replacing μ_{ijk} with $\overline{\mu}_{i\bullet\bullet}$, $\overline{y}_{ijk\bullet}$ with $\overline{y}_{i\bullet\bullet\bullet}$, and r with bcr in (6.12).

Comparisons among factor–level means are usually of great interest in experiments involving factors at two or more levels. Confidence intervals for mean differences such as $\overline{\mu}_{i\bullet\bullet} - \overline{\mu}_{j\bullet\bullet}$ can be constructed using the general expression

$$\hat{\theta} - t_{\alpha/2} s_e m^{1/2} < \theta < \hat{\theta} + t_{\alpha/2} s_e m^{1/2}, \tag{6.13}$$

where

$$\theta = \sum_i \sum_j \sum_k a_{ijk} \mu_{ijk}$$

is some linear combination of the model means,

$$\hat{\theta} = \sum_i \sum_j \sum_k a_{ijk} \overline{y}_{ijk\bullet}$$

is the corresponding linear combination of the response averages, and

$$m = \sum_i \sum_j \sum_k \frac{a_{ijk}^2}{r}.$$

This formula simplifies greatly for certain comparisons. For example, if a confidence interval is desired for some linear combination of the factor–level means for a single factor, the quantities in (6.13) can be rewritten as

$$\theta = \sum a_i \mu_i, \qquad \hat{\theta} = \sum a_i \overline{y}_{i\bullet}, \qquad m = \sum a_i^2 / r. \tag{6.14}$$

If a confidence interval on $\theta = \mu_i - \mu_j$ is desired, then $\hat{\theta} = \overline{y}_{i\bullet} - \overline{y}_{j\bullet}$, and $m = 2/r$ are inserted in (6.13) because $a_i = 1$, $a_j = -1$, and the other $a_k = 0$ for $k \neq i, j$ in (6.14).

It was noted in Section 6.1 that the main-effect and interaction parameters are not uniquely defined in terms of the model means. As an illustration of this, consider the single-factor model $y_{ij} = \mu + \alpha_i + e_{ij}$. Rather than using $\mu_i = \mu + \alpha_i$ to define the main-effect parameters, define them as follows:

$$\mu_1 = \mu, \qquad \mu_i = \mu + \alpha_i, \qquad i = 2, 3, \ldots, a.$$

This is an equally meaningful way to define main-effect parameters. The main-effect parameters α_2 to α_a in this representation measure differences in factor effects from the mean of the first level rather than from an overall mean, as when $\mu_i = \mu + \alpha_i$.

Because of the nonuniqueness of the definitions of main-effect and interaction parameters, the estimation of these parameters depends on how the parameters are defined. For example, estimation of the individual parameters α_i for single-factor models depends on which definition is used to relate the parameters to the model means. If $\mu_2 = \mu + \alpha_2$ then $\alpha_2 = \mu_2 - \mu$; while in the alternative definition used above $\alpha_2 = \mu_2 - \mu_1$.

Because model means are unique, estimates of functions of them are also unique, even though these functions can be represented in different ways using the main-effect and interaction model parameters. Estimation of model means was stressed above so the estimation procedures could be discussed without special conventions that depend on how the main effects and interactions relate to the model means.

Main-effect and interaction parameters for ANOVA models defined as in (6.2) are used only to focus more clearly on estimation and testing procedures. Relationships between model means and these main-effect and interaction parameters are established in (6.3). Insertion of the averages for their corresponding model means in (6.3) yields the following estimators of the parameters:

$$\begin{aligned}
\hat{\mu} = \bar{y}_{\bullet\bullet\bullet\bullet},\ \hat{\alpha}_i &= \bar{y}_{i\bullet\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet},\ \hat{\beta}_j = \bar{y}_{\bullet j\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet},\ \hat{\gamma}_k = \bar{y}_{\bullet\bullet k\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet}, \\
\widehat{(\alpha\beta)}_{ij} &= \bar{y}_{ij\bullet\bullet} - \bar{y}_{i\bullet\bullet\bullet} - \bar{y}_{\bullet j\bullet\bullet} + \bar{y}_{\bullet\bullet\bullet\bullet}, \text{ etc.}, \\
\widehat{(\alpha\beta\gamma)}_{ijk} &= \bar{y}_{ijk\bullet} - \bar{y}_{ij\bullet\bullet} - \bar{y}_{i\bullet k\bullet} - \bar{y}_{\bullet jk\bullet} \\
&\quad + \bar{y}_{i\bullet\bullet\bullet} + \bar{y}_{\bullet j\bullet\bullet} + \bar{y}_{\bullet\bullet k\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet}.
\end{aligned} \tag{6.15}$$

These quantities are the calculated main effects and interactions for the experimental factors.

In Section 5.3, main effects and interactions for two-level factors are defined as differences in factor–level averages; e.g., $\bar{y}_{2\bullet\bullet\bullet} - \bar{y}_{1\bullet\bullet\bullet}$. The expressions in (6.15) are alternative to those in Section 5.3 and can be directly related to them; e.g., $\bar{y}_{2\bullet\bullet\bullet} - \bar{y}_{1\bullet\bullet\bullet} = \hat{\alpha}_2 - \hat{\alpha}_1$. The advantage to the representation (6.15) is that it is immediately extendable to any number of factors having any number of levels. These effects and the parameter estimates in (6.15) are also directly related to the sums of squares in an ANOVA tables, as is demonstrated in Section 6.3.2.

While (6.12) or (6.13) can be used to construct individual confidence intervals for any of the model means, the overall confidence that all the intervals simultaneously contain the means is not $100(1-\alpha)\%$ when two or more confidence intervals are computed. This is because the same data are being used to form many confidence intervals separately. The derivation of the confidence interval in Section 3.1 is based on a single population or process mean and uses a single sample of independent observations from that population or process. For a set of confidence intervals for which one desires a confidence

TABLE 6.6 Interval Estimates for Flow-Rate Model Parameters

Filter	Average	95% Confidence Interval on Model Mean μ_i
A	0.233	(0.191, 0.275)
B	0.291	(0.249, 0.333)
C	0.247	(0.205, 0.289)
D	0.274	(0.232, 0.316)

Filter Pair	Difference in Averages	95% Confidence Interval on Difference in Main Effects ($\alpha_i - \alpha_j$ or $\mu_i - \mu_j$)
A–B	−0.058	(−0.117, 0.001)
A–C	−0.014	(−0.073, 0.045)
A–D	−0.041	(−0.100, 0.018)
B–C	0.044	(−0.015, 0.103)
B–D	0.017	(−0.042, 0.076)
C–D	−0.027	(−0.086, 0.032)

coefficient of $100(1 - \alpha)\%$, the simultaneous inference procedure discussed in Section 6.4.3 should be used.

Individual confidence-interval estimates of the means and the differences in the factor–level effects for the flow-rate data of Figure 6.1 are displayed in Table 6.6. The filter averages and the estimated error variance can be obtained from the summary information in Table 6.5. The confidence intervals for the means are calculated using (6.12) with the appropriate substitutions for the factor–level means μ_i, averages $\overline{y}_{i\bullet}$, and sample sizes ($r = 4$). The confidence intervals for the differences in the main effects is calculated using (6.13) and (6.14), which are equivalent to the two-sided interval in Table 3.6 for pairwise differences in means.

We again stress that these individual intervals are presented for illustration purposes and that each interval does not have a confidence coefficient of 95%, because several intervals were formed from the same set of data. It is interesting to note that all of the intervals for the differences in factor effects include zero. This suggests that the factor effects are not significantly different from one another. A test statistic appropriate for testing this hypothesis is presented in Section 6.3.

6.2.3 Quantitative Factor Levels

The estimation techniques for main effects and interactions that were presented in the last section are applicable to the levels of any fixed-effects factor. When

factor levels are quantitative, however, it is of interest to assess whether factor effects are linear, quadratic, cubic, or possibly of higher order. To make such an assessment, one must define linear combinations of the response averages that measure these effects.

Although polynomial effects can be calculated for most balanced and unbalanced designs, we restrict attention in this section to balanced experiments in which factor levels are equally spaced. Quantitative levels of a factor X are equally spaced if consecutive levels differ by the same constant amount, that is, if, algebraically, the value of the ith level x_i can be expressed as $x_i = c_0 + c_1 i$ for suitably chosen constants c_0 and c_1. When designs are balanced, the coefficients of the linear combinations of the response averages that measure the linear, quadratic, etc. effects of the factor levels on the response can be conveniently tabulated. Table A9 in the Appendix contains coefficients for such polynomials.

To illustrate the calculation of polynomial factor effects, consider an experiment conducted to evaluate the warping of copper plates. The response variable is a measurement of the amount of warping, and the factors of interest are temperature (50, 75, 100, and 125°C) and the copper content of the plates (40, 60, 80, and 100%). Both of the factors are equally spaced. Both are fixed effects, since the levels of each were specifically chosen for inclusion in the design. It is of interest, assuming a smooth functional relationship between the factor levels and the mean of the response variable, to characterize quantitatively the relationship between the factor levels and the mean amount of warping.

Table 6.7 lists the data from the experiment. Observe from the analysis-of-variance table that the interaction between copper content and temperature is not statistically significant (see Section 6.3). The two main effects are statistically significantly ($p < 0.001$ for copper content, $p = 0.002$ for temperature).

Also included are scaled linear, quadratic, and cubic factor effects for copper content (those for temperature are left as an exercise). Because the coefficients in Table A9 of the linear, quadratic, and cubic effects are not all of the same magnitude, some effects can appear to be larger simply because they have larger coefficients. The scaling is performed to make the magnitudes of the effects comparable. Using the notation of (6.14), the scaled effects are

$$\hat{\theta}_s = \frac{1}{(D/n)^{1/2}} \sum_i a_i \bar{y}_{i\bullet}, \qquad D = \sum_i a_i^2, \tag{6.16}$$

where the a_i are the coefficients for one of the polynomial effects taken from Table A9 and n is the number of responses used in the calculation of each of the averages. The sum of squares D of the coefficients is also included for each polynomial effect in Table A9.

TABLE 6.7 Analysis of Polynomial Effects of Copper Content on Warping*

Warping Measurements

		Copper Content (%)				
		40	60	80	100	Total
Temperature (°C)	50	17, 20	16, 21	24, 22	28, 27	175
	75	12, 9	18, 13	17, 12	27, 31	139
	100	16, 12	18, 21	25, 23	30, 23	168
	125	21, 17	23, 21	23, 22	29, 31	187
	Total	124	151	168	226	669

ANOVA

Source of Variation	Degrees of Freedom (*df*)	Sum of Squares	Mean Square	*F*-Value
Copper Content	3	698.34	232.78	34.33
Temperature	3	156.09	52.03	7.67
Content × temp.	9	113.78	12.64	1.86
Error	16	108.50	6.78	
Total	31	1076.71		

Polynomial Coefficients

Copper Content (%)	Coefficient			Average
	Linear	Quadratic	Cubic	
40	−3	1	−1	15.50
60	−1	−1	3	18.88
80	1	−1	−3	21.00
100	3	1	1	28.25
Scaled effect	25.51	5.47	4.04	
95% Confidence Interval	(19.99, 31.03)	(−0.05, 10.99)	(−1.48, 9.56)	

*Data from Johnson, N. L., and Leone, F. C. (1977), *Statistics and Experimental Design in Engineering and the Physical Sciences*, New York: John Wiley & Sons, Inc. Copyright 1977 John Wiley & Sons, Inc. Used by permission.

Confidence intervals, comparable to (6.13) and (6.14), for the scaled polynomial effects are

$$\hat{\theta}_s - t_{\alpha/2} s_e < \theta_s < \hat{\theta}_s + t_{\alpha/2} s_e, \tag{6.17}$$

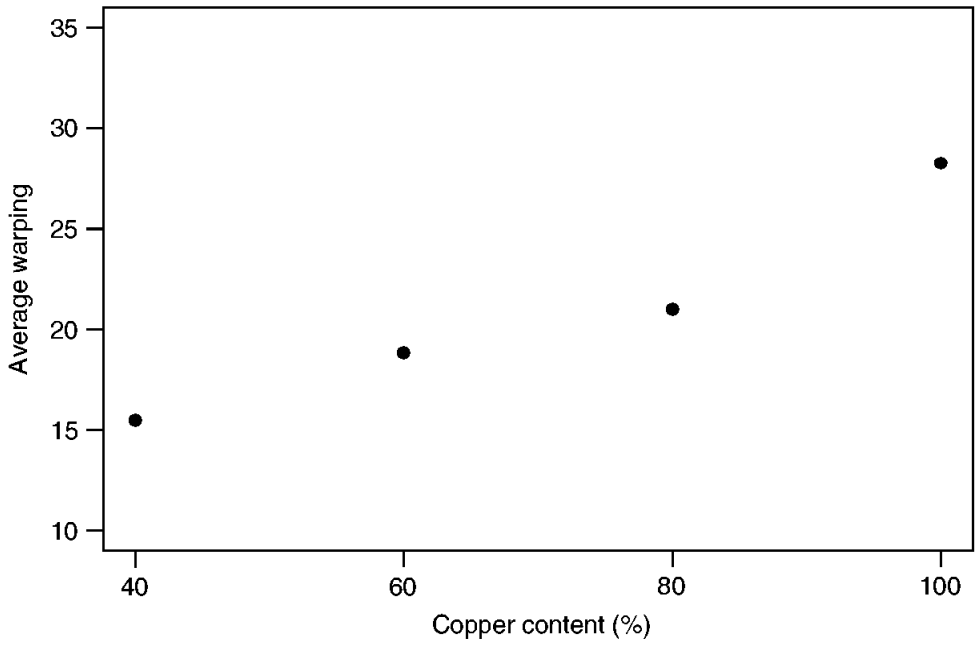


Figure 6.2 Average warping measurements.

when n is the number of responses entering into each average. The calculated confidence interval for each effect is shown in Table 6.7. Note from Table 6.7 that the scaled linear effect for copper content is quite large relative to the estimated model standard deviation $s = 2.60$. Note too that zero is included in the confidence intervals for the quadratic and the cubic parameter effects, but not for the linear effect. These two findings suggest that the dominant quantitative effect of copper content is linear. This suggestion is confirmed in Figure 6.2.

Polynomial effects can also be calculated for interactions between two equally spaced quantitative factors. The total number of such polynomial effects equals the product of the individual degrees of freedom for the main effects if the design is a balanced complete factorial. Each such polynomial effect is calculated by forming a table of coefficients, the elements of which are products of the main-effect coefficients.

In the construction of tables of coefficients for interaction effects, the rows of a table correspond to the levels of one of the factors and the columns correspond to the levels of the other factor. The body of the table contains the products, element by element, of the coefficients for each of the orthogonal polynomials.

For example, consider the copper-plate warping example. Linear, quadratic, and cubic main effects can be calculated for each factor. The nine degrees of freedom shown for the interaction sum of squares in Table 6.7 can be accounted for by nine joint polynomial effects. Letting copper content be factor A and temperature factor B, the nine quadratic effects are: linear $A \times$ linear B,

linear $A \times$ quadratic B, linear $A \times$ cubic B, quadratic $A \times$ linear B, quadratic $A \times$ quadratic B, quadratic $A \times$ cubic B, cubic $A \times$ linear B, cubic $A \times$ quadratic B, and cubic $A \times$ cubic B.

Table 6.8 lists the coefficients for each of the nine joint effects. To calculate scaled joint effects, multiply the coefficients in the table by the corresponding response averages, sum the products, and divide by the square root of the ratio of the sum of the squares of the coefficients to the number of repeat responses in each average, similarly to (6.16). The sum of the squares of the coefficients equals the product of the two D-values in Table A9 for the corresponding two main effects. This scaling of interaction effects is a straightforward extension of (6.16).

The scaled linear copper content by linear temperature interaction effect is

$$(A_L \times B_L) = \frac{9(18.5) + 3(18.5) - 3(23.0) - 9(27.5) + 3(10.5) + \cdots + 9(30.0)}{[(20)(20)/2]^{1/2}}$$

$$= -\frac{6.5}{(200)^{1/2}} = -0.46.$$

Note that this scaled effect is small relative to the linear main effect of copper content and to the estimated error standard deviation.

TABLE 6.8 Interaction Coefficients for Four-Level Factors

		Linear				Quadratic				Cubic			
		−3	−1	1	3	1	−1	−1	1	−1	3	−3	1
Linear	−3	9	3	−3	−9	−3	3	3	−3	3	−9	9	−3
	−1	3	1	−1	−3	−1	1	1	−1	1	−3	3	−1
	1	−3	−1	1	3	1	−1	−1	1	−1	3	−3	1
	3	−9	−3	3	9	3	−3	−3	3	−3	9	−9	3
Quadratic	1	−3	−1	1	3	1	−1	−1	1	−1	3	−3	1
	−1	3	1	−1	−3	−1	1	1	−1	1	−3	3	−1
	−1	3	1	−1	−3	−1	1	1	−1	1	−3	3	−1
	1	−3	−1	1	3	1	−1	−1	1	−1	3	−3	1
Cubic	−1	3	1	−1	−3	−1	1	1	−1	1	−3	3	−1
	3	−9	−3	3	9	3	−3	−3	3	−3	9	−9	3
	−3	9	3	−3	−9	−3	3	3	−3	3	−9	9	−3
	1	−3	−1	1	3	1	−1	−1	1	−1	3	−3	1

6.3 STATISTICAL TESTS

Tests of statistical hypothesis can be performed on the parameters of ANOVA models. These tests provide an alternative inferential methodology to the interval-estimation procedures discussed in the last section. In this section several commonly used statistical tests are discussed. We separate this discussion according to whether tests are desired for (a) a single parameter or a single function of the model parameters or (b) groups of parameters or parametric functions.

6.3.1 Tests on Individual Parameters

Tests of hypotheses on individual model means are straightforward extensions of the single-sample t-tests of Section 3.1. For example, in the ANOVA model for a three-factor balanced complete factorial experiment a test of $H_0: \mu_{ijk} = c$ versus $H_a: \mu_{ijk} \neq c$, where c is a specified constant (often zero), is based on the t-statistic

$$t = r^{1/2} \frac{\bar{y}_{ijk\bullet} - c}{s_e}, \tag{6.18}$$

where $s_e = (\mathrm{MS_E})^{1/2}$. Hypotheses about other model means can be tested by making the appropriate substitutions in (6.18).

Testing hypotheses about a single linear combination of model means is accomplished in a similar manner. The hypothesis

$$H_0: \theta = c \quad \text{vs} \quad H_a: \theta \neq c,$$

where $\theta = \sum_i \sum_j \sum_k a_{ijk} \mu_{ijk}$, is tested using the single-sample t-statistic

$$t = \frac{\hat{\theta} - c}{s_e m^{1/2}}, \tag{6.19}$$

where $\hat{\theta} = \sum_i \sum_j \sum_k a_{ijk} \bar{y}_{ijk\bullet}$, $m = \sum_i \sum_j \sum_k a_{ijk}^2 / r$, and r is the number of repeat tests for each factor–level combination. Note the equivalence of this test procedure with the confidence-interval approach using (6.13).

Tests of the statistical significance of polynomial effects can also be performed using (6.19) and the coefficients in Table A9. If the scaled form (6.16) of the polynomial effect is used, the t-statistic (6.19) is simply $\hat{\theta}_s / s_e$.

A common application of (6.19) occurs when one wishes to compare two factor level effects; for example $H_0: \alpha_1 - \alpha_2 = 0$ versus $H_a: \alpha_1 - \alpha_2 \neq 0$. In this case $\theta = \alpha_1 - \alpha_2$ and the t-statistic (6.19) is

$$t = \frac{\bar{y}_{1\bullet\bullet\bullet} - \bar{y}_{2\bullet\bullet\bullet}}{s_e (2/r)^{1/2}}.$$

6.3.2 F-Tests for Factor Effects

One of the difficulties with performing separate t-tests for each main effect or each interaction effect is that the overall chance of committing one or more Type I errors can greatly exceed the stated significance level for each of the individual tests. In part this is due to multiple testing using the same data set. There are a number of test procedures that can be used to simultaneously test the equality of all the main effects for a factor or for all the interaction effects of two or more factors. In this section we discuss F-tests based on the mean squares from an ANOVA table.

The numerators of the F-statistics in fixed effects ANOVA tables are the main effects and interaction mean squares. The sums of squares for these main effects and interactions in a balanced three-factor complete factorial experiment can be written as in Equations (6.7)–(6.9). Comparison of these sums of squares with the factor effects in (6.15) reveals that the sums of squares are functions of the estimated effects parameters when the parametrization (6.2) is used to relate the effects parameters to the model means; e.g.,

$$\mathrm{SS}_A = bcr \sum_i \hat{\alpha}_i^2, \qquad \mathrm{SS}_B = acr \sum_j \hat{\beta}_j^2, \qquad \mathrm{SS}_{AB} = cr \sum_i \sum_j (\widehat{\alpha\beta})_{ij}^2, \text{ etc.}$$

The sums of squares in an ANOVA table test the hypothesis that the parameters corresponding to the main effects and interactions are zero. For example, SS_A tests the hypothesis $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$ versus $H_a: \alpha_i \neq 0$ for at least one factor level i. The equivalent hypothesis in terms of the model means is $H_0: \overline{\mu}_{1\bullet\bullet} = \overline{\mu}_{2\bullet\bullet} = \cdots = \overline{\mu}_{a\bullet\bullet}$ versus $H_a: \overline{\mu}_{i\bullet\bullet} \neq \overline{\mu}_{j\bullet\bullet}$ for at least one pair of factor levels.

Under the hypothesis that a particular main effect or interaction is zero, the corresponding F-ratio should be around 1, because both the numerator and the denominator of the F-statistic are estimating the same quantity, the error variance. On the other hand, if the stated null hypothesis is false, the numerator mean square will tend to be larger than the error mean square. Thus, large F-ratios lead to rejection of the hypotheses of no factor effects.

The analysis-of-variance table for the flow-rate data is shown in Table 6.5. The main effect for the four filters has an F-ratio of 1.86. Comparison of this F-ratio with F critical values in Table A5 reveals that the filter effects are not statistically significant ($0.10 < p < 0.25$). Thus, the mean flow rates attributable to the filters do not significantly differ for the three filters. Stated another way, the response variability attributable to the filter means is not significantly greater than the variability attributable to uncontrolled experimental error.

The ANOVA table for the pilot-plant study is shown in Table 6.4. Using a significance level of $\alpha = 0.05$, the temperature-by-catalyst interaction and the main effect of concentration are statistically significant. In keeping with the

hierarchical modeling of the response, we ignore whether the main effects of temperature and catalyst are statistically significant. The reason we ignore the test for these main effects is that the significant interaction of these two factors indicates that they have a joint influence on the response; consequently, there is no need to examine the main effects.

The next step in the analysis of these data would be to examine which of the factor levels are affecting the response. For example, one would now like to know which combinations of temperature and catalyst produced significantly higher or lower yields than other combinations. The multiple-comparison procedures discussed in the next section should be used for this purpose. We defer detailed consideration of the effects of individual factor–level combinations until we discuss some of these procedures.

6.4 MULTIPLE COMPARISONS

Detailed exploration of data often stems from a desire to know what effects are accounting for the results obtained in an experiment. One's interest thus is directed to the comparison of specific factor–level means or of groups of means rather than strictly to the detection of a statistically significant main effect or interaction. In these situations, procedures utilizing multiple comparisons of means are appropriate.

Multiple comparisons of means frequently involve preselected comparisons that address specific questions of interest to a researcher. In such circumstances one sometimes has little interest in the existence of overall experimental effects; consequently, the summaries provided by an ANOVA table are of secondary interest, perhaps only to provide an estimate of the experimental-error variance. In contrast, researchers in many experimental settings do not know which factor effects may turn out to be statistically significant. If so, the F-statistics in an ANOVA table provide the primary source of information on statistically significant factor effects. However, after an F-test in an ANOVA table has shown significance, an experimenter usually desires to conduct further analyses to determine which pairs or groups of means are significantly different from one another.

Both of the above types of comparisons are examined in this section. Specific attention is given to comparisons involving quantitative factor levels and comparisons based on t-statistics.

6.4.1 Philosophy of Mean-Comparison Procedures

The estimation of linear combinations of means is detailed in this section to aid in the understanding of the philosophy of multiple-comparison procedures. There is a close connection between the material in this section and the discussion in Sections 6.2 and 6.3.

Analysis of linear combinations of means is a major objective of many experiments. For example, one may be interested in comparing the average fuel economy, $\bar{y}_1$, achieved by a test oil in a laboratory experiment with the averages, $\bar{y}_2$ and $\bar{y}_3$, of the fuel economies of two different reference oils. A linear combination of the sample means that would be used for the comparison is $\bar{y}_1 - (\bar{y}_2 + \bar{y}_3)/2$. Linear combinations of means such as these are termed *contrasts* (see Exhibit 6.5).

EXHIBIT 6.5

Contrast. A linear combination of k averages, denoted by

$$a_1\bar{y}_1 + a_2\bar{y}_2 + \cdots + a_k\bar{y}_k,$$

where $\bar{y}_i$ is the ith average and the a_i are constants, at least two of which are nonzero, is termed a contrast if the coefficients (the a's) sum to zero, that is, if

$$a_1 + a_2 + \cdots + a_k = 0.$$

In Chapter 5 we defined main effects and interactions as contrasts of factor–level averages (see Section 5.3 and the appendix to Chapter 5). In Section 6.2 we related main effects, interactions, and other factor effects to linear combinations of the model means and to linear combinations of the main-effect and the interaction model parameters. One generally wishes to draw inferences on linear combinations of means or model parameters using the corresponding linear combinations of factor–level averages. Contrasts are of special importance because only contrasts of the main-effect and interaction parameters can be estimated in an ANOVA model.

Consider the single-factor ANOVA model:

$$y_{ij} = \mu + \alpha_i + e_{ij}, \qquad i = 1, 2, \ldots, a, \quad j = 1, 2, \ldots, r, \tag{6.20}$$

where the model parameters are related to the factor–level means through the relationship $\mu_i = \mu + \alpha_i$.

Suppose that one is interested in estimating some linear combination of the factor–level means $\theta = \Sigma a_i \mu_i$. Estimation of this linear combination of means is accomplished by inserting the factor–level averages for the means as in (6.14). Observe that in terms of the model parameter in (6.20),

$$\theta = \sum a_i\mu_i = \mu \sum a_i + \sum a_i\alpha_i = \sum a_i\alpha_i \qquad \text{if} \quad \sum a_i = 0.$$

The reason that contrasts are so important in the comparison of model means is that the constant term μ appears in all linear combinations of model means

except for contrasts, in which case $\sum a_i = 0$. Thus, comparisons of the effects of individual factor levels must be made using contrasts.

The use of contrasts is not a limitation on comparisons of factor–level means. Note that a direct comparison of two means is accomplished by estimating the difference in the means. Such differences are contrasts. Comparisons of three or more means can be made by pairwise comparisons of differences of the means or any other contrasts that are deemed informative. We examine several such informative contrasts below.

The above comparison between the test oil and the two reference oils using the linear combination $\bar{y}_1 - (\bar{y}_2 + \bar{y}_3)/2$ is a contrast, since the sum of the coefficients is zero. To avoid fractional coefficients, it is common to rewrite the comparison as $2\bar{y}_1 - (\bar{y}_2 + \bar{y}_3)$. Because averages are estimators of fixed-effects portions of ANOVA models, these linear combinations of averages are estimators of the same linear combinations of model means.

For a fixed-effects model of the form (6.20) containing only oils as a factor, α_i denotes the fixed-effect model parameter corresponding to the ith oil. The above linear combination of averages is an estimator of

$$2(\mu + \alpha_1) - [(\mu + \alpha_2) + (\mu + \alpha_3)] = 2\alpha_1 - (\alpha_2 + \alpha_3).$$

Note that this linear combination of model parameters does not contain the constant term μ, and it is zero when all three factor levels have an equal effect ($\alpha_1 = \alpha_2 = \alpha_3$) on the response. These two characteristics are present in all contrasts of factor effects.

When making multiple comparisons of factor effects, an additional property of contrasts is needed. To make these comparisons statistically independent of one another, the contrasts involved must be mutually orthogonal (see Exhibit 6.6).

EXHIBIT 6.6

Orthogonal Contrasts. Two contrasts,

$$C_1 = \sum a_i \bar{y}_i \quad \text{and} \quad C_2 = \sum b_i \bar{y}_i$$

are said to be orthogonal if the sum of the products of the corresponding coefficients in the two contrasts is zero, that is, if

$$a_1 b_1 + a_2 b_2 + \cdots + a_k b_k = 0.$$

Three or more contrasts are said to be mutually orthogonal if all pairs of contrasts are orthogonal.

To illustrate the use of orthogonal contrasts, consider the data in Table 6.1 (see also Figure 5.3) on the chemical yield of a manufacturing process in a pilot plant. For simplicity, we will examine only two factors, catalyst (A) and concentration (B), each of which has two levels. If a 2^2 factorial experiment (see Chapter 5) had been conducted, three comparisons of interest among the four factor–level means μ_{ij} might be:

(i) the effect of the first catalyst compared with that of the second catalyst,
(ii) the effect of the high level of concentration compared with that of the low level of concentration, and
(iii) the difference in the effects of the two catalysts at the high concentration compared with their difference at the low concentration.

Denote the average responses for the four factor–level combinations as $\bar{y}_{11}, \bar{y}_{12}, \bar{y}_{21}$, and $\bar{y}_{22}$. In this representation, the first subscript refers to the catalyst (1 = catalyst 1, 2 = catalyst 2) and the second refers to the concentration (1 = 20%, 2 = 40%). The above three comparisons can now be made using these factor–level averages to estimate the corresponding model means. The comparisons of interest can then be expressed as:

(i) $(\bar{y}_{21} + \bar{y}_{22})/2 - (\bar{y}_{11} + \bar{y}_{12})/2$,
(ii) $(\bar{y}_{12} + \bar{y}_{22})/2 - (\bar{y}_{11} + \bar{y}_{21})/2$,
(iii) $(\bar{y}_{22} - \bar{y}_{12}) - (\bar{y}_{21} - \bar{y}_{11})$.

The coefficients in these three comparisons, apart from the divisor of 2 in the first two, are:

	Factor–Level Response Average			
Contrast	$\bar{y}_{11}$	$\bar{y}_{12}$	$\bar{y}_{21}$	$\bar{y}_{22}$
(i)	−1	−1	+1	+1
(ii)	−1	+1	−1	+1
(iii)	+1	−1	−1	+1

Notice that the sum of the coefficients of each of these comparisons is zero, indicating that each is a contrast. Further, the products of the corresponding coefficients of any two contrasts sum to zero, indicating that the three contrasts are mutually orthogonal. Finally, note that the above contrasts are the effects representations of the main effects and interaction for two two-level factors shown in Table 5.6.

The sums of squares corresponding to fixed effects in any ANOVA table can be partitioned into component sums of squares, each component of which corresponds to one of the degrees of freedom. This partitioning corresponds to the orthogonal contrasts that can be formed from the means that go into the sum of squares. For main effects, there are $k - 1$ mutually orthogonal contrasts that can be formed from the means for the k levels of a factor. The degrees of freedom for interactions are design dependent, but frequently (e.g., for complete factorial experiments) the number of degrees of freedom and the number of mutually orthogonal contrasts equal the product of the numbers of degrees of freedom for the main effects corresponding to the factors in the interaction.

For a set of averages each of which consists of n observations, the formula for the sum of squares corresponding to a contrast $C = \sum a_i \bar{y}_i$ is given by

$$\mathrm{SS}(C) = \frac{n\left(\sum a_i \bar{y}_i\right)^2}{\sum a_i^2}. \tag{6.21}$$

This sum of squares, because it has only one degree of freedom, is a mean square. Divided by the error mean square from the ANOVA table for these response variables, the ratio $\mathrm{SS}(C)/\mathrm{MS}_E$ is an F-statistic, which can be used to test the hypothesis

$$H_0: \sum a_i \mu_i = 0 \quad \text{vs} \quad H_a: \sum a_i \mu_i \neq 0, \tag{6.22}$$

where μ_i is the mean of the ANOVA model corresponding to the average $\bar{y}_i$. The degrees of freedom of this F statistic are $\nu_1 = 1$ and $\nu_2 = \nu$, where ν is the number of degrees of freedom of the error mean square.

This F-statistic is exactly equivalent to the use of a t-statistic with the effects calculated as in (6.14) or (6.15). For example, squaring the t-statistic (6.19) for H_0: $\theta = 0$ yields,

$$t^2 = \left(\frac{\hat{\theta}}{s_e m^{1/2}}\right)^2 = \frac{\mathrm{SS}(C)}{\mathrm{MS}_E}.$$

This equivalence allows the estimation procedures detailed in Section 6.2 or the testing procedure outlined in Section 6.3 to be applied to the general discussions in this section.

It is possible to obtain a nonsignificant overall F-statistic for a main effect or interaction in an ANOVA table, yet find that one or more of the component contrasts are statistically significant. This result generally occurs when one or two of the orthogonal contrasts are significant but the remaining ones are not. When totaled to give the sum of squares for the effect, the nonsignificant

contrasts dilute the significant one(s), yielding a nonsignificant main effect or interaction. Similarly, it is possible to have a significant overall F-statistic for a factor yet find that no pairwise comparison between any two averages is significant.

Mutual orthogonality is not a prerequisite for analyzing a set of contrasts. An experimenter may choose to examine whichever set of linear combinations of means is of interest or value. Orthogonal contrasts are usually selected because (a) contrasts eliminate the constant term from mean comparisons, (b) comparisons of interest among factor–level means can usually be expressed as contrasts, and (c) the sums of squares calculated from orthogonal contrasts are statistically independent for balanced experimental designs.

A critical question in making multiple comparisons concerns the choice of the significance level of the tests or the confidence level for interval estimates. Because several statistical inference procedures (several interval estimates or several tests) are to be made on the same data set, a distinction must be made between two types of error rates. We describe these Type I error rates (see Exhibit 6.7) in terms of significance levels for tests of hypotheses, but they are equally germane to interval estimation because of the equivalence between statistical tests and interval estimation that was established in the preceding chapters (e.g., Chapter 2). We return to this point below.

EXHIBIT 6.7 TYPE I ERROR RATES

Comparisonwise Error Rate. The probability of erroneously rejecting a null hypothesis when making a single statistical test.

Experimentwise Error Rate. The probability of erroneously rejecting at least one null hypothesis when making statistical tests of two or more null hypotheses using the data from a single experiment.

Comparisonwise Type I error rates essentially indicate the significance level associated with a single statistical test. All previous discussions of significance levels have been comparisonwise error rates. Experimentwise Type I error rates measure the significance level associated with multiple tests using the same data set. When making several statistical tests using the same data set, the experimentwise Type I error rate can be much larger than the significance level stated for each test.

An experimentwise error rate of 0.05 is much more stringent than a comparisonwise error rate of 0.05. Consider the comparison of k population means. If statistically independent contrasts are used to make $k - 1$ comparisons among the sample averages and the error standard deviation is assumed known, the experimentwise error rate E is related to the comparisonwise error rate (significance level of each comparison) through the formulas

$$E = 1 - (1 - \alpha)^{k-1} \quad \text{or} \quad \alpha = 1 - (1 - E)^{1/(k-1)}. \tag{6.23}$$

For example, suppose the comparisonwise significance level is selected to be $\alpha = 0.05$ and $k = 6$ means are to be compared. If all six population means are equal, the probability of incorrectly rejecting one or more of the five orthogonal comparisons is

$$E = 1 - (1 - 0.05)^5 = 0.23,$$

a value almost five times larger than the stated significance level for each test. If one desires an experimentwise significance level of $E = 0.05$, the individual comparisonwise significance levels should be

$$\alpha = 1 - (1 - 0.05)^{0.2} = 0.01.$$

This example illustrates that an experimentwise error rate yields fewer Type I errors than the same comparisonwise error rate. This advantage is counterbalanced by the fact that more Type II errors are likely to occur when controlling the experimentwise error rate. The formulas in (6.23) are not valid when several nonorthogonal comparisons among means are made with the same set of data. In such cases, the experimentwise error rate can be larger than that indicated by equation (6.23). Likewise, these formulas are not strictly valid when the same error mean square is used in the comparisons; however, these formulas are still used as approximations to the true error rates in such situations.

In practice, an experimenter must choose whether to control the comparisonwise or the experimentwise error rate. The more appropriate error rate to control depends on the degree to which one wants to control the Type I error rate. In situations involving sets of orthogonal contrasts, the comparisonwise error rate is often selected as the preferred choice when there are not a large number of comparisons to be made. When there are many comparisons to make or the comparisons are not based on orthogonal contrasts, the experimentwise error rate usually is controlled. This is especially true when using nonorthogonal contrasts because then the outcome of one comparison may affect the outcomes of subsequent comparisons.

The connection of this discussion with interval estimation is readily established. A Type I error occurs for a statistical test when the null hypothesis is erroneously rejected. The equivalent error for confidence intervals occurs when the confidence interval does not include the true value of the parameter. Thus comparisonwise error rates α and experimentwise error rates E are related to individual confidence levels $100(1 - \alpha)\%$ for one confidence interval and overall confidence levels $100(1 - E)\%$ for two or more intervals formed from one data set.

While an experimenter should be concerned about controlling overall error rates and confidence levels, it is important to stress that the main objective

of multiple comparison procedures is to learn as much about populations, processes, or phenomena as is possible from an experiment. A low error rate is not the sole purpose of an experiment; it is a guide to careful consideration of objectives and to an awareness of the proper use of statistical methodology. In a particular experiment an acceptably low error rate may be 20% or it may be 0.1%.

6.4.2 General Comparisons of Means

There are many ways to select specific comparisons of model means, although the choice of an experimental design effectively determines which contrasts are available for analysis. If one intends to conduct an analysis of variance of the data, one might choose to partition the degrees of freedom for one or more of the main effects or interactions into a specific set of mutually orthogonal, single-degree-of-freedom contrasts. In an analysis involving a quantitative factor one might decide to use a partitioning of the sum of squares into a linear effect, a quadratic effect, a cubic effect, and so forth. In studies involving a control group, one might choose to make select comparisons of the treatment groups with the control group, as well as comparing the treatment groups with one another. Finally, suggestions for comparisons might result from plots of the experimental results.

To illustrate how specific comparisons can be made, consider again the wire-die example given in Table 2.2. The purpose of the experiment was to compare the tensile strengths of wire samples taken from three different dies. Suppose that die 3 is a control die and the other two dies are new ones that have different geometric designs that could potentially allow higher production rates. Two comparisons which might be of interest are:

(1) The mean tensile strength of wire made from die 3 versus the mean tensile strength of wire made from dies 1 and 2, and

(2) the mean tensile strength of wire made from die 1 versus that for wire made from die 2.

Note that the first comparison examines whether the experimental dies differ from the standard [i.e., whether $\mu_3 = (\mu_1 + \mu_2)/2$], while the second one determines whether the two experimental dies differ [that is, whether $\mu_1 = \mu_2$].

The comparisons stated above can be made using contrasts among the three average tensile strengths:

(1) $C_1 = (\overline{y}_1 + \overline{y}_2)/2 - \overline{y}_3$, and

(2) $C_2 = \overline{y}_1 - \overline{y}_2$.

Writing these contrasts as

$$C_1 = \sum a_i \bar{y}_i \qquad \text{with} \quad (a_1, a_2, a_3) = (0.5, 0.5, -1),$$
$$C_2 = \sum b_i \bar{y}_i \qquad \text{with} \quad (b_1, b_2, b_3) = (1, -1, 0),$$

one can see that these contrasts are orthogonal. The sums of squares corresponding to these two contrasts constitute an additive partition of the sum of squares for the die factor.

Using equation (6.21) and the summary statistics given in Table 2.3, we find that the sums of squares for contrasts (1) and (2) are given by

$$\text{SS}(C_1) = 18\frac{\{(84.733 + 84.492)/2 - 82.470\}^2}{1 + 0.25 + 0.25} = 55.067,$$
$$\text{SS}(C_2) = 18\frac{(84.733 - 84.492)^2}{1 + 1} = 0.525.$$

These sums of squares add up to the sum of squares due to the die factor shown in Table 6.9:

$$\text{SS}_D = \text{SS}(C_1) + \text{SS}(C_2) = 55.067 + 0.525 = 55.592.$$

We can now test the significance of these two contrasts by dividing each by the error mean square in Table 6.9. The two F-statistics are

$$F(C_1) = \frac{\text{MS}(C_1)}{\text{MS}_E} = \frac{55.067}{2.181} = 25.249 \quad (p < 0.001),$$
$$F(C_2) = \frac{\text{MS}(C_2)}{\text{MS}_E} = \frac{0.525}{2.181} = 0.241 \quad (p > 0.50).$$

TABLE 6.9 Analysis-of-Variance Table for Tensile-Strength

Experiment Source	df	SS	MS	F
Dies	2	55.592	27.796	12.742
Error	51	111.254	2.181	
Total	53	166.846		

Because there are only two comparisons to be made, comparisons that are orthogonal contrasts, a comparisonwise error rate can be used to test these hypotheses. Any reasonable choice of a significance level will result in the rejection of the first hypothesis and the nonrejection of the second one. Thus, there is not sufficient evidence from these data to conclude that the mean tensile strengths of wire made from dies 1 and 2 differ from one another. There is, however, sufficient evidence to conclude that the mean tensile strength of wire made from die 3 is different from the average of the means for the other two dies.

As mentioned above, multiple comparisons need not be orthogonal. They can be any comparisons that are of interest to an experimenter, provided that the data and the model allow them to be analyzed. In the tensile-strength example we could have chosen to compare each of the test dies with the standard die. The two contrasts then would have the form

(3) (control versus die 1) $C_3 = \overline{y}_1 - \overline{y}_3$, and
(4) (control versus die 2) $C_4 = \overline{y}_2 - \overline{y}_3$.

These are not orthogonal contrasts, because the coefficients for the two linear combinations of the three average tensile strengths, when multiplied together, do not sum to zero:

$$(+1)(0) + (0)(+1) + (-1)(-1) = 1.$$

Nevertheless, F-statistics appropriate for testing that the corresponding contrasts of the factor–level means are zero would be calculated in the same manner as those for orthogonal contrasts. Using equation (6.21) and the summary statistics in Table 2.3,

$$\mathrm{SS}(C_3) = 18\frac{(84.733 - 82.470)^2}{1+1} = 46.088,$$

$$\mathrm{SS}(C_4) = 18\frac{(84.492 - 82.470)^2}{1+1} = 36.796.$$

Using the MS_E from Table 6.9, the F-statistics are

$$F(C_3) = \frac{\mathrm{MS}(C_3)}{\mathrm{MS}_E} = \frac{46.088}{2.182} = 21.127 \quad (p < 0.001),$$

$$F(C_4) = \frac{\mathrm{MS}(C_4)}{\mathrm{MS}_E} = \frac{36.796}{2.182} = 16.858 \quad (p < 0.001).$$

Because these two contrasts are not orthogonal, it is appropriate to use an experimentwise error rate rather than a comparisonwise error rate. Letting $E =$

0.05, we use equation (6.23) to obtain a comparisonwise significance level of 0.025 for testing these two contrasts. Because both significance probabilities are less than 0.001, we reject both of the hypotheses; that is, we conclude that each of the test dies produces wire that has an average tensile strength different from that of the standard die. An examination of the averages in Table 2.3 indicates that the experimental dies produce wire that has significantly higher average tensile strengths than the standard die.

The steps involved in multiple comparisons of means are summarized in Exhibit 6.8 in a general way.

As mentioned in the previous section, there is an equivalence between the multiple-comparison testing procedures described in this section and the use of multiple confidence-interval procedures. Multiple confidence-interval procedures provide important information on the precision of calculated effects and are often more informative than a series of statistical tests. The philosophy of multiple comparisons described in this section is applicable to multiple confidence intervals if the individual confidence intervals are calculated using a confidence coefficient of $100(1-\alpha)\%$, where α is determined from (6.23). The overall confidence coefficient of the multiple interval is then approximately $100(1-E)\%$.

EXHIBIT 6.8 MULTIPLE COMPARISONS

1. State the comparisons (hypotheses) of interest in a form similar to (6.22).
2. Calculate the individual sums of squares for the comparisons using equation (6.21). (Equivalently, the t-statistics discussed in Sections 6.2 and 6.3 can be used.)
3. Divide the comparison sums of squares by the mean squared error from the ANOVA table to form F-statistics for testing the respective hypotheses.
4. Choose an appropriate comparisonwise or experimentwise error rate based on
   (a) whether the comparisons are orthogonal contrasts, and
   (b) the number of comparisons to be made.
5. Compare the significance probabilities with the selected significance level and draw the resulting conclusions.

In Section 6.4.3 we introduce alternative multiple-comparison techniques to the general ones described in this section. Prior to doing so, we wish to return to the investigation of quantitative factor levels that was discussed in Section 6.2. We return to this topic because of the importance of modeling the effects of quantitative factor levels on a response when the response can be assumed to be a smooth function of the factor levels. Testing of the significance of polynomial effects is a frequent application of multiple-comparison procedures.

The procedures that were illustrated above are applicable whether the experimental factor is qualitative or quantitative. When a quantitative factor is included in an experiment, interest often centers on determining whether the quantitative factor has a linear, quadratic, cubic, etc. effect on the response. To make such a determination, one must define linear combinations of the response averages that measure these effects.

The coefficients a_i used in equation (6.21) to measure an rth-degree polynomial effect on a response are obtained from a linear combination of the first r powers of the quantitative factor:

$$a_i = b_1 x_i + b_2 x_i^2 + \cdots + b_r x_i^r \tag{6.24}$$

where a_i is the coefficient on $\bar{y}_i$ for the rth-degree polynomial effect and x_i^k is the ith level of the factor raised to the kth power. When these coefficients a_i are selected so that the rth polynomial (e.g., cubic, $r = 3$) is orthogonal to all lower-order polynomials (e.g., linear, $r = 1$, and quadratic, $r = 2$), the resulting coefficients are referred to as *orthogonal polynomials*.

When a factor has equally spaced levels in a balanced design, the fitting of orthogonal polynomial terms can be accomplished similarly to the calculation of orthogonal contrasts. Tables of coefficients a_i for orthogonal polynomials are available in Table A9 of the Appendix. These tables provide the a_i needed in equation (6.21) to compute the effects (linear, quadratic, cubic, etc.) of the factor on the response (see also Section 6.2.3). The tables are constructed so that the coefficients are orthogonal contrasts; consequently, all the procedures described above for contrasts involving qualitative factors are applicable.

Consider an experiment conducted to examine the force (lb) required to separate a set of electrical connectors at various angles of pull. The force exerted is the response variable, and the factor of interest is the pull angle. This factor has four equally spaced levels: 0, 2, 4, and 6°. The experiment was conducted using five connectors and five repeat tests for each angle with each connector. The factor–level averages, based on 25 observations per angle, were as follows:

Angle (deg)	Average Force (lb)
0	41.94
2	42.36
4	43.82
6	46.30

Just as there are only $k - 1$ orthogonal contrasts for a qualitative factor having k levels, one only can fit a polynomial expression up to degree $k - 1$ for a quantitative factor having k equally spaced levels. Thus, for the pull angle,

we can fit up to a cubic polynomial. The coefficients for the linear, quadratic, and cubic terms in the polynomial are, from Table A9 of the appendix,

$$\begin{aligned} \text{Linear}(L): &\quad (-3, -1, 1, 3), \\ \text{Quadratic (Q)}: &\quad (1, -1, -1, 1), \\ \text{Cubic (C)}: &\quad (-1, 3, -3, 1). \end{aligned}$$

Using these coefficient with the above averages in equation (6.21) yields the following sums of squares for the orthogonal polynomials:

$$\text{SS}(L) = 264.26, \quad \text{SS}(Q) = 26.52, \quad \text{SS}(C) = 0.01.$$

The appropriate mean square for testing the significance of these polynomial terms is the interaction mean square between connectors (B) and the pull angles (A), $\text{MS}_{AB} = 37.92$ with 12 degrees of freedom (see Chapter 10). Dividing each of the above sums of squares by this mean square yields the following F-statistics:

$$F(L) = 6.97, \quad F(Q) = 0.70, \quad F(C) = 0.0003.$$

Because this is a small number of mutually orthogonal contrasts, we select a comparisonwise error rate of 0.05 and use an F-table with 1 and 12 degrees of freedom. Only the linear effect of pull angle is statistically significant ($p < 0.001$). Hence, as the pull angle increases, the separation force increases at approximately a linear rate.

It should be noted that comparisons can be composed of groups of orthogonal contrasts rather than merely of individual orthogonal contrasts. When using orthogonal contrasts the sum of squares for the group comparison is obtained by adding the sums of squares for the individual contrasts. The mean square for the group comparison is obtained by dividing this sum of squares by the number of contrasts (that is, degrees of freedom). An F-statistic for testing the significance of this group comparison can be calculated by dividing this mean square by the error mean square.

Figure 6.3 is a plot of the average force measurements as a function of pull angle. This figure well illustrates the desirability of accompanying statistical tests with estimation and plotting procedures. The above tests indicated that the only statistically significant polynomial effect of pull angle is linear. However, the plot clearly suggests a quadratic effect. The reason the quadratic effect is not statistically significant is that orthogonal polynomials adjust each effect for lower-order effects to make the contrasts orthogonal. Consequently, once the force averages are adjusted for the linear effect of pull angle, the quadratic effect is not statistically significant.

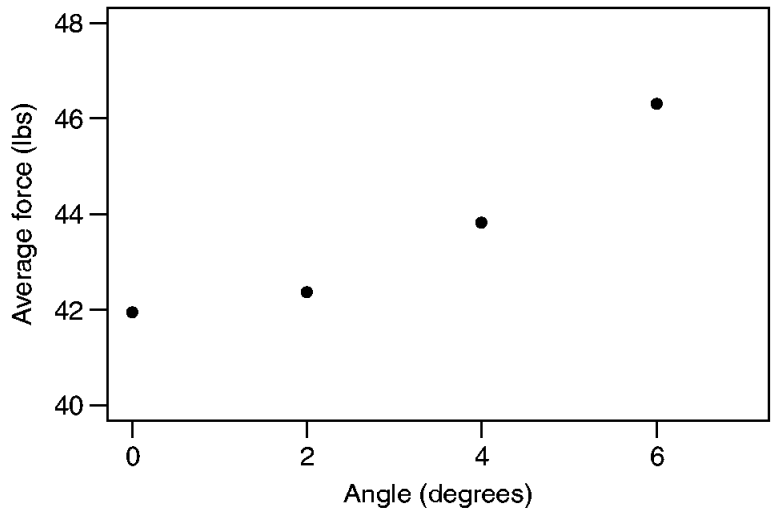


Figure 6.3 Average force versus pull angle.

As an indication that there is a statistically significant quadratic effect due to pull angle, we can calculate the combined effect of the linear and the quadratic effects. The sum of squares for the combined effects is the sum of the component sums of squares (because the effects are based on orthogonal contrasts), and the mean square is this combined sum of squares divided by 2. The resulting F-statistic, with 2 and 12 degrees of freedom, is 3.83. This F-statistic has a p-value of about 0.05. As an alternative to this analysis, one can fit polynomial regression models (Chapter 15) to the force values and assess the desirability of including a quadratic effect due to pull angle.

6.4.3 Comparisons Based on t-Statistics

The most commonly used multiple comparison tests are those based on t-statistics. Chief among these are the tests for contrasts presented in the last section. As mentioned above, rather than using an F-statistic, a contrast can be tested using a t-statistic. The tests are exactly equivalent. For testing the hypotheses shown in equation (6.22), the test statistic formed from (6.14) can be written as

$$t = \frac{\sum a_i \bar{y}_i}{\mathrm{SE}_c}, \tag{6.25}$$

where SE_c, the estimated standard error of the contrast, is given by

$$\mathrm{SE}_c = \left[\mathrm{MSE} \left(\sum \frac{a_i^2}{r} \right) \right]^{1/2}. \tag{6.26}$$

Contrasts tested using (6.25) and (6.26) often involve the pairwise comparisons of means. One of the oldest and most popular techniques for making multiple pairwise comparisons of means is Fisher's least-significant-difference (LSD) procedure (see Exhibit 6.9). It is termed a *protected* LSD test if the procedure is applied only after a significant F-test is obtained using an analysis of variance. An *unprotected* LSD procedure occurs when pairwise comparisons of means are made regardless of whether a main effect or interaction F-test is statistically significant. The technique consists of applying the two-sample Student's t-test (i.e., see Section 3.4) to all $k(k-1)/2$ possible pairs of the k factor–level means.

EXHIBIT 6.9 FISHER'S LSD

Two averages, $\bar{y}_i$ and $\bar{y}_j$, in an analysis of variance are declared to be significantly different if

$$|\bar{y}_i - \bar{y}_j| > \text{LSD}, \tag{6.27}$$

where

$$\text{LSD} = t_{\alpha/2}(\nu)[\text{MS}_E(n_i^{-1} + n_j^{-1})]^{1/2}$$

In this expression n_i and n_j denote the numbers of observations used in the calculation of the respective averages, ν denotes the number of degrees of freedom for error, and $t_{\alpha/2}(\nu)$ denotes a critical value for the t-distribution having ν degrees of freedom and an upper-tail probability of $\alpha/2$.

Fisher's LSD procedure consists of pairwise t-tests, each with a significance level of α. Individual factor–level means or interaction means can be compared using the appropriate response averages for each. Fisher's protected LSD procedure is usually applied with a comparisonwise significance level that equals that for the F-statistic used to make the overall comparison of factor or interaction levels. The unprotected LSD procedure should be used with a significance level that is chosen to control the experimentwise error rate. Bonferroni comparisonwise error rates (see below) are often used: the comparisonwise error rate is α/m, where m is the number of pairwise comparisons to be made.

Fisher's LSD procedure is simple and convenient to use. It can be applied with unequally repeated individual or joint factor levels. It can be used to provide confidence intervals on mean differences; e.g.,

$$(\bar{y}_i - \bar{y}_j) - \text{LSD} \leqslant \mu_i - \mu_j \leqslant (\bar{y}_i - \bar{y}_j) + \text{LSD} \tag{6.28}$$

is a $100(1-\alpha)\%$ confidence interval for $\mu_i - \mu_j$. All of these features make its use attractive.

The major disadvantage of the LSD technique is that its error rate is not satisfactory for testing all possible pairs of mean differences when there are a moderate or large number to be compared. In repeated testing of this type, the experimentwise error rate can be much greater than the desired significance level. The experimentwise error rate can be better controlled by using a Bonferroni procedure (see Exhibit 6.10) to select the comparisonwise significance level. Bonferroni comparisons control the experimentwise error rate by adjusting the t critical value to compensate for many tests on the same set of data. The inequality (6.12) can again be used to make pairwise comparisons of means, but the comparisonwise (two-tailed) significance level is α/m, where α is the desired experimentwise error rate and m is the number of comparisons to be made. Bonferroni pairwise comparisons should not be used when there are a very large number of comparisons to be made, because the comparisonwise significance level can become too small to be of value. In these situations multiple-range tests (see the references at the end of this chapter) offer a compromise between the desired experimentwise error rate and an unacceptably small comparisonwise error rate.

EXHIBIT 6.10 BONFERRONI COMPARISONS

1. Let $\theta = \sum a_i \mu_i$ represent one of m linear combinations of the means μ_i for which one is interested in testing $H_0 : \theta = 0$ vs $H_a : \theta \neq 0$.
2. Reject H_0 if $|\hat{\theta}| = |\sum a_i \bar{y}_i|$ exceeds

$$\text{BSD} = t_{\alpha/2m} \left[\text{MS}_E \sum a_i^2 / n_i \right]^{1/2},$$

where n_i is the number of observations used in the calculation of each $\bar{y}_i$, and $t_{\alpha/2m}$ is an upper-tail t critical value based on ν degrees of freedom, the number of degrees of freedom for MS_E, and an upper-tail probability of $\alpha/2m$.

Bonferroni techniques can be used to test any contrasts or linear combinations of interest using the t-statistic from equation (6.25) or those discussed in Sections 6.2 and 6.3. Confidence intervals can also be constructed using these t-statistics.

Using the wire-die example, we can apply Bonferroni's procedure to the comparisons of the three dies in the tensile-strength study. Using an experimentwise significance level of $\alpha = 0.05$, the two-tailed comparisonwise significance level is $\alpha/m = 0.05/3 = 0.017$. We can choose to be slightly more conservative and use a two-tailed significance level of $\alpha = 0.01$ to obtain a

critical value from Table A3 in the appendix. Doing so produces the following results:

Die:	3	2	1
Average:	82.470	84.492	84.733

In this analysis, each average is based on 18 observations (see Table 2.3), the mean squared error from the ANOVA table (Table 6.4) is MSE = 2.181, and the t critical value is 2.68; consequently, the BSD cutoff value for the mean differences is BSD = 1.32. The line underscoring the two test dies is used to indicate that dies 1 and 2 do not have significantly different averages. The absence of a line underscoring the average for die 3 and for either of the other indicates that dies 1 and 2 are significantly different from the standard die, die 3.

6.4.4 Tukey's Significant Difference Procedure

Tukey's procedure controls the experimentwise error rate for multiple comparisons when all averages are based on the same number of observations (see Exhibit 6.11). The stated experimentwise error rate is very close to the correct value even when the sample sizes are not equal. The technique is similar to Fisher's LSD procedure. It differs in that the critical value used in the TSD formula is the upper $100\alpha\%$ point for the difference between the largest and smallest of k averages. This difference is the range of the k averages, and the critical point is obtained from the distribution of the range statistic, not from the t-distribution.

EXHIBIT 6.11 TUKEY'S TSD

Two averages, $\overline{y}_i$ and $\overline{y}_j$, based on n_i and n_j observations respectively, are significantly different if

$$|\overline{y}_i - \overline{y}_j| > \text{TSD},$$

where

$$\text{TSD} = q(\alpha; k, \nu)\left(\text{MS}_E \frac{n_i^{-1} + n_j^{-1}}{2}\right)^{1/2},$$

in which $q(\alpha; k, \nu)$ is the studentized range statistic, k is the number of averages being compared, MS_E is the mean squared error from an ANOVA fit to the data based on ν degrees of freedom, and α is the experimentwise error rate.

Tukey's TSD is stated in terms of an analysis of variance, but it can be applied in more general situations, as can all of the multiple-comparison procedures discussed in this chapter. All that is needed is a collection of k statistically independent averages and an independent estimate of the common population variance based on ν degrees of freedom. The estimate of σ^2 would replace MS_E in these applications.

Tukey's TSD can provide simultaneous confidence intervals on all pairs of mean differences $\mu_i - \mu_j$. These confidence intervals collectively have a confidence coefficient of $100(1-\alpha)\%$; that is, in the long run, the confidence intervals will simultaneously cover all the mean differences with a probability of $1-\alpha$. The simultaneous confidence intervals are defined as

$$(\bar{y}_i - \bar{y}_j) - \text{TSD} \leqq \mu_i - \mu_j \leqq (\bar{y}_i - \bar{y}_j) + \text{TSD},$$

where TSD is given in Exhibit 6.11.

As an example of a multiple-range test using Tukey's TSD reconsider the wire-tensile-strength study. Using an experimentwise error rate of 0.05, the TSD factor for this example is

$$\text{TSD} = 3.418(2.181/18)^{1/2} = 1.19,$$

where 3.418 is the studentized-range critical value obtained from Table A14 in the appendix (linearly interpolating to approximate the critical point for $\nu = 51$ degrees of freedom). Testing the equality of every pair of means, we arrive at the same conclusion as with the Bonferroni procedure in the last section: there is not enough evidence to conclude that the mean tensile strengths of wire produced from dies 1 and 2 are different, but there is sufficient evidence to conclude that the mean for die 3 is different from those for each of the other two.

6.5 GRAPHICAL COMPARISONS

An indispensable complement to any quantitative assessment of the statistical significance of factor effects is a graphical display of changes in response averages with changes in factor levels. Graphical displays provide a visual description of the effects of factor–level changes and thereby facilitate a better understanding of the population or process being investigated. A wide variety of such displays are available.

The cube plot was introduced in Figure 5.3 (Section 5.2) and was applied to the pilot-plant chemical yield data. Inserting the average responses in the geometric representation of the experimental design space, as illustrated with the cube plot, is a very simple and effective way to summarize how the

mean response changes with changes in the factor levels. Dominant factors are often clearly visible when averages on opposite faces of the cube plot substantively differ from one another. For example, the larger yields on the right face than on the left one of the cube plot in Figure 5.3 is compelling evidence of the importance of temperature on average yield, specifically the main effect for temperature. Often the visual display of a cube plot is used in formal presentations after the quantitative analysis of variance has identified or confirmed the existence of statistically significant factor effects. The cube plot is especially insightful for displaying main effects for experiments with two or three factors, each of which has two or three levels. Main effects for experiments with more factors or more levels per factor are not as easily visualized with a cube plot nor are interaction effects. We will now present three additional graphical displays that are very effective for visualizing a broader variety of factor effects: *trellis plots* using multipanel conditioning, *interaction plots*, and *least significant interval plots*.

An experiment was undertaken to study how the weld strength of a wire-grid welder is affected by electrode position (1, 2A, 2B, 3, 4, 5, 6), electrode polarity (positive, negative), and grid-wire type (coded 1 through 5). The position order corresponds to physical locations on the welding machine. Weld strength (lbs) was measured on a prototype welding machine for two repeat tests at each of the seventy combinations of the factor levels. Note the number of factor levels for electrode positions (7) and for grid wire (5) would cause a cube plot to be very cluttered. Factor effects would be very difficult to discern from so cluttered a plot.

A *trellis plot* displays data on a grid of panels. Each panel contains a graph of a subset of the data for one level of a factor or for one combination of levels of two or more factors. Scatterplots, boxplots, point plots, or many other types of plots can be effective in displaying factor effects in trellis plots. Figure 6.4 shows a trellis scatterplot for the weld-strength study. Each panel is a scatterplot of the average (of the two repeats) weld strength versus the grid-wire types for one combination of position and electrode polarity. The averages in a panel are said to be *conditioned* on the factor level(s) corresponding to the panel. Thus, the averages in each panel are conditioned on combined levels of electrode position and electrode polarity.

This plot shows that there are apparent position differences; for example, compare the averages for position 3 with those of positions 1, 2A, 2B, and 5. There are also possible position-by-polarity interactions; for example, the averages for positions 3 and 5 are approximately the same for the positive polarity but are quite different for the negative polarity. Because of how difficult it is to assess differences without some measure of uncertainty, these observations need to be confirmed with ANOVA calculations and mean comparisons.

Table 6.10 shows the ANOVA table for the weld-strength data. The type of grid wire does not appear to be exerting a substantive effect on mean weld

TABLE 6.10 Analysis-of-Variance Table for the Weld Strength Study

Source	df	SS	MS	F	p-Value
Position	6	6,060,968	1,010,161	18.29	0.000
Pole	1	76,612	76,612	1.39	0.243
Grid Type	4	151,944	37,986	0.69	0.603
Pos × Pole	6	1,875,143	312,524	5.66	0.000
Pos × Type	24	1,282,631	53,443	0.97	0.517
Pole × Type	4	214,297	53,574	0.97	0.430
Pos × Pole × Type	24	1,465,637	61,068	1.11	0.361
Error	70	3,867,015	55,243		
Total	139	14,994,247			

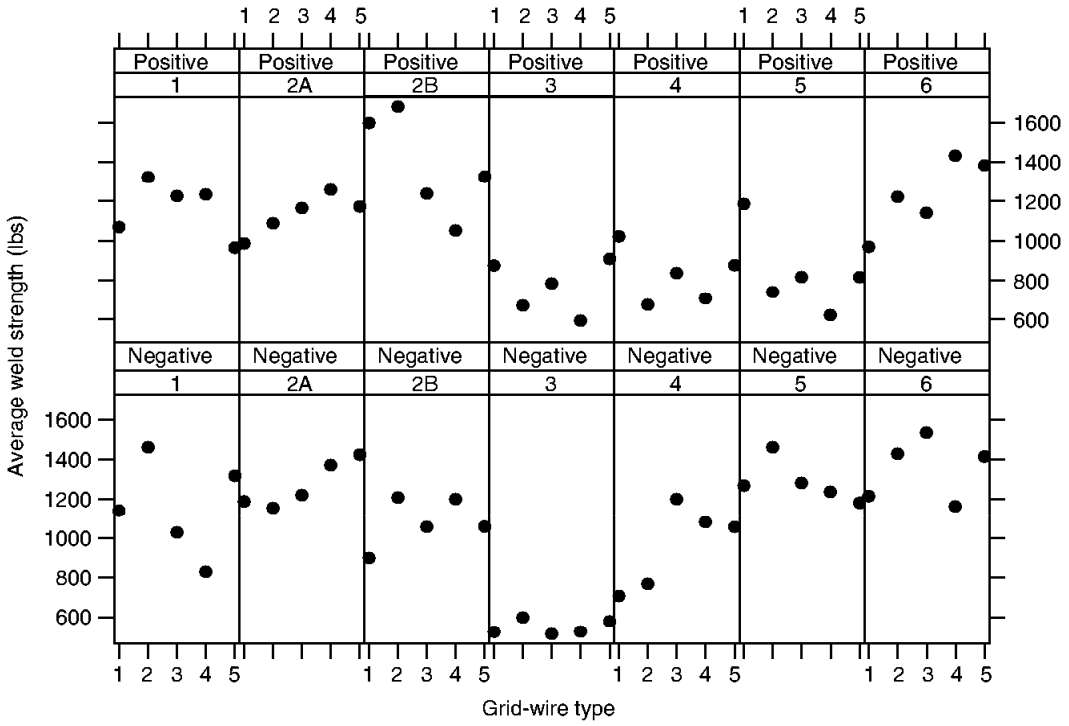


Figure 6.4 Average weld strength, conditioned on electrode position and polarity.

strength because this factor does not occur in any of the statistically significant main effects or interactions. This is consistent with the visual impression left from Figure 6.4 because there does not appear to be any consistent pattern of average weld strength with grid-wire type across the panels. In some of the panels the trend is upward, in some it is downward, in some it is fairly flat.

The two-factor interaction of position-by-polarity (pos × pole) is judged to be statistically significant ($p < 0.001$) in Table 6.10. Because electrode position and electrode polarity combine for a significant joint effect, *there is no need to further test the respective main effects*; that is, the statistically significant interaction reveals that these two factors jointly affect the mean weld strength. This conclusion is also consistent with the impressions left by Figure 6.4 in that there appear to be strong changes in the average weld strength by position, sometimes large changes in the same direction for both polarities, sometimes large changes in opposite directions for the two polarities. A plot that makes these differences clearer is the interaction plot. Interaction plots were first introduced in Section 5.2 and are discussed further in this section for completeness.

Interaction plots are graphs of average responses for combined levels of two or more factors. Levels of one of the factors are plotted on the horizontal axis and dashed lines connect the averages for levels of the other factor(s). The dashed lines are used to clarify changes in the averages with changes in the factor levels. They are not used to suggest continuous trends with factor–level changes unless it is known that the response is a continuous function of (quantitative) factor levels. Interaction plots for two-factor interactions are summarized in Exhibit 6.12. Interaction plots for three or more factors consist of two-factor interaction plots for each combination of the remaining factor levels in the interaction (cf. Figure 5.5).

EXHIBIT 6.12 INTERACTION PLOTS FOR TWO FACTORS

1. Calculate the average response for each combination of the levels of the two factors.
2. Place labels for the levels of one of the two factors on the horizontal axis. Construct a suitable scale for the response averages on the vertical axis.
3. For each level of the second factor, plot the two-factor response averages $\overline{y}_{ij}$ versus the levels of the factor whose labels are on the horizontal axis; that is, for each j plot the averages $\overline{y}_{ij}$ versus the levels i of the factor on the horizontal axis.
4. Connect the plotted averages for each level j of the second factor be a dashed line.

An interaction plot of the average weld strengths for position and polarity is shown in Figure 6.5. Plotting the averages and connecting the averages associated with consecutive positions with dashed lines clearly identifies similarities and differences in the averages for each electrode as position changes. Average weld strength is similar for both electrodes at positions 1, 2A, 4, and 6. There are greater differences at positions 2B, 3, and 5, with the average

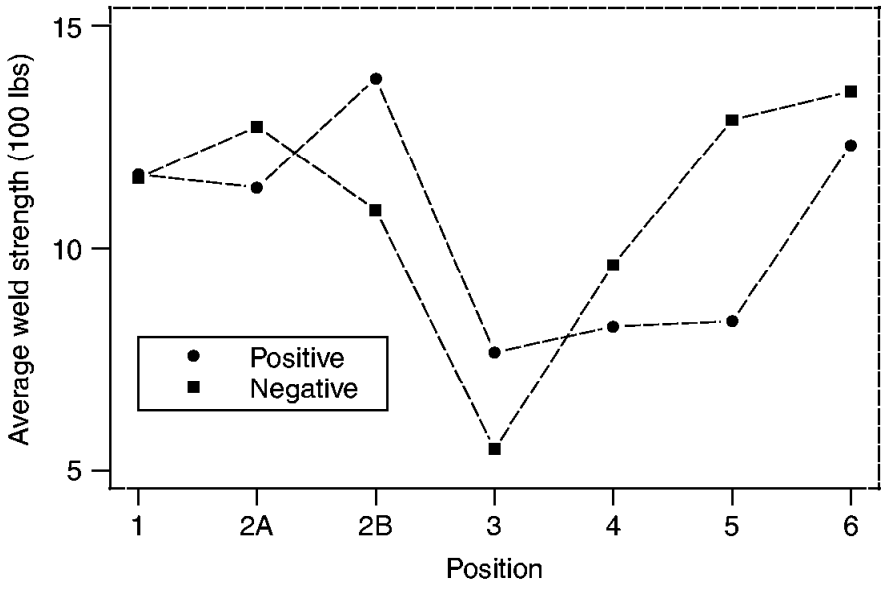


Figure 6.5 Average weld-strength measurements.

TABLE 6.11 Weld Strength Averages (lbs) by Electrode Position and Electrode Polarity

Position	Negative	Positive
1	1,155.8	1,163.5
2A	1,271.9	1,134.5
2B	1,084.6	1,379.3
3	547.5	765.3
4	962.4	823.4
5	1,286.3	835.4
6	1,350.5	1,230.1

weld strength for the positive electrode being greater than that for the negative electrode at positions 2B and 3, and less at position 5.

The final step in the analysis of the position-by-polarity effects on the average weld strength is to determine which of the apparent differences in the averages are significantly different from one another. Table 6.11 lists the average weld-strength measurements for all combinations of the position and polarity factor levels. Each average in the table is calculated from two repeat measurements for each of five grid wire types, a total of ten measurements for each average. Using Fisher's protected LSD (Exhibit 6.9) and a comparison-wise significance level of 0.05, any two averages in this table are significantly

different if their difference (in absolute value) exceeds

$$\text{LSD} = 1.994[55,243(1/10 + 1/10)]^{1/2} = 209.6.$$

Because there are k = (12)(13)/2 = 78 pairwise comparisons among the twelve interaction averages, a Bonferroni cutoff value (Exhibit 6.10) using an experimentwise error rate of $\alpha = 0.05/78 = 0.00064$ is preferable to the protected LSD. Then, any two averages in the table are significantly different if their difference exceeds

$$\text{BSD} = 2.810[55,243(1/10 + 1/10)]^{1/2} = 295.4.$$

Any pairwise comparison of interest can be made using the BSD cutoff criteria.

A simple adaptation of the interaction plot, the *least significant interval* (LSI) plot, permits comparisons of interest to be made graphically. LSI plots are graphs of averages with "error" or "uncertainty" bars extended a length equal to half the LSD, BSD, or TSD cutoff values above and below the individual plotted averages. Any two averages for which the error bars do not overlap are significantly different from each other. LSI plots can be made for main effects or for interactions. Exhibit 6.13 details the construction of LSI plots for main effect averages or interaction averages.

EXHIBIT 6.13 LEAST SIGNIFICANT INTERVAL PLOT

1. Plot each of k averages (vertical axis) versus its level (main effect plot) or its combination of levels (interaction plot).
2. Calculate the LSD (Exhibit 6.9), BSD (Exhibit 6.10), or TSD (Exhibit 6.11) cutoff value.
3. Extend vertical bars above and below each average a distance of LSI = LSD/2, LSI = BSD/2, or LSI = TSD/2, respectively.
4. Any two averages for which the uncertainty limits indicated by the vertical bars do not overlap are declared statistically significant.

Figure 6.6 displays the weld-strength averages from Table 6.11 in an LSI plot using the Bonferroni cutoff value. For this plot LSI = BSD/2 = 147.7. Only the electrode polarity averages for position 5 are judged to be significantly different using this criterion. For comparison purposes, Figure 6.7 is an LSI plot for the same interaction averages using the LSD cutoff. In this plot the electrode polarity averages for positions 2B, 3, and 5 are judged to be statistically significant, with position 5 having the negative polarity average significantly greater than the positive polarity average. The opposite is true for the other two positions. However, with so many comparisons being made

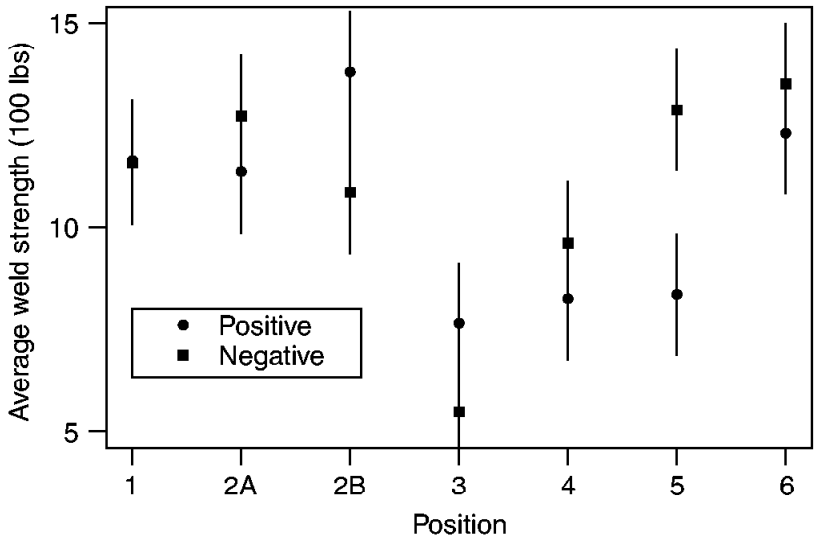


Figure 6.6 Bonferroni LSI comparisons of weld-strength measurements.

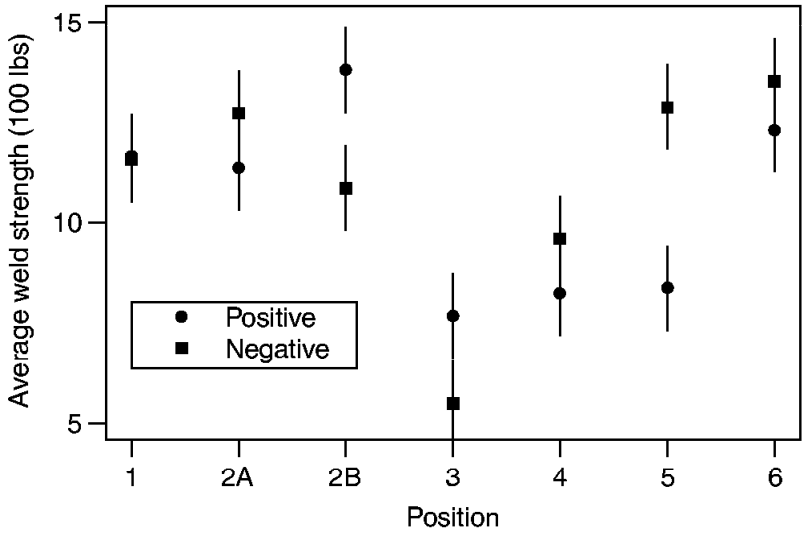


Figure 6.7 Fisher's LSD comparisons of weld-strength measurements.

at a comparisonwise 0.05 significance level, one should be concerned with the high likelihood of making one or more Type 1 errors using the LSD cutoff values.

The procedures discussed in this section provide techniques for making detailed inferences on means beyond the overall conclusions arrived at in an analysis of variance. Two general approaches have been discussed. When a researcher seeks answers to particular questions about the factors in the experiment that can be posed in terms of orthogonal contrasts, t or F statistics can be used to make the comparisons. It is common to simply perform t or

F tests using a comparisonwise significance level. If a number of tests are to be made, an adjusted comparisonwise level (e.g., Bonferroni) should be used. Confidence-interval procedures using comparisonwise confidence levels are equally effective and often more informative.

When comparisons involving nonorthogonal contrasts or when many comparisons are to be made, one of the techniques that offer experimentwise error-rate protection should be used. There are numerous techniques that can be used in this situation. Some of the more useful procedures have been described in Sections 6.4.3 and 6.4.4. Selecting the most useful procedure generally depends on the choice of controlling the experimentwise or the comparisonwise error rate.

If one wishes to control the comparisonwise error rate, the perferred methods discussed in this chapter are t-tests, or Fisher's protected and unprotected LSD procedures. Fisher's protected LSD tests offer more protection from Type I errors than the other two and are less conservative than a procedure based on controlling the experimentwise error rate.

If one desires to control the experimentwise error rate, the useful methods include Bonferroni t-tests and Tukey's procedure. Both of these techniques have strong advocates. Bonferroni tests have the advantage of using an ordinary t-statistic. Their main disadvantage is the small size of the comparisonwise significance level when making a large number of comparisons.

Our experiences lead us to recommend Fisher's protected LSD and Tukey's TSD for multiple pairwise comparisons, depending on whether one wishes to control the comparisonwise or the experimentwise error rates. Each offers sufficient protection against Type I errors, and they are relatively easy to implement. Because they can both be written as t-statistics (they only differ in the choice of a critical value), they use a statistic that is familiar to most users of statistical methodology.

REFERENCES

Text References

Analysis-of-variance assumptions, models, and calculations are covered in a traditional manner in the texts listed at the end of Chapter 4. The following texts are on a more advanced level and emphasize data-analysis methods. The first two describe analysis-of-variance calculations in detail, the second text uses the Statistical Analysis Software (SAS) *library, including designs in which the data are unbalanced.*

Johnson, N. L. and Leone, F. C. (1964). *Statistical and Experimental Design for Engineers and Scientists*, New York: John Wiley & Sons, Inc.

Milliken, G. A. and Johnson, D. E. (1982). *Analysis of Messy Data—Volume I: Designed Experiments*, Belmont, NJ: Wadsworth Publishing Co.

Searle, S. R. (1971). *Linear Models*, New York: John Wiley & Sons, Inc.

Most advanced statistical textbooks contain sections devoted to a discussion of multiple-comparison tests. Three excellent resources that provide comprehensive discussions of the multiple-comparison procedures discussed in this chapter, and other multiple-comparison techniques, are:

Chew, V. (1977). "Comparisons among Treatment Means in an Analysis of Variance," Washington: United States Department of Agriculture, Agricultural Research Services (ARS-H-6).

Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods*. Boca Raton, FL: CRC Press, Inc.

Miller, R. G. (1981). *Simultaneous Statistical Inference*, New York: Springer-Verlag.

Information on trellis plots can be found in:

Cleveland, W. C. (1993). *Visualizing Data*, Summit, New Jersey: Hobart Press.

Further details on the graphical least significant interval procedure can be found in:

Andrews, H. P., Snee, R. D., and Sarner, M. H. (1980). "Graphical Display of Means," *The American Statistician*, **34**, 195–199.

Data References

The example describing the force needed to separate electrical connectors in Section 16.2 is from Chew (1977).

The copper-plate warping example is extracted from the text by Johnson and Leone (p. 98).

The chemical pilot-plant data are taken from

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*, New York: John Wiley and Sons, Inc.

EXERCISES

1 In a nutrition study, three groups of month-old laboratory mice were fed supplements to their usual feed. The percentage weight gains over the next month were recorded and are shown below. Assess whether the supplements differ in their ability to produce weight gain.

Group 1:	7.1 7.0 7.0 8.6 8.2 6.6 6.8 6.7 7.7
Group 2:	7.6 7.3 7.3 8.3 7.6 6.6 6.7 6.8 7.0
Group 3:	9.2 8.3 9.1 9.0 8.9 9.0 9.2 7.6 8.1

2 The following data are elasticity measurements on skin that has been exposed to varying intensities of light for a fixed time period. The light intensities are equally spaced but have been coded for convenience. Evaluate whether there are significant differences in the elasticity measurements for the various light intensities.

Light Intensity (Coded)	Elasticity Measurements						
1	0.54	1.98	0.65	0.52	1.92	1.48	0.97
2	1.76	1.24	1.82	1.47	1.39	1.25	1.29
3	2.05	2.18	1.94	2.50	1.98	2.17	1.83
4	7.92	4.88	9.23	6.51	6.77	4.25	3.72

3 Using the data in Exercise 2, construct a confidence interval on the mean elasticity measurement for the highest light intensity.

4 Treating the light intensities in Exercise 2 as equally spaced factor levels, estimate the linear, quadratic, and cubic effects of light intensity. Which, if any, of these effects are statistically significant? Which, if any, of these effects appears to be the most dominant? Make a plot of the averages to confirm your conclusions visually.

5 A consumer testing agency tested the lifetimes of five brands of dot-matrix computer printer ribbons using a single printer. The results are tabulated below. Assess the performance of these printer ribbons.

Brand	Lifetime (hr)			
A	20.5	17.7	20.0	19.2
B	24.0	26.2	21.2	26.1
C	27.4	35.2	31.2	28.2
D	17.1	18.1	18.5	16.7
E	36.5	33.9	26.9	27.0

6 The ballistic limit velocity for 10-mm rolled homogeneous armor was measured for projectiles with blunt, hemispherical, and conical nose shapes fired at 0° and 45° obliquity to the target. Three test runs were made with each type of projectile, with the firing order of the 18 test runs randomized. Analyze the effects of nose shape and firing angle on the ballistic velocities. Use at least two of the graphical techniques discussed in this chapter to support your conclusions.

Nose Shape	Ballistic Velocities (m/sec)	
	Angle: 0°	45°
Blunt	938	1162
	942	1167
	943	1163
Conical	889	1151
	890	1145
	892	1152
Hemispherical	876	1124
	877	1125
	881	1128

7 Determine the linear, quadratic, and cubic effects of temperature for the copper-plate warping experiment in Table 6.7. Place individual confidence intervals on the mean for each of these polynomial effects. What inferences appear reasonable from these results?

8 Calculate the polynomial interaction effects between copper content and temperature for the copper-plate warping example. Which, if any, of these effects are statistically significant?

9 The table below contains elasticity measurements similar to those reported in Exercise 2. These measurements are taken on four subjects who were specifically chosen to represent four types of skin classes. Monthly measurements were taken on the last day of five consecutive months. Assess whether there are significant effects due to either skin group or time.

Skin Group	Elasticity				
	Month 1	Month 2	Month 3	Month 4	Month 5
A	1.21	1.32	1.44	1.38	1.26
B	0.89	1.11	1.26	1.05	0.82
C	3.67	4.69	4.88	4.33	4.02
D	2.22	2.36	2.58	2.46	2.13

10 Calculate polynomial effects for the main effect for time in the previous exercise. Which polynomial effects, if any, are statistically significant?

11 Refer to Exercise 14, Chapter 3. A second question of interest to process engineers is whether the four processes for manufacturing aluminum tops for food containers are equivalent to one another. These processes are considered equivalent if they possess the same process means and process standard deviations. Examine whether the processes can be considered equivalent, using the data provided below.

Process	**Sphericity Index Measurements**			
A	0.723	0.721	0.707	0.723
B	0.734	0.789	0.796	0.761
C	0.811	0.752	0.902	0.864
D	0.721	0.786	0.742	0.777

12 Data were collected on efficiencies of transfer machines over a five-week period. Based on the data in the following table, are there machine differences? Does there appear to be a time effect? Support your analysis with graphical displays. Write three questions about this study that you would like to ask the engineer who collected the data.

Week	Transfer 21	Transfer 22	Transfer 23
1	46	35	52
	37	40	50
	46	36	43
2	46	63	52
	35	64	62
	47	60	56
3	60	84	58
	47	68	47
	58	78	54
4	50	53	59
	40	59	54
	44	71	68
5	33	46	43
	17	44	31
	24	33	23

13 Construct an effects representation for the main effects of a single four-level factor using two two-level factors (see the appendix to Chapter 5). Show that the three main effects so constructed are orthogonal contrasts. Show that the sum of squares for the main effect of pull angle in the

electric-connector study (Section 6.4.2) equals the sum of the individual sums of squares for these three contrasts.

14 Show that the orthogonal polynomials for a four-level factor can be expressed as linear combinations of the three main-effects constructed in Exercise 13, algebraically confirming that effects representations and orthogonal polynomials are two ways to partition a sum of squares into single-degree-of-freedom components.

15 Five brands of automobile tires are being tested to evaluate their stopping distances (ft) on wet concrete surfaces. Four tires of each brand were mounted on a mid-sized sedan. The vehicle then accelerated to a speed of 60 mph and the brakes were applied. The data below show ten test runs for each tire brand:

Measured Stopping Distances (FT)

	Tire Brand				
Run	***A***	***B***	***C***	***D***	***E***
1	194.1	188.7	185.0	183.0	194.6
2	184.4	203.6	183.2	193.1	196.6
3	189.0	190.2	186.0	183.6	193.6
4	188.8	190.3	182.8	186.3	201.6
5	188.2	189.4	179.5	194.4	200.2
6	186.7	206.5	191.2	198.7	211.3
7	194.7	203.1	188.1	196.1	203.7
8	185.8	193.4	195.7	187.9	205.5
9	182.8	180.7	189.1	193.1	201.6
10	187.8	206.4	193.6	195.9	194.8

Perform a single-factor analysis of variance on these data. Assess the statistical significance of these tire brands using the F-statistic for the main effect of tire brands. What assumptions are necessary to conduct this test?

16 Construct individual 95% confidence intervals for the mean stopping distances for each of the tire brands in Exercise 3. Use the error standard deviation estimate from the analysis of variance in each of the intervals. Construct simultaneous 95% confidence intervals for the means using a Bonferroni procedure. How do the intervals compare?

17 Construct a set of orthogonal contrasts that can be used to compare the mean stopping times in Exercise 3. Calculate the F or t statistic for testing

that each of the contrasts in the means is zero. Interpret the results of these tests in the context of the exercise.

18 Perform Fisher's LSD and Tukey's TSD comparisons on all pairs of the average stopping distances. Interpret the results of these comparisons.

19 Construct an LSI plot for the mean comparisons in Exercise 18. Visually compare the conclusions drawn using Fisher's LSD with the LSI plot.

20 A study was conducted to investigate the effects of three factors on the fragmentation of an explosive device. The factors of interest are pipe diameter, a restraint material, and the air gap between the device and the restraint material. Three repeat tests were conducted on each combination of the factor levels. The data are shown below. Construct an ANOVA table for these data, and perform appropriate tests on the main effects and interactions. Use at least two of the graphical techniques discussed in this chapter to support your conclusions.

		Pipe Diameter			
		0.75 in.		1.75 in.	
	Material:	Sand	Earth	Sand	Earth
Air Gap	0.50 in.	0.0698	0.0659	0.0625	0.0699
		0.0698	0.0651	0.0615	0.0620
		0.0686	0.0676	0.0619	0.0602
	0.75 in.	0.0618	0.0658	0.0589	0.0612
		0.0613	0.0635	0.0601	0.0598
		0.0620	0.0633	0.0621	0.0594

21 Perform suitable multiple-comparison procedures to determine which combinations of the factor levels in Exercise 20 have significantly different average fragment thicknesses.

22 Two companies were contacted by a major automobile manufacturer to design a new type of water pump for their off-the-road vehicles. Each company designed two water-pump prototypes. Ten pumps of each design were tested in four-wheel-drive vehicles, and the number of miles driven before failure of each water pump was recorded. Investigate the following specific comparisons among the response means:

(a) mean failure mileage for the two companies (ignoring design types),

(b) mean failure mileage for the two designs, separately for each company, and

(c) mean failure mileage for the four designs.

Company A		Company B	
Design 1	**Design 2**	**Design 3**	**Design 4**
31,189	24,944	24,356	27,077
31,416	24,712	24,036	26,030
30,643	24,576	24,544	26,573
30,321	25,488	26,233	25,804
30,661	24,403	23,075	25,906
30,756	24,625	25,264	27,190
31,316	24,953	25,667	26,539
30,826	25,930	21,613	27,724
30,924	24,215	21,752	26,384
31,168	24,858	26,135	26,712

23 High-pressure and high-stress tests were used in a project for an aerospace research firm to identify materials suitable for use in syngas coolers. This particular study focused on the effects of temperature on the corrosion of a copper material. Specimens from the copper material were tested for 1000 hours of exposure at 11,000 psi. Eight specimens were tested at each temperature level. The data collected are shown below. Plot the average corrosion percentages versus the temperature levels. Use orthogonal polynomials to assess the polynomial effects of temperature on corrosion.

	Copper Corrosion (%)							
Temp. (°C)	**Specimen 1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
300	5.1	5.2	5.8	4.0	5.5	4.7	5.5	4.3
350	6.5	8.2	2.0	6.1	8.0	4.3	7.2	10.6
400	5.9	4.1	6.4	5.4	3.7	5.5	7.6	7.5
450	11.2	9.8	13.6	12.7	15.1	8.8	13.0	12.7
500	13.8	16.1	18.9	15.7	17.0	15.1	17.3	17.0

CHAPTER 7

Fractional Factorial Experiments

Complete factorial experiments cannot always be conducted because of economic, time, or other constraints. Fractional factorial experiments are widely used in such circumstances. In this chapter fractional factorial experiments for two-level and three-level factors are detailed. The following topics are emphasized:

- *confounding and design resolution, properties of fractional factorial experiments that determine which factor–level combinations are to be included in the experiment,*
- *construction of fractional factorial experiments in completely randomized designs,*
- *efficient fractional factorials for experiments with two-level factors, three-level factors, and a combination of two- and three-level factors,*
- *the use of fractional factorial experiments in saturated and supersaturated screening experiments, and*
- *sequentially building experiments using fractional factorials.*

Fractional factorial experiments are important alternatives to complete factorial experiments when budgetary, time, or experimental constraints preclude the execution of complete factorial experiments. In addition, experiments that involve many factors are routinely conducted as fractional factorials because it is not necessary to test all possible factor–level combinations to estimate the important factor effects, generally the main effects and low-order interactions.

A study in which a fractional factorial experiment was necessitated is summarized in Table 7.1. The experiment was designed to study the corrosion rate of a reactor at a chemical acid plant. There were six factors of interest, each having two levels. Initially it was unknown whether changing any of these six

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

TABLE 7.1 Factor Levels for Acid-Plant Corrosion-Rate Study

Factor	Levels
Raw-material feed rate	3000, 6000 pph
Gas temperature	100°, 220°C
Scrubber water	5, 20%
Reactor-bed acid	20, 30%
Exit temperature	300, 360°C
Reactant distribution point	East, west

process variables (factors) would have an effect on the reactor corrosion rate. To minimize maintenance and downtime, it was desirable to operate this acid plant under process conditions that produce a low corrosion rate. If one were to test all possible combinations of levels of the six process variables, 64 test runs would be required.

A practical difficulty with this investigation was that the plant needed to cease commercial production for the duration of the study. An experimental design that required 64 test runs would have been prohibitively expensive. A small screening experiment was designed and executed in which only gross factor effects were identified. The cost of experimentation was greatly reduced. In addition, substantial production savings resulted from the identification and subsequent adjustment of factors that have a large effect on the corrosion rate.

In this chapter we discuss fractional factorial experiments. To provide a systematic mechanism for constructing such experiments, we begin the chapter by discussing confounding and design resolution in Sections 7.1 and 7.2. In Sections 7.3–7.5, two- and three-level fractional factorial experiments conducted in completely randomized designs are discussed. Section 7.6 outlines the use of fractional factorial experiments in saturated and supersaturated screening experiments and in sequential experimentation.

7.1 CONFOUNDING OF FACTOR EFFECTS

Whenever fractional factorial experiments are conducted, some effects are confounded with one another. The goal in the design of fractional factorial experiments is to ensure that the effects of primary interest are either unconfounded with other effects or, if that is not possible, confounded with effects that are not likely to have appreciable magnitudes. In this section the confounding of effects in fractional factorial experiments is discussed. The efficient design of fractional factorial experiments utilizes the planned confounding of factor effects in their construction.

To make some of the concepts related to confounding and design resolution clear, reconsider the pilot-plant example introduced in Section 5.2. In this study, the yield of a chemical process was investigated as a function of three design factors: operating temperature of the plant (factor A), concentration of the catalyst used in the process (factor B), and type of catalyst (factor C). The graphical (Figure 5.4) and numerical (Table 5.7) analyses of these data suggest that temperature and type of catalyst have a strong joint effect on the chemical yield and that concentration of the catalyst has a strong main effect.

Suppose now that the experimenters wish to design an experiment to confirm these results on a full-scale manufacturing plant. Because of the time required to make changes in the process, suppose that only four test runs can be conducted during each eight-hour operating shift. The ability to run only four of the eight possible factor–level combinations during a single shift introduces the possibility that some of the factor effects could be confounded with the effects of different operators or operating conditions on the shifts.

Confounding was defined in Table 4.1 as a situation in which an effect cannot unambiguously be attributed to a single main effect or interaction. In Section 5.3 and the appendix to Chapter 5, factor effects were defined in terms of differences or *contrasts* (linear combinations whose coefficients sum to zero) of response averages. The expanded definition in Exhibit 7.1 relates the confounding of effects explicitly to their computation. According to this definition, one effect is confounded with another effect if the calculations of the two effects are identical, apart from a possible sign reversal. Effects that are confounded in this way are called *aliases*. The examples presented in this and the next section will make this point clear.

EXHIBIT 7.1

Confounded effects. Two or more experimental effects are confounded (aliased) if their calculated values can only be attributed to their combined influence on the response and not to their unique individual influences. Two or more effects are confounded if the calculation of one effect uses the same (apart from sign) difference or contrast of the response averages as the calculation of the other effects.

When an experiment in designed, it is imperative that an experimenter know which effects are confounded, and with what other effects they are confounded. This knowledge is necessary to ensure that conclusions about the effects of interest will not be compromised because of possible influences of other effects.

In the manufacturing-plant example there are three factors, each at two levels. Table 7.2 lists the effects representation (Section 5.3) for each of the main effects and interactions. Table 7.2 also contains a column of coefficients for the calculation of the *constant* effect, represented by the letter I. The

TABLE 7.2 Effects Representation for Manufacturing-Plant Factorial Experiment*

Factor–Level Combination	Constant (I)	A	B	C	AB	AC	BC	ABC	Response
1	+1	−1	−1	−1	+1	+1	+1	−1	y_{111}
2	+1	−1	−1	+1	+1	−1	−1	+1	y_{112}
3	+1	−1	+1	−1	−1	+1	−1	+1	y_{121}
4	+1	−1	+1	+1	−1	−1	+1	−1	y_{122}
5	+1	+1	−1	−1	−1	−1	+1	+1	y_{211}
6	+1	+1	−1	+1	−1	+1	−1	−1	y_{212}
7	+1	+1	+1	−1	+1	−1	−1	−1	y_{221}
8	+1	+1	+1	+1	+1	+1	+1	+1	y_{222}

*A = Temperature ($-1 = 160°$C, $+1 = 180°$C); B = Concentration ($-1 = 20\%$, $+1 = 40\%$); C = Catalyst ($-1 = C_1$, $+1 = C_2$).

constant effect is simply the overall average response, a measure of location for the response in the absence of any effects due to the design factors.

For simplicity in this discussion, suppose only eight test runs are to be made ($r = 1$) at the manufacturing plant, so that a typical response for the ith level of temperature (factor A), the jth level of concentration (factor B), and the kth level of catalyst (factor C) is denoted y_{ijk}. The last column of Table 7.2 algebraically lists the responses for each factor–level combination.

Note that the signs on the constant effect are all positive. The constant effect is simply the overall response average, and it is obtained by summing the individual factor–level responses (because $r = 1$) and dividing by the total number of factor–level combinations, $2^k = 8$ for $k = 3$ factors. All other effects are obtained by taking the contrasts of the responses indicated in the respective columns and dividing by half the number of factor–level combinations, $2^{k-1} = 4$ (see Section 5.3).

Suppose the response values in Table 7.2 are the responses that would be observed in the absence of any shift effect. Suppose further that the first four factor–level combinations listed in Table 7.2 are taken during the first shift and the last four are taken during the second shift. Observe that the first four combinations have temperature (factor A) at its lower level (160°C) and the second four all have temperature at its upper (180°C) level. The influence of the operating conditions and operators on each shift is simulated in Table 7.3 by adding a constant amount, s_1 for the first shift and s_2 for the second shift, to each observation taken during the respective shifts. The average responses for each shift are thereby increased by the same amounts.

One way to confirm the confounding of the main effect for temperature and the main effect due to the shifts is to observe that these two main effects

TABLE 7.3 Influence of Shifts on Responses from a Factorial Experiment for the Manufacturing-Plant Chemical-Yield Study

Factor–Level Combination	Effects Representation*			Shift	Observed Response
	A	B	C		
1	−1	−1	−1	1	$y_{111} + s_1$
2	−1	−1	1	1	$y_{112} + s_1$
3	−1	1	−1	1	$y_{121} + s_1$
4	−1	1	1	1	$y_{122} + s_1$
				Average	$\bar{y}_{1..} + s_1$
5	1	−1	−1	2	$y_{211} + s_2$
6	1	−1	1	2	$y_{212} + s_2$
7	1	1	−1	2	$y_{221} + s_2$
8	1	1	1	2	$y_{222} + s_2$
				Average	$\bar{y}_{2..} + s_2$

*A = Temperature (−1 = 160°C, 1 = 180°C); B = Concentration (−1 = 20%, 1 = 40%); C = Catalyst (−1 = C_1, 1 = C_2).

have the same effect representation. The main effect for temperature is the difference between the first four and the last four factor–level combinations in Table 7.3. But this is the same difference in averages that one would take to calculate the effect of the shifts. Thus, the main effect representation for factor A in Table 7.3 is, apart from changing all of the signs, also the main effect representation for the shift effect.

A second way of confirming the confounding of the main effect for temperature and the main effect for shift is to symbolically compute the main effect using the effects representation and the responses shown in Table 7.3. Doing so yields the following result:

$$\begin{aligned} M(A) &= (\bar{y}_{2\bullet\bullet} + s_2) - (\bar{y}_{1\bullet\bullet} + s_1) \\ &= \bar{y}_{2\bullet\bullet} - \bar{y}_{1\bullet\bullet} + s_2 - s_1. \end{aligned}$$

While $\bar{y}_{2\bullet\bullet} - \bar{y}_{1\bullet\bullet}$ is the actual effect of temperature, the calculated effect is increased by an amount equal to $s_2 - s_1$, the shift effect.

The other two main effects and all the interaction effects have an equal number of positive and negative signs associated with the responses in each shift. When these effects are calculated, any shift effect will cancel, because half the responses in each shift will add the shift effect and the other half of

the responses will subtract the shift effect. For example,

$$\begin{aligned} M(B) &= \tfrac{1}{4}[(y_{121} + s_1) + (y_{122} + s_1) + (y_{221} + s_2) \\ &\quad + (y_{222} + s_2) - (y_{111} + s_1) - (y_{112} + s_1) \\ &\quad - (y_{211} + s_2) - (y_{212} + s_2)] \\ &= \bar{y}_{\bullet 2\bullet} - \bar{y}_{\bullet 1\bullet}. \end{aligned}$$

In this example, the complete factorial experiment is partitioned so that half of the complete factorial is conducted during each of the two shifts. The fraction chosen for testing during each shift is not the best one that could be selected given the information known from the pilot study. It would be preferable to design the experiment so that one of the interactions that showed little influence on the response in the pilot study would be confounded with the shift effect. In this way, all the effects believed to appreciably influence the response could be estimated even though the responses are split between the two shifts. This type of planned confounding is the goal of the statistical design of fractional factorial experiments.

In general, when designing fractional factorial experiments one seeks to confound either effects known to be negligible relative to the uncontrolled experimental error variation, or in the absence of such knowledge, high-order interactions, usually those involving three or more factors. Confounding of high-order interactions is recommended because frequently these interactions either do not exist or are negligible relative to main effects and low-order interactions. Thus, in the absence of the information provided by the pilot-plant study, it would be preferable to confound the three-factor interaction in the manufacturing-plant experiment rather than the main effect for temperature. To do so one merely needs to assign all factor–level combinations that have one sign (e.g., those with +1: combinations 1, 4, 6, 7) in the three-factor interaction ABC in Table 7.2 to the first shift, and those with the other sign (those with −1: combinations 2, 3, 5, 8) to the second shift.

Not only does confounding occur when a complete factorial experiment is conducted in groups of test runs (e.g., shifts), it also occurs when only a portion of all the possible factor–level combinations are included in the experiment. Confounding occurs because two or more effects representations (apart from a change in all the signs) are the same. A calculated effect then represents the combined influence of the effects. In some instances (e.g., one-factor-at-a-time experiments) the confounding pattern may be so complex that one cannot state with assurance that any of the calculated effects measure the desired factor effects. This again is why planned confounding, confounding in which important effects either are unconfounded or are only confounded with effects that are believed to be negligible, is the basis for the statistical constructions of fractional factorial experiments.

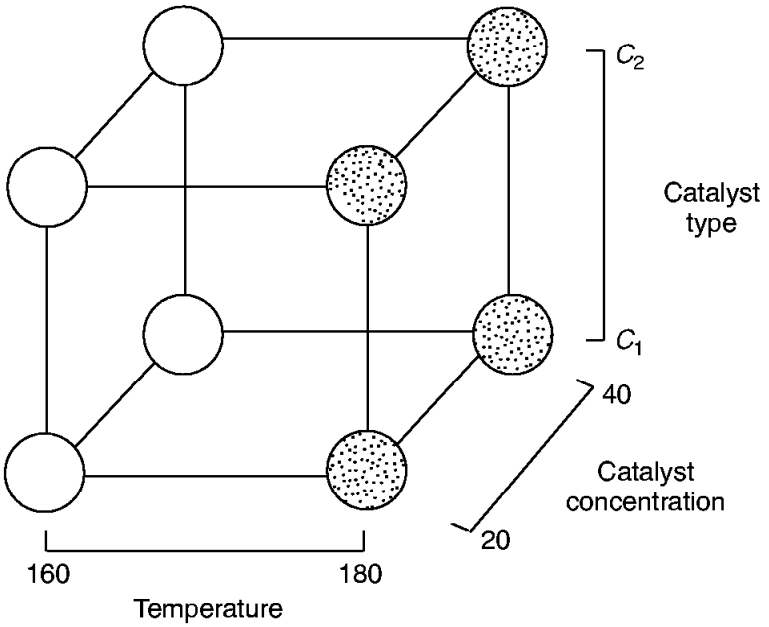


Figure 7.1 Fractional factorial for chemical yield experiment. Only the unshaded combinations are included.

Consider again the manufacturing-plant experiment. Suppose that because of the expense involved, only four test runs could be conducted. While this is a highly undesirable experimental restriction for many reasons, it does occur. Suppose further that only the first four test runs are conducted, as shown graphically in the cube plot in Figure 7.1. Then the effects representations for the various experimental effects are those shown in the first four rows of Table 7.2. These representations are redisplayed in Table 7.4, grouped in pairs between the dashed lines.

The effects representations are grouped in pairs because the effects in each pair are calculated, apart from a change in all the signs, from the same linear combination of the responses. Each effect in a pair is the alias of the other

TABLE 7.4 Effects Representation for a Fractional Factorial Experiment for the Manufacturing-Plant Chemical-Yield Study

Factor–Level Combination	Factor Effects								
	I	A	B	AB	C	AC	BC	ABC	Yield
1	1	−1	−1	1	−1	1	1	−1	y_{111}
2	1	−1	−1	1	1	−1	−1	1	y_{112}
3	1	−1	1	−1	−1	1	−1	1	y_{121}
4	1	−1	1	−1	1	−1	1	−1	y_{122}

effect in the pair. For example, the main effect for B and the interaction AB are, respectively,

$$\begin{aligned} M(B) &= \tfrac{1}{2}(-y_{111} - y_{112} + y_{121} + y_{122}) \\ &= \bar{y}_{\bullet 2\bullet} - \bar{y}_{\bullet 1\bullet} \end{aligned}$$

and

$$\begin{aligned} I(AB) &= \tfrac{1}{2}(+y_{111} + y_{112} - y_{121} - y_{122}) \\ &= -M(B). \end{aligned}$$

The negative sign in the equation $I(AB) = -M(B)$ indicates that the two effects representations are identical except that one has all its signs opposite those of the other.

Similarly, $M(C) = -I(AC)$ and $I(BC) = -I(ABC)$. Calculation of the constant effect, denoted CONST, differs from the calculation of the other effects only in that its divisor is twice the others. Nevertheless, it too is aliased, because CONST $= -M(A)/2$. In terms of the effects representation for these factor effects, one would express this confounding pattern as

$$I = -A, \qquad B = -AB, \qquad C = -AC, \qquad BC = -ABC.$$

Thus, this fractional factorial experiment results in each experimental effect being aliased with one other experimental effect. Moreover, no information is available on the main effect of temperature (factor A), because only one of its levels was included in the design. This is why temperature is aliased with the constant effect.

By constructing a fractional factorial experiment with a different set of four factor–level combinations, a different confounding pattern would result. Every choice of four factor–level combinations will result in each factor effect being aliased with one other effect. One can, however, choose the combinations so that no main effect is aliased with the constant effect. For example, by choosing combinations 2, 3, 5, and 8, the confounding pattern is

$$I = ABC, \qquad A = BC, \qquad B = AC, \qquad C = AB.$$

With this design, which is illustrated in Figure 7.2, all information is lost on the three-factor interaction, because all the factor–level combinations in the design have the same sign on the effects representation for the interaction ABC.

This latter design would not be advisable for the manufacturing-plant example because both the main effect of concentration, $M(B)$, and the joint effect of temperature and catalyst type, $I(AC)$, are known to have appreciable effects on the chemical yield, at least for the pilot plant. The preceding design

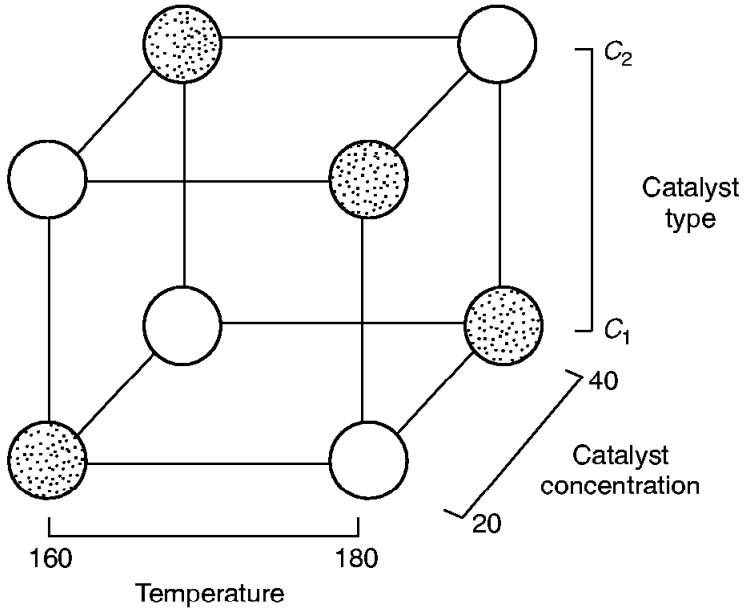


Figure 7.2 Fractional factorial for chemical yield experiment. Only the unshaded combinations are included.

would confound these two effects, because $B = AC$. This design would, however, be suitable if prior experimentation had shown that all interactions were zero or at least negligible relative to the uncontrolled experimental error. If this were the experimental situation, any large calculated main effects could be attributed to individual effects of the respective factors.

Although not commonly done or recommended, an interesting alternative half fraction can be used to confirm the magnitudes of the effects already found for the chemical yield experiment. If the experimenters believed that the pilot plant results would also be true for the manufacturing plant and simply wanted to estimate the magnitudes of the concentration main effect and the temperature by catalyst interaction effect, the half-fraction with defining contrast I = BC (combinations 1, 4, 5, 8; see Table 7.2) could be run. This half fraction would have the main effect for concentration aliased with that for catalyst and the temperature by catalyst interaction aliased with the temperature by concentration interaction. In each of these two alias pairs the second effect was found in the pilot-plant experiment to be nonsignificant. Obviously, the aliasing of two main effects is not to be recommended, but in an extreme situation the knowledge of this confounding pattern could be the only effective experimental strategy.

In the remainder of this chapter the construction of fractional factorial experiments in completely randomized designs will be detailed. The focus in these discussions is on the planned confounding of effects so that the effects of most interest are either unconfounded or only confounded with effects that are believed to be negligible relative to the uncontrolled experimental error. When

prior knowledge about the magnitude of effects is not available, attention is concentrated on the confounding of main effects and low-order interactions only with high-order interactions.

7.2 DESIGN RESOLUTION

An important guide that is used in selecting fractional factorial experiments is the concept of *design resolution.* (see Exhibit 7.2). Design resolution identifies for a specific design the order of confounding of main effects and interactions.

EXHIBIT 7.2

Design Resolution. An experimental design is of resolution R if all effects containing s or fewer factors are unconfounded with any effects containing fewer than $R - s$ factors.

Design resolution is defined in terms of which effects are unconfounded with which other effects. One ordinarily seeks fractional factorial experiments that have main effects ($s = 1$) and low-order interactions (say those involving $s = 2$ or 3 factors) unconfounded with other main effects and low-order interactions — equivalently, in which they are confounded only with high-order interactions.

The resolution of a design is usually denoted by capital Roman letters, for example III, IV, V. The symbol R in the above definition denotes the Arabic equivalent of the Roman numeral, for example $R = 3, 4, 5$.

A design in which main effects are not confounded with other main effects but are confounded with two-factor and possibly higher-order interactions is a resolution-III (also denoted R_{III}) design. Such a design is of resolution III because effects containing $s = 1$ factor are unconfounded with effects containing fewer than $R - s = 3 - 1 = 2$ factors. The only effects containing fewer than $s = 2$ factors are other main effects. As should be clear from the confounding pattern described above, the four-combination (2, 3, 5, 8) fractional factorial experiment with $I = ABC$ for the chemical yield experiment is a R_{III} design.

At least one main effect is confounded with a two-factor interaction in a R_{III} design; otherwise the design would be at least R_{IV}. In a resolution-IV design, main effects are unconfounded with effects involving fewer than $R - s = 4 - 1 = 3$ factors, that is, main effects and two-factor interactions. Two-factor interactions ($s = 2$) are unconfounded with effects involving fewer than $R - s = 2$ factors, that is, main effects ($s = 1$). There are some two-factor interactions confounded with other two-factor interactions in R_{IV} designs.

An interesting feature of fractional factorial experiments is that a design of resolution R is a complete factorial experiment in any R − 1 factors. Thus, a properly chosen half fraction of a three-factor factorial experiment is a complete factorial experiment in any two of the three factors (the defining contrast must involve the three-factor interaction so that the design is R_{III}; see Section 7.3). This is illustrated graphically in Figure 7.3 in which the four points on the cube are projected onto planes representing two of the three factors. Examination of the columns for A, B, and C in Table 7.2 for combinations having +1 values for the ABC interaction also reveals the complete factorial arrangement for each pair of factors.

A design of resolution V has main effects and two-factor interactions ($s = 1, 2$) unconfounded with one another. Some main effects are confounded with four-factor and higher-order interactions; some two-factor interactions are confounded with three-factor and higher-order interactions. In Resolution-VII designs, main effects, two-factor interactions, and three-factor interactions are mutually unconfounded with one another.

This discussion of design resolution is intended to provide a quick means of assessing whether an experimental design allows for the estimation of all important effects (assumed to be main effects and low-order interactions) without the confounding of these effects with one another. Techniques for determining the resolution of a design are presented in the next section.

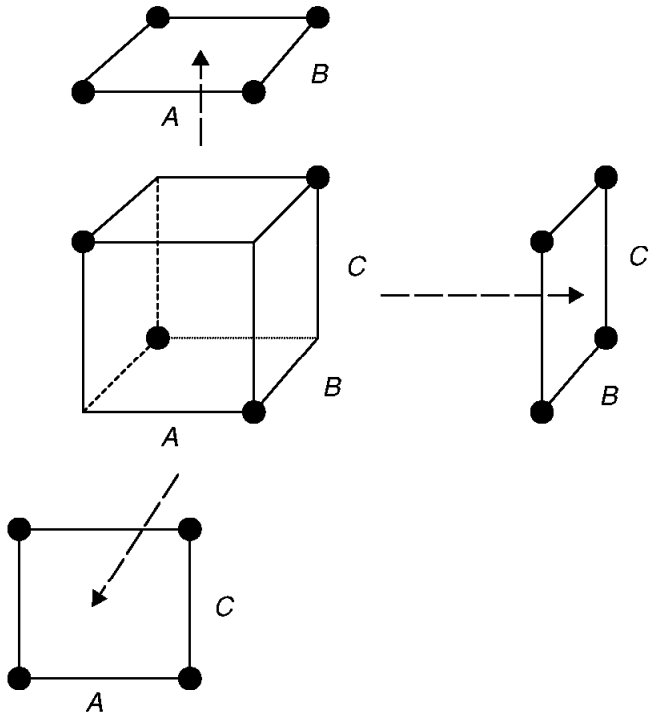


Figure 7.3 Projections of a half fraction of a three-factor complete factorial experiment ($I = ABC$).

7.3 TWO-LEVEL FRACTIONAL FACTORIAL EXPERIMENTS

The number of test runs needed for complete factorial experiments increases quickly as the number of factors increases, even if each factor is tested at only two levels. The intentional confounding of factor effects using fractional factorials can reduce the number of test runs substantially for large designs if some interaction effects are known to be nonexistent or negligible relative to the uncontrolled experimental error variation.

The next two subsections of this section detail the construction of two-level fractional factorial experiments in completely randomized designs. The basic procedures are introduced in Section 7.3.1 for half-fractions. These procedures are extended to smaller fractions ($\frac{1}{4}$, $\frac{1}{8}$, etc.) in Section 7.3.2. A common terminology used to designate a $(1/2)^p$ Fraction of a complete 2^k factorial experiment is to refer to the resulting factor–level combinations as a 2^{k-p} fractional factorial experiment.

7.3.1 Half Fractions

Half fractions of factorial experiments are specified by first stating the *defining contrast* for the fraction. The defining contrast is the effect that is aliased with the constant effect. It can be expressed symbolically as an equation, the *defining equation*, by setting the aliased effect equal to I, which represents the constant effect. For the manufacturing-plant example in the last section, two half fractions of the complete 2^3 experiment were discussed. The defining equations for the two fractional factorial experiments were, respectively, $I = -A$ and $I = ABC$.

Ordinarily the effect chosen for the defining contrast is the highest-order interaction among the factors. Once the defining contrast is chosen, the half fraction of the factor–level combinations that are to be included in the experiment are those that have either the positive or the negative signs in the effects representation of the defining contrast (see Exhibit 7.3). The choice of the positive or the negative signs is usually made randomly (e.g., by flipping a coin).

As an illustration of this procedure, consider the acid-plant corrosion-rate study discussed in the introduction to this chapter. Suppose that a half fraction of the 2^6 complete factorial experiment is economically feasible. If no prior information is available about the existence of interaction effects, the defining contrast is the six-factor interaction, *ABCDEF*. The effects representation for this defining contrast is determined by multiplying the respective elements from the effects representations of the six main effects. The sign (+ or −) of the elements in the effects representation is randomly chosen and the corresponding 32 factor–level combinations to be included in the experiment are thereby identified.

EXHIBIT 7.3 DESIGNING HALF FRACTIONS OF TWO-LEVEL FACTORIAL EXPERIMENTS IN COMPLETELY RANDOMIZED DESIGNS

1. Choose the defining contrast: the effect that is to be aliased with the constant effect.
2. Randomly decide whether the experiment will contain the factor–level combinations with the positive or the negative signs in the effects representation of the defining contrast.
3. Form a table containing the effects representations of the main effects for each of the factors. Add a column containing the effects representation of the defining contrast.
4. Select the factor–level combinations that have the chosen sign in the effects representation of the defining contrast. Randomly select any repeat tests that are to be included in the design.
5. Randomize the assignment of the factor–level combinations to the experimental units or to the test sequence, as appropriate.

Table 7.5 lists the 32 combinations for the half fraction of the acid-plant corrosion-rate experiment using the defining equation $I = -ABCDEF$. The half fraction is graphically displayed in Figure 7.4. If a few repeat tests can be included in the experiment, they will be randomly selected from those shown in Table 7.5. The test sequence of these factor–level combinations should be randomized prior to the execution of the experiment.

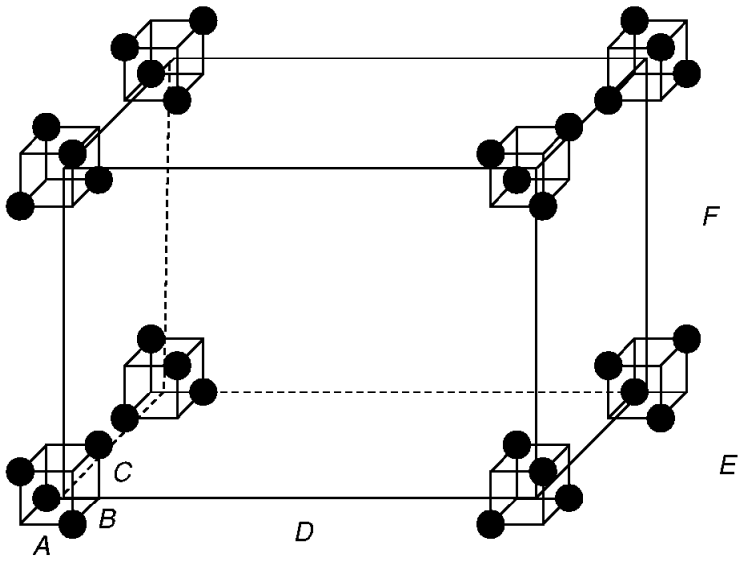


Figure 7.4 Half-fraction (R_{VI}) of a 2^6 experiment: $I = -ABCDEF$.

TABLE 7.5 Half Fraction of the Complete Factorial Experiment for the Acid-Plant Corrosion-Rate Study

	Effects Representation						
Factor–Level Combination	*A*	*B*	*C*	*D*	*E*	*F*	*ABCDEF*
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
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
17	1	−1	−1	−1	−1	−1	−1
18	1	−1	−1	−1	1	1	−1
19	1	−1	−1	1	−1	1	−1
20	1	−1	−1	1	1	−1	−1
21	1	−1	1	−1	−1	1	−1
22	1	−1	1	−1	1	−1	−1
23	1	−1	1	1	−1	−1	−1
24	1	−1	1	1	1	1	−1
25	1	1	−1	−1	−1	1	−1
26	1	1	−1	−1	1	−1	−1
27	1	1	−1	1	−1	−1	−1
28	1	1	−1	1	1	1	−1
29	1	1	1	−1	−1	−1	−1
30	1	1	1	−1	1	1	−1
31	1	1	1	1	−1	1	−1
32	1	1	1	1	1	−1	−1

	Levels	
Factor	−1	+1
A = Raw-material feed rate (pph)	3000	6000
B = Gas temperature (°C)	100	220
C = Scrubber water (%)	5	20
D = Reactor-bed acid (%)	20	30
E = Exit temperature (°C)	300	360
F = Reactant distribution point	East	West

The resolution of a half fraction of a complete factorial experiment equals the number of factors included in the defining contrast. Half fractions of highest resolution, therefore, are those that confound the highest-order interaction with the constant effect. The highest resolution a half fraction can attain is equal to the number of factors in the experiment.

The confounding pattern of a half fraction of a complete factorial is determined by symbolically multiplying each side of the defining equation by each of the factor effects. The procedure given in Exhibit 7.4 is used to determine the confounding pattern of effects.

EXHIBIT 7.4 CONFOUNDING PATTERN FOR EFFECTS IN HALF FRACTIONS OF TWO-LEVEL COMPLETE FACTORIAL EXPERIMENTS

1. Write the defining equation of the fractional factorial experiment.
2. Symbolically multiply both sides of the defining equation by one of the factor effects.
3. Reduce both sides of the equation using the following algebraic convention: For any factor, say X,

$$X \cdot I = X \quad \text{and} \quad X \cdot X = X^2 = I.$$

4. Repeat steps 2 and 3 until each factor effect is listed in either the left or the right side of one of the equations.

The symbolic multiplication of two effects is equivalent to multiplying the individual elements in the effects representations of the two effects. Note that the constant effect has all its elements equal to 1. Consequently, any effect multiplying the constant effect remains unchanged. This is the reason for the convention $X \cdot I = X$. Similarly, any effect multiplying itself will result in a column of ones because each element is the square of either $+1$ or -1. Consequently, an effect multiplying itself is a column of $+1$ values: $X \cdot X = I$.

As an illustration of this procedure for determining the confounding pattern of effects, recall the first of the two half fractions of the manufacturing-plant example discussed in the last section. Because the first four factor–level combinations were used, those with the negative signs on the main effect for A, the defining equation is $I = -A$. Multiplying this equation by each of the other effects, using the algebraic convention in step 3 of Exhibit 7.4, results in the following confounding pattern for the factor effects:

$$I = -A, \quad B = -AB, \quad C = -AC, \quad BC = -ABC.$$

This is the same confounding pattern that was found in the previous section by calculating the individual effects. One can use this same procedure to confirm

the confounding pattern reported in the last section for the half fraction with defining equation $I = ABC$.

The latter half fraction for the manufacturing-plant example has three factors in the defining contrast. Because of this, the completely randomized design is of resolution III. Main effects are unconfounded with other main effects, but main effects are confounded with two factor interactions. The confounding pattern makes this explicit; however, the confounding pattern did not have to be derived to make this determination. Simply knowing the defining contrast allows one to ascertain the resolution of the design and therefore the general pattern of confounding.

As a second illustration, consider again the acid-plant corrosion-rate study. The highest-resolution half fraction has as its defining contrast *ABCDEF*, the six-factor interaction. This is a resolution-VI design, because six factors occur in the defining contrast. Because the design is R_{VI}, all main effects and two-factor interactions are unconfounded with one another, but some three-factor interactions are confounded with other three-factor interactions.

Suppose one chooses the negative signs in the six-factor interaction to select the factor–level combinations, as in Table 7.5. Then the defining equation is $I = -ABCDEF$. If one wishes to write out the explicit confounding pattern of the effects, the defining equation will be multiplied by all of the effects until each effect appears on one or the other side of one of the equations. For example, the confounding pattern of the main effects is as follows:

$$A = -AABCDEF = -BCDEF,$$
$$B = -ABBCDEF = -ACDEF,$$
$$C = -ABCCDEF = -ABDEF,$$
$$D = -ABCDDEF = -ABCEF.$$
$$E = -ABCDEEF = -ABCDF,$$
$$F = -ABCDEFF = -ABCDE.$$

Note that one does not now have to evaluate the confounding pattern of the five-factor interactions, because each of them already appears in one of the above equations. Confounding patterns for two-factor and higher-order interactions are determined similarly. By determining the confounding pattern of all the factor effects, one can confirm that the design is R_{VI}.

7.3.2 Quarter and Smaller Fractions

Quarter and smaller fractions of two-level complete factorial experiments are constructed similarly to half fractions. The major distinction is that more than one defining contrast is needed to partition the factor–level combinations.

Consider designing an experiment that is to include a quarter fraction of a six-factor complete factorial experiment. For example, suppose only 16 of the 64 test runs can be conducted for the acid-plant corrosion-rate study. By using two defining contrasts, $\frac{1}{4}$ of the factor–level combinations can be selected. Suppose the two defining contrasts chosen are $ABCDEF$ and ABC. Suppose further that the negative sign is randomly assigned to the first defining contrast and a positive sign is randomly selected for the second defining contrast. This means that a factor–level combination is only included in the experiment if it has both a negative sign on the effects representation for the six-factor interaction $ABCDEF$ and a positive sign on the effects representation for the three-factor interaction ABC. Table 7.6 lists the 16 factor–level combinations that satisfy both requirements. Figure 7.5 graphically displays the 16 combinations.

Although only two defining equations need to be chosen to select the factor–level combinations for a quarter fraction of a complete factorial experiment, a third equation is also satisfied. Note in the last column of Table 7.6 that all the factor–level combinations that satisfy the first two defining equations, $I = -ABCDEF$ and $I = ABC$, also satisfy $I = -DEF$. A third defining equation is always satisfied when two defining contrasts are chosen. This third implicit defining contrast can be identified by symbolically multiplying the other two contrasts. In this example,

$$(-ABCDEF)(ABC) = -AABBCCDEF = -DEF.$$

A concise way of expressing the defining contrasts is $I = -ABCDEF = ABC(= -DEF)$, with the implied contrast in parentheses being optional, because it can be easily be determined from the other two.

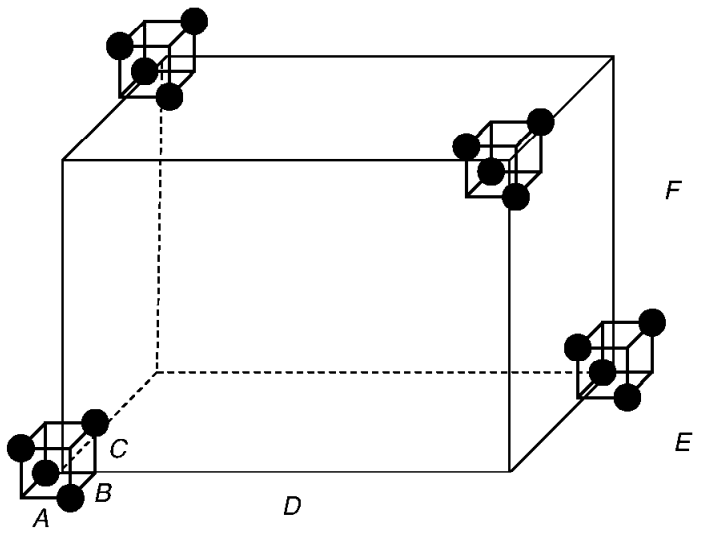


Figure 7.5 Quarter-fraction (R_{III}) of a 2^6 experiment: $I = -ABCDEF = ABC(= -DEF)$.

TABLE 7.6 Quarter Fraction of the Complete Factorial Experiment for the Acid-Plant Corrosion-Rate Study

Factor–Level Combination	Effects Representation								
	A	*B*	*C*	*D*	*E*	*F*	*ABCDEF*	*ABC*	*DEF*
1	−1	−1	1	−1	−1	−1	−1	1	−1
2	−1	−1	1	−1	1	1	−1	1	−1
3	−1	−1	1	1	−1	1	−1	1	−1
4	−1	−1	1	1	1	−1	−1	1	−1
5	−1	1	−1	−1	−1	−1	−1	1	−1
6	−1	1	−1	−1	1	1	−1	1	−1
7	−1	1	−1	1	−1	1	−1	1	−1
8	−1	1	−1	1	1	−1	−1	1	−1
9	1	−1	−1	−1	−1	−1	−1	1	−1
10	1	−1	−1	−1	1	1	−1	1	−1
11	1	−1	−1	1	−1	1	−1	1	−1
12	1	−1	−1	1	1	−1	−1	1	−1
13	1	1	1	−1	−1	−1	−1	1	−1
14	1	1	1	−1	1	1	−1	1	−1
15	1	1	1	1	−1	1	−1	1	−1
16	1	1	1	1	1	−1	−1	1	−1

Factor	Levels	
	−1	+1
A = Raw-material feed rate (pph)	3000	6000
B = Gas temperature (°C)	100	220
C = Scrubber water (%)	5	20
D = Reactor-bed acid (%)	20	30
E = Exit temperature (°C)	300	360
F = Reactant distribution point	East	West

The resolution of quarter fractions of complete factorials constructed in this fashion equals the number of factors in the smallest of the defining contrasts, including the one implied by the two chosen for the design. Thus, in the above example, the resulting quarter fraction is a resolution-III design because the smallest number of factors in the defining contrasts is three (*ABC* or *DEF*). A higher-resolution fractional factorial design can be constructed by using, apart from sign, the defining equations $I = ABCD = CDEF(= ABEF)$. This would be a R_{IV} design, since the smallest number of factors in the defining contrast is four.

Just as there are three defining contrasts in a quarter fraction of a complete factorial experiment, each of the factor effects in the experiment is aliased with three additional effects. The confounding pattern is determined by symbolically multiplying the defining equations, including the implied one, by each of the factor effects. For example, if the defining equations for a quarter fraction of a six-factor factorial are

$$I = ABCD = CDEF = ABEF,$$

then the main effects are confounded as follows:

$$\begin{aligned}
A &= BCD = ACDEF = BEF,\\
B &= ACD = BCDEF = AEF,\\
C &= ABD = DEF = ABCEF,\\
D &= ABC = CEF = ABDEF,\\
E &= ABCDE = CDF = ABF,\\
F &= ABCDF = CDE = ABE.
\end{aligned}$$

As required by the fact that this is a resolution-IV design, the main effects are not confounded with one another. One can show in the same manner that the two-factor interactions are confounded with one another.

Another way to appreciate the aliasing of effects in fractional factorials, apart from the equation form given above, is to write each set of aliased effects as a sum of effects. Using this form of expression, the contrasts for the main effects are actually estimating the combined effects:

$$\begin{aligned}
&A + BCD + ACDEF + BEF\\
&B + ACD + BCDEF + AEF\\
&C + ABD + DEF + ABCEF\\
&D + ABC + CEF + ABDEF\\
&E + ABCDE + CDF + ABF\\
&F + ABCDF + CDE + ABE.
\end{aligned}$$

The general procedure for constructing fractional factorial experiments for two-level factors is a straightforward generalization of the procedure for quarter fractions. A $(\frac{1}{2})^p$ fraction requires that p defining contrasts be chosen, none of which is obtainable by algebraic multiplication of the other contrasts. An additional $2^p - p - 1$ implicit defining contrasts are then determined by algebraically multiplying the p chosen contrasts. The resolution of the resulting completely randomized fractional factorial experiment equals the number of factors in the smallest defining contrast, including the implicit ones. The

confounding pattern is obtained by multiplying the defining equations by the factor effects.

Table 7A.1 in the appendix to this chapter is a list of defining equations, apart from sign, for fractional factorial experiments involving from 5 to 11 factors. Included in the table are the defining equations and the resolutions of the designs. Resolution-IV and resolution-V designs are sought because the primary interest in multifactor experiments is generally in the analysis of main effects and two-factor interactions. The last column of the table is an adaptation of the defining equations to the construction of fractional factorial experiments using added factors. This type of construction is explained in the appendix to this chapter.

7.4 THREE-LEVEL FRACTIONAL FACTORIAL EXPERIMENTS

Complete factorial experiments with factors having three or more levels can be burdensome and inefficient in terms of the number of test runs. For example, a four-factor experiment in which each factor has three levels requires a minimum of $3^4 = 81$ test runs, while a six-factor experiment requires a minimum of $3^6 = 729$ test runs. Repeat test runs in either of these experiments adds to the total number of test runs. Moreover, a very large number of test runs in these experiments are used to estimate high-order interactions, which are not of primary interest in assessing the factor effects. For example, there are 48 degrees of freedom for three- and four-factor interactions in a 3^4 complete factorial. In essence, 48 of the 81 test runs are used to estimate these high-order interactions. A total of 656 of the 729 test runs in a 3^6 complete factorial are used to estimate third-order and higher interactions. This is the reason for the inefficiency of these complete factorials.

As with two-level factorial experiments, intentional confounding of factor effects in these large experiments using fractional factorials can greatly reduce the number of test runs without sacrificing information on the main effects and low-order interactions. However, fractional factorials for experiments with factors having 3, 5, 6, 7, etc. levels are exceedingly more difficult to construct than if all factors have either 2 levels or numbers of levels that are powers of 2 (see below). They are also much less efficient in terms of reducing the number of test runs. In this section, these issues are highlighted through the construction of fractional factorials of 3^k experiments.

The basic procedure for constructing a one-third fraction of a 3^k experiment is presented in Section 7.4.1. To facilitate the construction of fractional factorials of 3^k experiments, tables of *orthogonal arrays* are discussed in Section 7.4.2. A number of the most useful fractional factorials for experiments with some or all factors having three levels can be obtained directly from construction of such tables. Fractional factorial for experiments in which some factors have 2 levels and some have 3 levels are discussed in section 7.5.

Several alternatives to the use of 3^k complete and fractional factorials are discussed in other portions of this book. In the Appendix to this chapter, fractional factorial experiments for factors having numbers of levels equal to a power of 2 are constructed similarly to fractional factorials of 2^k experiments. These fractional factorials have many of the benefits and efficiencies of fractional factorials of 2^k experiments. In Chapter 17, two-level complete and fractional factorials with quantitative factor levels are augmented with test runs at one or more specified additional levels. These experiments, which include central composite designs and Box–Behnken designs, are often more efficient than fractions of 3^k designs.

7.4.1 One-Third Fractions

As discussed in Section 6.1.3, a main effect for a q-level factor has $q - 1$ degrees of freedom. An implication of this fact is that each main effect can be written as a multiple of the sum of $q - 1$ squares of statistically independent contrasts of the response averages (averaged across repeats and other factor levels):

$$SS_A = \text{const} \times \sum_{i=1}^{q-1} (c_i' \bar{y}_{i\cdot})^2,$$

where the c_i are vectors of contrasts. For a three-level factor in a balanced complete factorial, there are many ways of writing the two contrasts that comprise the main effect. For example, the main effect sum of squares can be written as a function of the two statistically independent orthogonal contrasts $\bar{y}_1 - \bar{y}_3$ and $\bar{y}_1 - 2\bar{y}_2 + \bar{y}_3$, or as a linear combination of any other two orthogonal contrasts of the three factor–level averages. Note that the contrast vectors $c_1 = (1 \quad 0 \quad -1)'$ and $c_2 = (1 \quad -2 \quad 1)'$ that are used to calculate the contrasts in the averages are orthogonal: $c_1'c_2 = 0$. Because of this, statistical theory can be used to show that these contrasts of the averages are statistically independent. This is the essence of the importance of degrees of freedom. The degrees of freedom indicate the number of statistically independent contrasts needed to quantitatively determine all the possible effects of a factor. In addition, degrees of freedom indicate the number of statistically independent contrasts needed to calculate the sum of squares that is used to assess the statistical significance of the factor effects. For more discussion of factor effects, see the Appendix to Chapter 5.

To fractionate a balanced three-level factorial experiment, defining equations must account for the two degrees of freedom in each main effect. Since the main effects can always be represented by two orthogonal (and statistically independent, if the design is balanced) contrasts, all interactions involving three-level factors can be formed as products of these main effect

contrasts. It is the interaction contrasts that are ordinarily used to fractionate the experiment.

By convention, the two orthogonal contrasts for a main effect A are designated A and A^2. This is merely a notational convention — it does not imply that the contrast A^2 is the square of the contrast A. In the above example, this convention would imply that $A = (1 \quad 0 \quad -1)'$ and $A^2 = (1 \quad -2 \quad 1)'$. Similarly, the interaction between factors A and B is designated by two terms, AB and AB^2, each with *two* degrees of freedom (the interaction between the two factors has a total of 4 degrees of freedom). A three-way interaction between factors A, B, and C will have four components: $ABC, ABC^2, AB^2C, AB^2C^2$. Each of these components represents an interaction between a main-effect component and a two-way interaction component, and each will have two degrees of freedom. This pattern easily extends to higher-order interactions.

As with fractional factorials of two-level factors, the defining contrast for the three-level fractional factorial is the contrast that is aliased with the constant term. In constructing a one-third fraction, a choice must be made as to which of three fractions of the complete factorial is chosen. This is accomplished by following the procedure described in Exhibit 7.5.

As an illustration, suppose a one-third fraction of a 3^3 experiment is desired. Further, suppose the defining equation is chosen to be $I = AB^2C^2$. In the appendix to this chapter modular arithmetic is used as an alternative to the general method of Section 7.3 to fractionate two-level complete factorials. That method basically replaces the products of factor effects with a sum of effects and selects fractions based on whether the sum is 0 or 1 mod (2). This alternative is very advantageous for selecting fractions of three-level factors, using sums of factor effects modulo 3.

EXHIBIT 7.5 DESIGNING ONE-THIRD FRACTIONS OF THREE-LEVEL FACTORIAL EXPERIMENTS IN COMPLETELY RANDOMIZED DESIGNS

1. Let x_j denote a coded level for factor j, where the levels are coded as 0, 1, or 2.
2. Write the defining contrast as $T = a_1x_1 + a_2x_2 + \cdots + a_kx_k$, where $a_j = 0$, 1, or 2 is the power on factor j in the algebraic expression for the defining contrast.
3. Randomly decide whether $T = 0$ mod (3), $T = 1$ mod (3), or $T = 2$ mod (3) in the defining relation. Note that $T = r$ mod (3) only if $(T - r)/3$ is an integer.
4. Form a table listing the factor–level combinations in coded form. Add a column for the values of T in the defining contrast.
5. Select the test runs that have the chosen value of T. Randomly add any repeat tests that are to be included in the design.
6. Randomize the assignment of the factor–level combinations to the experiment units or to the test sequence, as appropriate.

To implement this method, the first step in Exhibit 7.5 is to designate the levels of each factor as 0, 1, or 2. Because the algebraic expression for the defining contrast is $I = AB^2C^2$, the second step is to write the defining contrast in equation form as

$$T = x_1 + 2x_2 + 2x_3.$$

If one randomly selects the fraction for which $T = 0$ mod (3), then the factor–level combinations for this one-third fraction are listed in Table 7.7. This one-third fraction is displayed graphically in Figure 7.6. Note the geometrically even distribution of these nine test runs on edges, corners, and faces.

TABLE 7.7 One-Third Fraction of a 3^3 Factorial Experiment

Combination	A	B	C	$T = x_1 + 2x_2 + 2x_3$
1	0	0	0	0
2	0	1	2	0
3	0	2	1	0
4	1	0	1	0
5	1	1	0	0
6	1	2	2	0
7	2	0	2	0
8	2	1	1	0
9	2	2	0	0

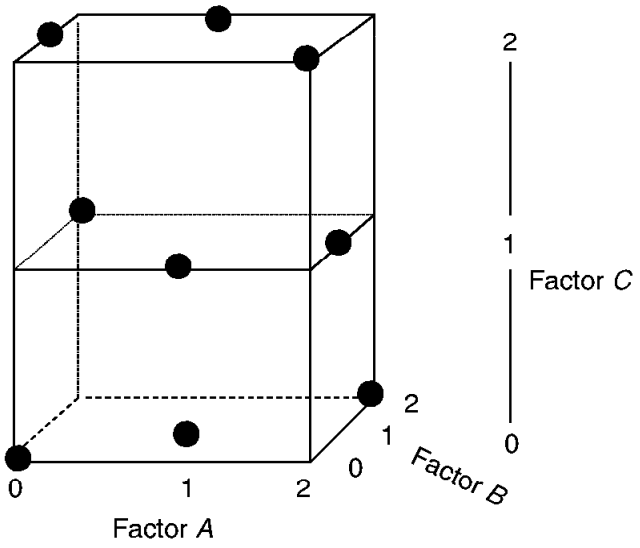


Figure 7.6 One-third fractional factorial experiment.

There are two other one-third fractions that could have been selected by using either of the defining contrasts $T = x_1 + 2x_2 + 2x_3 = 1 \mod (3)$ or $T = x_1 + 2x_2 + 2x_3 = 2 \mod (3)$. Similarly, one could have chosen any of the other three components of the ABC interaction, such as ABC, ABC^2, or AB^2C, to form the fractional factorial. Because there would be three defining relations for each component, this yields a total of 12 different one-third fractions that are associated with the three-way interaction for three three-level factors. The design in Table 7.7 is of Resolution-III because this is the number of factors included in the defining contrast. The confounding pattern is determined as it is in two-level fractional factorials, by multiplying each side of the defining equation by each of the factor effects. The procedure to follow is given in Exhibit 7.6. For example, the confounding pattern of the main effects for factor A for this example is:

$$A = A(AB^2C^2) = A^2B^2C^2 = (A^2B^2C^2)^2 = A^4B^4C^4 = ABC$$
$$A^2 = A^2(AB^2C^2) = A^3B^2C^2 = B^2C^2 = (B^2C^2)^2 = B^4C^4 = BC.$$

EXHIBIT 7.6 CONFOUNDING PATTERN FOR EFFECTS IN ONE-THIRD FRACTIONS OF THREE-LEVEL COMPLETE FACTORIAL EXPERIMENTS

1. Write the defining equation of the fractional factorial experiment.
2. Symbolically multiply both sides of the defining equation by one of the factor effects components.
3. Reduce the right-side of the equation using the following algebraic conventions. For any factor X,

$$X \cdot I = X \quad X \cdot X = X^2 \quad X \cdot X \cdot X = X^3 = I.$$

4. By convention, force the exponent on the first letter of the multiplied effect to equal one and reduce the remaining exponents of the effect to modulo 3 units. If the exponent on the first letter does not equal one, square the entire term and reduce the exponents modulo 3 to obtain an exponent of one.
5. Repeat steps 2–4 until each factor effect is listed in either the left or right side of one of the equations.

Thus, the two degrees of freedom for the main effect of factor A are aliased with one of the degrees of freedom of the BC and ABC interactions. This is typical of what happens with fractions of three-level factorial experiments: the confounding pattern is far more complicated than that of fractions of two-level factorial experiments. As with two-level fractional factorials, it is only when there are several factors that fractional factorials of three-level complete

factorials have main effects unaliased with one another and with two-factor interactions.

7.4.2 Orthogonal Array Tables

The complexity involved in fractionating 3^k experiments has led many researchers to develop tables containing the test runs for these designs. Having such tables removes the need to construct the design and to identify the confounding pattern, but one must still be aware of the design resolution and, in some cases, of the confounding patterns. In an *orthogonal array* table, balance is achieved because each level of a factor occurs an equal number of times with each level of each of the other factors. Note that all complete factorials in which there are an equal number of repeats for each factor–level combination are orthogonal arrays. Some fractional factorials are orthogonal arrays, some are not.

Table 7.8 displays an example of both an orthogonal and a nonorthogonal array for three two-level factors. Notice in the orthogonal array each level of factor A occurs once with each level of factors B and C. However, in the nonorthogonal array, the first level of A does not occur an equal number of times in combination with each level of C, and the second level of A does not occur with the second level of B.

Orthogonal arrays are highly desirable in the design of experiments because they possess many important statistical properties, such as the statistical independence of estimated factor effects. All the complete and fractional factorials presented in this chapter have been orthogonal arrays. Certain sets of orthogonal arrays have gained widespread acceptance and usage in experimentation due to the emphasis placed on them by the Japanese engineer Genichi Taguchi, who was instrumental in promoting their use using simple tables.

An illustration of the ease with which tables of orthogonal arrays can be used in practical applications is the L_9 (3^4) array shown in Table 7.9. L-arrays, many of which are Latin Square or Greco-Latin Square design layouts

TABLE 7.8 Example of an Orthogonal and a Nonorthogonal Array

Orthogonal Array			Nonorthogonal Array		
Factors: A	B	C	Factors: A	B	C
1	1	1	1	1	2
1	2	2	1	2	1
2	1	2	2	1	2
2	2	1	1	2	1

TABLE 7.9 An L_9 (3^4) Orthogonal Array

Run No.	Factor *A*	Factor *B*	Factor *C*	Factor *D*
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	2	1	2	3
5	2	2	3	1
6	2	3	1	2
7	3	1	3	2
8	3	2	1	3
9	3	3	2	1

(see Chapter 9), are highly efficient in terms of the number of test runs and are effective whenever fractional factorials of Resolution-III (main effects are unaliased only with other main effects) can be conducted. The L_9 (3^4) array in Table 7.9 can be used in experiments in which 4 factors each having 3 levels (1, 2, and 3) are to be investigated. The experimental design consists of only nine test runs. It is a 3^{4-2} fractional factorial experiment and is Resolution-III. In general, these design layouts are labeled $L_n(3^k)$ arrays where the number of test runs n is some power p of 3, and $k = (n - 1)/2$ represents the greatest number of factors that can be investigated using the design layout.

Many of the orthogonal arrays advocated in practice consist of very few test runs. These design layouts are often *saturated* (see Exhibit 7.7) in that they can be used with up to one fewer factors than test runs — the maximum number of factors possible. The key design property that is often overlooked is that the resulting designs are only of Resolution-III. The presence of any interaction effects bias the main effects. Orthogonal arrays for three-level fractional factorial experiments with higher resolution are available; see, for example, Gunst and Mason (1991).

EXHIBIT 7.7 SATURATED DESIGNS

An experimental design is saturated if the degrees of freedom for all effects (constant, main effects, interactions) equal the number of unique factor–level combinations included in the experiment (excluding repeats). A saturated 2^k experiment in which only main effects are to be analyzed has $n = k + 1$. In saturated designs that have no repeat tests, the overall mean, main effects, and interactions (if any) can be estimated but experimental error variation cannot be estimated.

7.5 COMBINED TWO- AND THREE-LEVEL FRACTIONAL FACTORIALS

Experiments with some factors having two levels and the remainder having three are an excellent compromise between (a) the high efficiency of two-level designs and (b) the greater generality and the ability to detect curvature when factors are quantitative that are afforded by three-level designs. Even so, complete factorials in experiments with combinations of two- and three-level factors can still require large numbers of test runs.

Procedures for constructing fractional factorials for experiments with combinations of two-level and three-level factors are similar to the procedures used in constructing 2^k and 3^k fractional factorial experiments. To illustrate, consider an experiment that is to investigate the effects of five fuel properties on NO_x (oxides of nitrogen) exhaust emissions of a heavy-duty diesel engine. The five fuel properties to be used as factors are (1) fuel density, (2) base cetane, (3) delta cetane, a specified change in cetane from the base amount, and (4) monoaromatics and (5) polyaromatics, which are associated with the hydrogen content of the fuel. It is desired to use three levels of the cetane content of the fuel but only two levels (e.g., low and high) of the other four fuel properties. Thus, there are a total of $2^4 \times 3^1 = 48$ potential test fuels that must be blended if a complete factorial experiment is to be conducted. Because blending specialized fuels and testing them on an engine are very time-consuming and costly, it is decided to fractionate the design. A second reason for deciding to fractionate the design is that the primary goal of this experiment is to identify one or two key fuel properties that are very influential on the NO_x emissions and then conduct more extensive studies on these properties.

Among the many possible fractional factorials that can be used to reduce the number of test runs for this experiment is the one shown in Table 7.10. This

TABLE 7.10 One-Sixth Fraction of a 3×2^4 Complete Factorial Experiment

Test No.	*A*	*B*	*C*	*D*	*E*
1	1	1	1	1	1
2	1	2	2	2	2
3	2	1	1	2	2
4	2	2	2	1	1
5	3	1	2	1	2
6	3	2	1	2	1
7	2	1	2	2	1
8	2	2	1	1	2

Resolution-III design layout has one column for a three-level factor and four columns for two-level factors. This experimental layout is a one-sixth fraction of a 3×2^4 complete factorial experiment. It should be stressed again that this highly fractionated experiment, while very efficient in terms of the number of test runs, risks the possibility of biased main effects if any interaction effects dominate the main effects.

The fractional factorial in Table 7.10 was constructed from a 2^{6-3} Resolution-III fractional factorial. Two of the two-level factors were recoded into a single three-level factor. The middle level for the three-level factor was created from two combinations of the four combined levels of the original two factors. The two columns for the original two factors were replaced by a single column for the three-level factor, the first column in Table 7.10. This procedure can be generalized to obtain many fractional factorials for combined two-and three-level complete factorial experiments. One simply uses any pair of columns for two two-level factors to represent each three-level factor. Alternatively, there are many tables of orthogonal arrays for combined two- and three-level factors.

Beyond two- and three-level factors, very little can be said in general about the properties of fractional factorial experiments. One ordinarily must evaluate each specific fractional factorial layout to determine its properties. The properties of two-level factorial experiments can, however, be extended to factors whose number of levels is a power of two. This extension is briefly outlined in the Appendix to this chapter.

7.6 SEQUENTIAL EXPERIMENTATION

This chapter and Chapter 5 provide a basic set of statistical designs and fundamental design principles that can be applied in a wide variety of settings. The vast majority of statistically designed experiments use complete and fractional factorial experiments in completely randomized designs. There are, of course, many additional types of statistical designs that are available for use when circumstances warrant. Over the next several chapters of this book many of the most important and widely used alternatives to factorial experiments in completely randomized designs will be discussed. Prior to doing so, it is important to emphasize that experimental designs are statistical tools that are part of an overall effective strategy of experimentation.

The reality of experimentation is that experiments are performed to answer questions. These questions can often be posed in terms of the effects of factors on responses and in terms of the limits of uncertainty. Experiments only provide new knowledge to the extent that they are able to answer the posed questions. Statistical experimental design is only an aid to experimentation when designs permit these questions to be answered efficiently in terms of

experimental resources (time, budget, personnel, equipment, etc.). In addition, statistical designs are by their nature intended to prevent ambivalent conclusions due to avoidable (by the design) biases or because of no or invalid estimates of uncertainly.

Within an overall strategy of experimentation is a statistical strategy of experimentation. One does not simply select factors, levels, and a design. A comprehensive design strategy could end up with one single definitive experiment, but equally often the strategy would involve a sequence of experiments with decisions about further experimentation being made after each experiment in the sequence. In this section, some insight into how a sequence of experiments might be planned is given within the context of complete and fractional factorial experiments. The first step in a sequence of experiments is sometimes a screening experiment to identify key factors that will be examined more comprehensively in subsequent experiments. Screening experiments are discussed in Section 7.6.1 and then further comments are made on sequential experiments in Section 7.6.2.

7.6.1 Screening Experiments

Screening experiments are conducted when a large number of factors are to be investigated but limited resources mandate that only a few test runs be conducted. Screening experiments are conducted to identify a small number of dominant factors, often with the intent to conduct a more extensive investigation involving only these dominant factors.

Saturated (see Exhibit 7.7) Resolution-III fractional factorials are commonly used as screening experiments because they allow the estimation of main effects with very few test runs. Often R_{III} designs result in experiments that clearly identify the dominant factors without the major time and expense that would be required for higher-resolution designs.

A special class of two-level fractional factorial experiments that is widely used in screening experiments was proposed by Plackett and Burman (see the references). These experiments have Resolution-III when conducted in completely randomized designs and are often referred to as *Plackett–Burman* designs.

The designs discussed by Plackett and Burman are available for experiments that have the number of test runs equal to a multiple of four. Table 7.A.2 lists design generators for experiments having 12, 16, 20, 24, and 32 test runs. The rows of the table denote the design factors, and the elements in each column are coded factor levels: a minus sign denotes one level of a factor and a plus sign denotes the other factor level. To allow for an estimate of experimental error, it is recommended that the design have at least six more test runs than the number of factors included in the experiment.

The particular design selected depends on both the number of factors in the experiment and the number of test runs. Each design generator can be used to construct a saturated screening design having up to one fewer factor than the number of test runs. For example, the 16-run design generator can be used to construct screening designs for 1–15 factors; however, we again recommend that no more than 10 factors be used with a 16-run design.

To illustrate the construction of screening experiments using the design generators in Table 7.A.2, consider the construction of a twelve-run design. The first column of Table 7.A.2 is used as the first row of a (nonrandomized) twelve-run screening design. Succeeding rows are obtained by taking the sign in the first position of the previous row and placing it last in the current row, and then shifting all other signs forward one position. After all cyclical arrangements have been included in the design (a total of one less than the number of test runs), a final row of minus signs is added. The complete twelve-run design is shown in Table 7.11.

If, as recommended, fewer factors than test runs are included in the design, several columns of the generated design are discarded. Any of the columns can be eliminated. Elimination can be based on experimental constraints or be random. All rows (test runs) for the retained columns are included in the design.

The acid-plant corrosion-rate study was actually conducted as a screening experiment. The 12-run screening experiment shown in Table 7.11 satisfied the recommended requirement of six more test runs than factors. For ease of discussion, the first six columns of Table 7.11 are used; the remaining columns are discarded. The combination of factor levels to be used in the experimental

TABLE 7.11 Twelve-Run Screening Design (Not Randomized)

	Factor No.										
Run No.	1	2	3	4	5	6	7	8	9	10	11
1	+	+	−	+	+	+	−	−	−	+	−
2	+	−	+	+	+	−	−	−	+	−	+
3	−	+	+	+	−	−	−	+	−	+	+
4	+	+	+	−	−	−	+	−	+	+	−
5	+	+	−	−	−	+	−	+	+	−	+
6	+	−	−	−	+	−	+	+	−	+	+
7	−	−	−	+	−	+	+	−	+	+	+
8	−	−	+	−	+	+	−	+	+	+	−
9	−	+	−	+	+	−	+	+	+	−	−
10	+	−	+	+	−	+	+	+	−	−	−
11	−	+	+	−	+	+	+	−	−	−	+
12	−	−	−	−	−	−	−	−	−	−	−

TABLE 7.12 Actual Factor Levels for a 12-Run Screening Design for the Acid-Plant Corrosion-Rate Study

	Factor					
Factor–Level Combination	Feed Rate (pph)	Gas Temp (°C)	Scrubber Water (%)	Bed Acid (%)	Exit Temp (°C)	Distribution Point
1	6000	220	5	30	360	West
2	6000	100	20	30	360	East
3	3000	220	20	30	300	East
4	6000	220	20	20	300	East
5	6000	220	5	20	300	West
6	6000	100	5	20	360	East
7	3000	100	5	30	300	West
8	3000	100	20	20	360	West
9	3000	220	5	30	360	East
10	6000	100	20	30	300	West
11	3000	220	20	20	360	West
12	3000	100	5	20	300	East

program are shown in Table 7.12. To minimize the possibility of bias effects due to the run order, the experimental test sequence should be randomized as should the assignment of the factors to the columns.

Screening designs often lead to further experimentation once dominant factors are identified. More generally, one often can plan experiments so that preliminary results determine the course of the experimentation.

7.6.2 Designing a Sequence of Experiments

Sequential experimentation is highly advisable when one embarks on a new course of experimental work in which little is known about the effects of any of the factors on the response. It would be unwise, for example, to design a complete factorial experiment in seven or more factors in such circumstances. The result of such a large test program might be that a small number of main effects and two-factor interactions are dominant. If so, a great amount of time and effort would have been wasted when a much smaller effort could have achieved the same results.

If an experiment is contemplated in which a large number of factors are to be investigated, a key question that should be asked is whether the project goals will be met if only the dominant factor effects are identified. If so, a screening

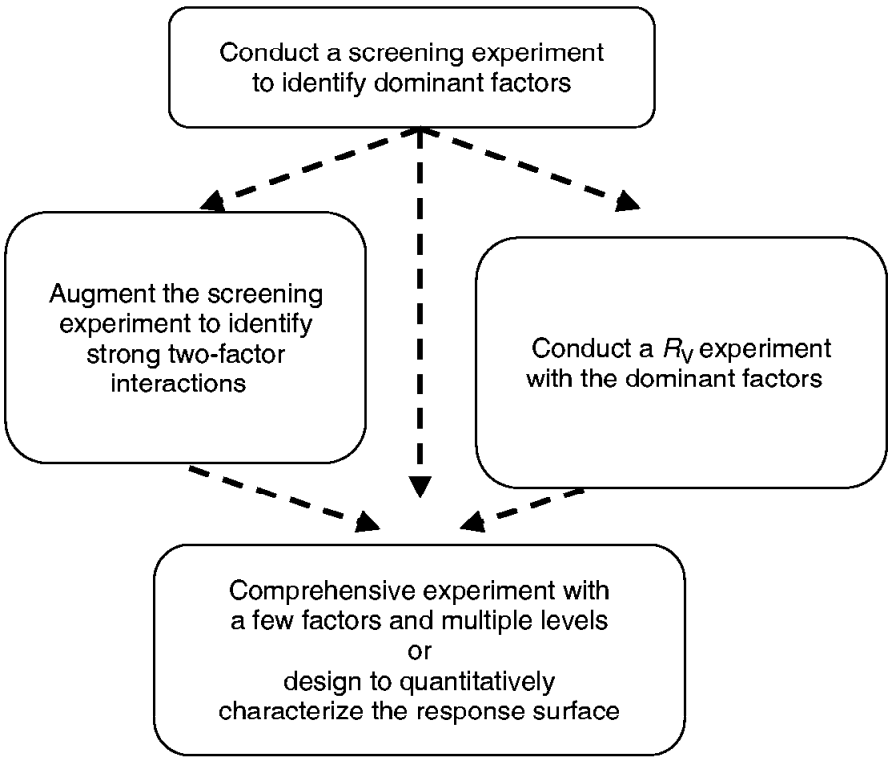


Figure 7.7 A simple strategy for a sequence of experiments.

experiment should be performed, followed by a complete or fractional factorial experiment involving only the dominant factors. If not, perhaps a Resolution-V fractional factorial experiment would suffice.

Figure 7.7 is a simple schematic of one useful strategy for conducting a sequence of experiments when a moderate to large number of factors is to be investigated. The first stage of the sequence of experiments would consist of a screening experiment to identify the most dominant factors. Plackett–Burman screening designs or any other R_{III} designs can be used at this stage. Many alternatives to these classes are also available, including *supersaturated designs* for which the number of test runs is less than the number of main effect degrees of freedom. See Lin (1993, 1995) for more information on supersaturated designs and Abraham, Chipman, and Vijayan (1999) for concerns about these designs and the resulting analyses of the main effects.

When dominant main effects are identified, a number of alternatives are available for the second stage of the sequence of experiments. If a R_{III} design is a quarter, eighth, etc. fraction of a complete factorial, adding other fractions of the complete factorial can increase the resolution of the composite design. If a sufficient number of main effects have been identified as dominant, this may be the safest alternative for the identification of strong two-factor

interactions. This approach may be wasteful of resources if only a few main effects are identified as dominant. Other alternatives are available, including *foldover designs.*

Foldover designs augment a R_{III} design with another fraction having an equal number of test runs and specific reversals of factor levels. These designs are intended to unalias specific interactions from main effects and one another. These designs are especially useful if the confounding pattern leads one to be concerned that an apparent significant main effect might be due to an aliased two-factor interaction involving two of the other significant main effects.

One of the most popular techniques for constructing foldover designs from two-level R_{III} fractional factorials is to use as the second stage experiment the original design with the levels of all the factors in the first fraction exactly reversed. This combined design is R_{IV} or higher when the two fractions are combined. Thus, the R_{III} initial experiment in which main effects are aliased with two-factor interactions has been augmented to produce a R_{IV} design in which main effects are not aliased with two-factor interactions.

A second popular method for constructing foldover designs from two-level R_{III} fractional factorials is to use as the second stage experiment the original design with the levels of only one of the factors exactly reversed. This composite design allows the main effect for the factor that has its levels reversed to be estimated along with all two-factor interactions involving that factor. The estimators of these effects are not aliased with any other main effects or two-factor interactions.

In general, a foldover design can be formed from any fraction of a factorial experiment where more information is desired on the main effects and interactions. However, the use of foldovers with designs of Resolution-V or higher is rare because the main effects and two-factor interactions are unconfounded with one another in these experiments, thus, little would be gained by using them.

The final sequential segment in Figure 7.7 concerns the comprehensive evaluation of a few factors. Complete factorials are excellent for this purpose, often with more than two levels for each of the remaining factors. If the factors have quantitative levels, statistical modeling of the response as a function of the factor levels can lead to a geometric representation for the response. Highly efficient response surface designs (see Chapter 17) are available for this purpose.

APPENDIX: FRACTIONAL FACTORIAL DESIGN GENERATORS

1. Added Factors

For a large number of factors, a time-saving device for constructing fractional factorial experiments is the technique of *added factors.* Rather than examining the effects representation of all the factor–level combinations and then

selecting a portion of them for the design, new factors are added to complete factorial experiments involving fewer than the desired number of factors to construct fractional factorial experiments involving a large number of factors. For example, consider setting a sixth factor equal to a five-factor interaction. By setting the levels of this sixth factor equal to those in the effects representation of the five-factor interaction, one would construct a half-fraction of a complete factorial in six factors from the complete factorial in five factors. Symbolically, this substitution can be expressed (apart from the sign) as

$$F = ABCDE.$$

Algebraically multiplying both sides of this equation by F produces the defining equation for a half-fraction of a six-factor factorial: $I = ABCDEF$.

This approach can be effectively utilized to construct fractional factorial experiments from complete factorials of fewer factors. In Table 7A.1, the last factor listed in each defining equation can be equated to the product of the previous factors without changing the contrasts, simply by multiplying both sides of each defining equation by the factor. This process can be implemented on a computer by the following procedure.

Let $x_1, x_2, \ldots, x_k$ denote the k factors in a fractional factorial experiment. For each factor level in the complete factorial experiment, let $x_j = 0$ if the jth factor is at its lower level (-1), and let $x_j = 1$ if the factor is at its upper level $(+1)$. Replace the defining contrasts for the fractional factorial experiment desired by an equal number of sums of the form

$$x_w = x_u + \cdots + x_v,$$

where x_w represents the added factor and $x_u, \ldots, x_v$ represent the other factors involved in the defining contrast. These sums are calculated for each factor–level combination in the complete factorial for the factors on the right side of the added-factor equation.

For example, to construct a half fraction of a six-factor factorial experiment, let $x_1, \ldots, x_6$ represent, respectively, factors $A, \ldots, F$. An effects-representation table for the complete factorial experiment in the first five factors can be constructed by setting the elements in the columns for $x_1, \ldots, x_5$ equal to 0 or 1 rather than -1 or 1, respectively. The added factor is then

$$x_6 = x_1 + x_2 + x_3 + x_4 + x_5.$$

For each factor–level combination in the complete five-factor factorial experiment, compute the value of x_6 from the above equation. If x_6 is even (0 mod 2), assign the sixth factor its upper level. If x_6 is odd (1 mod 2), assign the

sixth factor its lower level. The 32 factor–level combinations in the half fraction of this 2^6 factorial are identified by replacing the zeros in the table with the lower levels of each factor and the ones with the upper levels of each.

This process is valid for smaller fractions. If one desires to construct a $\frac{1}{16}$ fraction of an eleven-factor factorial experiment, a Resolution-V design can be obtained using the added factors listed in Table 7A.1:

$$x_8 = x_1 + x_2 + x_3 + x_7, \qquad x_9 = x_2 + x_3 + x_4 + x_5, \qquad \text{etc.}$$

For each factor–level combination in a complete seven-factor factorial experiment, one would compute x_8, x_9, x_{10}, and x_{11}. As these values for x_j are computed, if the sign on the corresponding defining contrast is +, the value of $x_j \pmod 2$ is used. If the sign is $-$, $x_j + 1 \pmod 2$ is used. Once the values of $x_8, \ldots, x_{11}$ are determined, all the zeros and ones for the eleven factors are replaced by the lower and upper levels of the respective factors.

2. Factors with More Than Two Levels

Beyond two- or three-level factors very little can be said in general about the properties of fractional factorial experiments. One ordinarily must evaluate each design to determine its properties. The properties of two-level factorial experiments can, however, be extended to factors whose number of levels is a power of two. In doing so, it is imperative that one understand the nature of effects for factors having more than two levels (see the appendix to Chapter 5).

Consider as an illustration an experiment in which three factors are to be investigated, factor A having four levels and factors B and C each having two levels. As recommended in the appendix to Chapter 5, define two new factors A_1 and A_2, each having two levels, to represent the four levels of factor A. One can construct a half fraction of the complete factorial experiment in these factors using the defining equation (apart from sign)

$$I = A_1A_2BC.$$

In interpreting the properties of this design, one might conclude that the design is of resolution IV, because four factors are present in the defining equation. One must recall, however, that A_1A_2 represents one of the main effects for the four-level factor A. Thus, this design is not R_{IV}; it is R_{III}, because the confounding pattern for the main effects is

$$A_1 = A_2BC, \quad A_2 = A_1BC, \quad A_1A_2 = BC, \quad B = A_1A_2C, \quad C = A_1A_2B.$$

This design is not R_{IV} because, for example, the main effect A_1A_2 is confounded with the two-factor interaction BC.

TABLE 7A.1 Selected Fractional Factorial Experiments

Number of Factors	Number of Test Runs	Fraction	Resolution	Defining Equations	Added Factors
5	8	$\frac{1}{4}$	III	$I = ABD$	$4 = 12$
				$I = ACE$	$5 = 13$
6	8	$\frac{1}{8}$	III	$I = ABD$	$4 = 12$
				$I = ACE$	$5 = 13$
				$I = BCF$	$6 = 23$
	16	$\frac{1}{4}$	IV	$I = ABCE$	$5 = 123$
				$I = BCDF$	$6 = 234$
7	8	$\frac{1}{16}$	III	$I = ABD$	$4 = 12$
				$I = ACE$	$5 = 13$
				$I = BCF$	$6 = 23$
				$I = ABCG$	$7 = 123$
	16	$\frac{1}{8}$	IV	$I = ABCE$	$5 = 123$
				$I = BCDF$	$6 = 234$
				$I = ACDG$	$7 = 134$
	32	$\frac{1}{4}$	IV	$I = ABCDF$	$6 = 1234$
				$I = ABDEG$	$7 = 1245$
8	16	$\frac{1}{16}$	IV	$I = BCDE$	$5 = 234$
				$I = ACDF$	$6 = 134$
				$I = ABCG$	$7 = 123$
				$I = ABDH$	$8 = 124$
	32	$\frac{1}{8}$	IV	$I = ABCF$	$6 = 123$
				$I = ABDG$	$7 = 124$
				$I = BCDEH$	$8 = 2345$
	64	$\frac{1}{4}$	V	$I = ABCDG$	$7 = 1234$
				$I = ABEFH$	$8 = 1256$
9	16	$\frac{1}{32}$	III	$I = ABCE$	$5 = 123$
				$I = BCDF$	$6 = 234$
				$I = ACDG$	$7 = 134$
				$I = ABDH$	$8 = 124$
				$I = ABCDJ$	$9 = 1234$
	32	$\frac{1}{16}$	IV	$I = BCDEF$	$6 = 2345$
				$I = ACDEG$	$7 = 1345$
				$I = ABDEH$	$8 = 1245$
				$I = ABCEJ$	$9 = 1235$
	64	$\frac{1}{8}$	IV	$I = ABCDG$	$7 = 1234$
				$I = ACEFH$	$8 = 1356$
				$I = CDEFJ$	$9 = 3456$
	128	$\frac{1}{4}$	VI	$I = ACDFGH$	$8 = 13467$
				$I = BCEFGJ$	$9 = 23567$

(*continued overleaf*)

TABLE 7A.1 (*continued*)

Number of Factors	Number of Test Runs	Fraction	Resolution	Defining Equations	Added Factors
10	16	$\frac{1}{64}$	III	$I = ABCE$	$5 = 123$
				$I = BCDF$	$6 = 234$
				$I = ACDG$	$7 = 134$
				$I = ABDH$	$8 = 124$
				$I = ABCDEJ$	$9 = 1234$
				$I = ABK$	$10 = 12$
	32	$\frac{1}{32}$	IV	$I = ABCDF$	$6 = 1234$
				$I = ABCEG$	$7 = 1235$
				$I = ABDEH$	$8 = 1245$
				$I = ACDEJ$	$9 = 1345$
				$I = BCDEK$	$10 = 2345$
	64	$\frac{1}{16}$	IV	$I = BCDFG$	$7 = 2346$
				$I = ACDFH$	$8 = 1346$
				$I = ABDEJ$	$9 = 1245$
				$I = ABCEK$	$10 = 1235$
	128	$\frac{1}{8}$	V	$I = ABCGH$	$8 = 1237$
				$I = BCDEJ$	$9 = 2345$
				$I = ACDFK$	$10 = 1346$
11	16	$\frac{1}{128}$	III	$I = ABCE$	$5 = 123$
				$I = BCDF$	$6 = 234$
				$I = ACDG$	$7 = 134$
				$I = ABDH$	$8 = 124$
				$I = ABCDJ$	$9 = 1234$
				$I = ABK$	$10 = 12$
				$I = ACL$	$11 = 13$
	32	$\frac{1}{64}$	IV	$I = ABCF$	$6 = 123$
				$I = BCDG$	$7 = 234$
				$I = CDEH$	$8 = 345$
				$I = ACDJ$	$9 = 134$
				$I = ADEK$	$10 = 145$
				$I = BDEL$	$11 = 245$
	64	$\frac{1}{32}$	IV	$I = CDEG$	$7 = 345$
				$I = ABCDH$	$8 = 1234$
				$I = ABFJ$	$9 = 126$
				$I = BDEFK$	$10 = 2456$
				$I = ADEFL$	$11 = 1456$
	128	$\frac{1}{16}$	V	$I = ABCGH$	$8 = 1237$
				$I = BCDEJ$	$9 = 2345$
				$I = ACDFK$	$10 = 1346$
				$I = ABCDEFGL$	$11 = 1234567$

TABLE 7.A.2 Screening Design Generators

Factor No.	No. of Test Runs					Factor No.
	12	16	20	24	32	
1	+	+	+	+	−	1
2	+	−	+	+	−	2
3	−	−	−	+	−	3
4	+	−	−	+	−	4
5	+	+	+	+	+	5
6	+	−	+	−	−	6
7	−	−	+	+	+	7
8	−	+	+	−	−	8
9	−	+	−	+	+	9
10	+	−	+	+	+	10
11	−	+	−	−	+	11
12		−	+	−	−	12
13		+	−	+	+	13
14		+	−	+	+	14
15		+	−	−	−	15
16			−	−	−	16
17			+	+	−	17
18			+	−	+	18
19			−	+	+	19
20				−	+	20
21				−	+	21
22				−	+	22
23				−	−	23
24					−	24
25					+	25
26					+	26
27					−	27
28					+	28
29					−	29
30					−	30
31					+	31

REFERENCES

Test References

Extensive discussion of complete and fractional factorial experiments appears in the texts at the end of Chapter 4.

An extensive list of fractional factorial experiments in completely randomized and blocking designs is contained in Box, Hunter, and Hunter (1978) *and in the following references*:

Anderson, V. L. and McLean, R. A. (1984). *Applied Factorial and Fractional Designs*, New York: Marcel Dekker, Inc.

Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*, New York: John Wiley and Sons, Inc.

Gunst, R. F. and Mason, R. L. (1991). *How To Construct Fractional Factorial Experiments*, Milwaukee, WI: Quality Press.

Included in the above three references are fractional experiments for three-level factors and combinations of two- and three-level factors.

The material in Table 7A.1 is compiled from [see also Box, Hunter, and Hunter (1978)]:

Box, G. E. P and Hunter, J. S. (1961). "The 2^{k-p} Fractional Factorial Designs, Part I, II," *Technometrics*, **3**, 311–351, 449–458.

The fractional factorial screening designs proposed by Plackett and Burman are derived in the following article:

Plackett, R. L. and Burman, J. P. (1946). "The Design of Optimum Multifactorial Experiments," *Biometrika*, **33**, 305–325.

These screening designs are also discussed in the book by Cochran and Cox cited above.

A useful description of supersaturated designs and tables and algorithms associated with them is given in the following two papers:

Lin, D. J. (1993). "A New Class of Supersaturated Designs," *Technometrics*, **35**, 28–31.

Lin, D. J. (1995). "Generating Systematic Supersaturated Designs," *Technometrics*, **37**, 213–225.

Cautions over the indiscriminant use of supersaturated designs and difficulties with the assessment of main effects are noted in:

Abraham, B, Chipman, H., and Vijayan, K. (1999). "Some Risks in the Construction and Analysis of Supersaturated Designs," *Technometrics*, **41**, 135–141.

Some useful graphical methods for displaying fractional factorial designs can be found in the following:

Barton, R. R. (1998). "Design-Plots for Factorial and Fractional Factorial Designs," *Journal of Quality Technology*, **30**, 40–54.

EXERCISES

1 Construct a half fraction of the complete factorial experiment for the experiment on the intensity of light described in Exercise 12 of Chapter 5. What is the defining contrast for the half fraction? What is the resolution of the design? What is the confounding pattern for the effects?

2 Construct a quarter fraction of the complete factorial experiment on light intensity. What are the defining contrasts? What is the resolution of the design? What is the confounding pattern for the effects?

3 Construct a half fraction (of highest resolution) of a four-factor, two-level complete factorial experiment. Show that this design is a complete factorial in any three of the factors.

4 Construct a Resolution-V design for an eight-factor, two-level experiment in which the number of test runs cannot exceed 70. Verify that this a complete factorial experiments in any one set of four factors.

5 An experiment is to have three factors each at two levels and one factor at four levels. Suppose a maximum of 20 test runs can be made. Construct a half fraction of the four-factor experiment. Specify the resolution of the design, and write out the confounding pattern.

6 A ruggedness test is to be conducted on laboratory procedures for the determination of acid concentration in a solution. The factors of interest are the temperature and the humidity of the laboratory, the speed of the blender used to mix the solution, the mixing time, the concentration of acid in the solution, volume of the solution, the size of the flask containing the solution, the presence or absence of a catalyst, and the laboratory technicians. Each of these factors is to be investigated at two levels. Design a screening experiment consisting of no more than 20 test runs to investigate the main effects of these factors.

7 Can a fractional factorial experiment of greater than Resolution-III be designed for the experiment in Exercise 6? If so, state the defining contrasts and the resolution of the design.

8 An experiment is to be conducted to determine factors that would increase the displacement efficiency of material used in cementing operations in wellbore drilling holes in oil fields. It is suspected that any of a number of factors might affect the displacement efficiency, including the following:

Factor	Levels
Hole angle	45, 60°
Cement viscosity	30, 40 centipoise
Flow rate	50, 175 gal/min
Production-string rotation speed	0, 60 rpm
Type of wash material	Plain, caustic water
Wash volume	5, 10 barrels/annular volume
Production string eccentricity	Concentric, 30° offset

Design a screening experiment to investigate the effects of these factors. What properties does this screening design possess (e.g., resolution, confounding).

9 The endurance strength (psi) of several different stainless steels is being studied under varying fatigue cycles and temperatures. It is suspected that the temperature and fatigue cycles to which each specimen is subjected are jointly contributing to the strength of the material. Four stainless steels (302 SS, 355 SS, 410 SS, 430 SS) are to be tested under four temperatures and two fatigue cycles. Design an experiment to study the effects of these factors on the strength of stainless steel. Design for three alternatives:

(a) any reasonable number of test runs can be conducted,

(b) at most 25 test runs can be conducted,

(c) at most 15 test runs can be conducted.

Discuss the properties of these three designs. Are there any that you would recommend against? Why (not)?

10 Construct a one-third fraction of an experiment having four factors, each with three levels. Present the defining contrasts for the one-third fraction. What is the resolution of the design? What is the confounding pattern for the effects?

11 The $L_9(3^4)$ orthogonal array presented in Table 7.9 is a fractional factorial experiment. Construct a one-ninth fraction of a 3^4 experiment, and compare the results with the orthogonal array in Table 7.9.

12 A study was conducted to determine the percent of paint removed from aircraft exteriors by using various chemical paint-strippers. The selected paint coatings (and their associated primers) were to be applied to 1-ft square metal samples (1 inch thick) in a test lab under varying temperature and humidity levels. Some samples were to be painted with a single film of paint while others were to be painted with two layers of paint. Some of the painted samples were to be treated with ultraviolet light to age the paint, while others were untreated. The following six factors and factor levels were identified for the project:

Factor	Levels
Chemical stripper	Type 1, Type 2, Type 3
Coating manufacturer	Manf. 1, Manf. 2, Manf. 3
Temperature	Lo, Hi
Relative humidity	Lo, Hi
Coating film thickness	Single layer, Multilayer
Coating age condition	New, Aged

To reduce the size of the test matrix, a fractionated experiment was desired. To solve this problem, construct a one-eighth fraction of a 2^8 experiment so that the resulting fraction is a Resolution III design. What is the defining contrast? Following the steps outlined in Section 7.5, construct a design similar to the one given in Table 7.10.

13 Construct a quarter fraction of a 2^5 experiment. By reversing the signs of the first factor, construct a second fraction. Combine the two fractions to obtain a foldover design. Show that the first factor and all two-way interactions with the first factor are not aliased with any of the other main factor effects.

14 Using the first quarter fraction constructed in Exercise 13, construct a second quarter fraction by reversing the signs of all five factors. Combine the two fractions to obtain a foldover design. Show that the combined design is of Resolution-IV or higher.

15 Suppose an experimenter constructs a half-fraction of a 2^5 experiment using as the defining contrast $I = -ABCD$. Show that AB and $-CD$ are confounded in this design by using the effects representation of them.

16 Design an experiment to investigate the effects of different drilling tool configurations on the wear of drill bits. The response variable is the wear measurement, and the five factors and their levels are given below.

Factor	Levels
Rotational Drill Speed	60, 75 rpm
Longitudinal Velocity	50, 100 fpm
Drill Pipe Length	200, 400 ft
Drilling Angle	30, 60 degrees
Tool Joint Geometry	Straight, ellipsoidal edges

Construct a design using no more than 25 test runs, including at least 5 repeat runs. List the test runs and effects. What is the resolution of your design?

17 Construct a one-quarter fraction of a 2^5 experiment. Because this design requires only 8 runs (which is a multiple of four), do you think it could be used as a Plackett–Burman screening design? Why or why not?

18 Suppose the particulate emissions from an two-cycle outboard boat motor of a specified size are being collected in an experiment involving six factors, each at two levels. The factors include the following: engine lubrication mix (A or B), sampling dilution ratio (1 or 2), water quality (clean or dirty), water type (salt water or fresh water), water injection spray rate (low or high), and water injection spray quality (course or fine). If the experimenter desires to run the experiment in only eight runs, select an appropriate design. What is the resolution of the design? List the defining contrasts.

19 Show that the design in Exercise 1 is an orthogonal array. Do you think that all fractional factorial designs are based on orthogonal arrays? Explain your answer.

20 Suppose you had eight two-level factors to consider in an experiment. Discuss the pros and cons of using a design of Resolution-III versus selecting one of Resolution-IV. Describe situations where each design would be preferred.

CHAPTER 8

Analysis of Fractional Factorial Experiments

In this chapter, procedures that are commonly used to analyze fractional factorial experiments are presented. The procedures are similar to the analysis-of-variance methods introduced in Chapter 6 for completely randomized designs having fixed effects. The procedures are extended in this chapter to unbalanced, completely randomized designs. The major topics discussed in this chapter are:

- *analysis of unbalanced, completely randomized designs with fixed effects,*
- *analysis of two-level, and three-level, fractional factorial experiments,*
- *analysis of fractional experiments containing a mix of two-level and three-level factors, and*
- *analysis of screening experiments.*

General procedures for analyzing multifactor experiments were presented in Chapter 6 for balanced fixed-effects experiments conducted in completely randomized designs. Those formulas and analyses are not necessarily correct when designs are unbalanced. For example, fractional factorial experiments are, by definition, unbalanced because some factor–level combinations occur in the experiment while others do not. In addition, if a fractional factorial experiment contains repeats, not all of the factor–level combinations need be repeated an equal number of times. In this chapter, methods for analyzing data from general unbalanced designs are presented, with special emphasis on fractional factorial experiments.

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

Statistical methods that are useful for analyzing unbalanced, completely randomized designs are necessarily more complex than those for balanced designs and must be chosen with the specific goals of an experiment in mind. Sections 8.1 and 8.2 present two general approaches that can be applied to data from unbalanced designs. Applications of these general approaches to data taken from fractional factorial experiments with all factors having only two levels are given in Section 8.3. Analysis of three-level, fractional factorial experiments is detailed in Section 8.4, with special emphasis on decomposing the main effect sums of squares for each quantitative factor into single-degree-of-freedom sums of squares. A brief discussion of methods for fractional factorial experiments that contain both two-level and three-level factors is presented in Section 8.5. Application of the general method of Section 8.1 to screening experiments is given in Section 8.6.

8.1 A GENERAL APPROACH FOR THE ANALYSIS OF DATA FROM UNBALANCED EXPERIMENTS

In this section we present a general method for analyzing the fixed-effects factors. This method can be applied to many experimental designs, including certain unbalanced completely randomized designs. The method is most useful when the design is intended for the analysis of only main effects, such as Resolution-III fractional factorial designs and certain types of blocking designs (e.g., balanced incomplete block designs; see Section 9.3). This method can also be applied to nonexperimental data such as often accompany the fitting of regression models (see Chapters 14–16).

Two statistical models can be readily compared with respect to their ability to adequately account for the variation in the response if one of the models has terms that are a subset of the terms in the other model. Such *subset hierarchical models* (see Exhibit 8.1) allow one to assess main effects and interactions by comparing error sums of squares from different model fits. Error sums of squares can be calculated for any model, whether the design is balanced or not, and regardless of whether the model terms represent design factors or covariates.

EXHIBIT 8.1

Subset Hierarchical Models. Two statistical models are referred to as *subset hierarchical models* when the following two conditions are satisfied:

(1) each model contains only hierarchical terms (see Exhibit 6.4), and

(2) one model contains terms that are a subset of those in the other model.

With two subset hierarchical models, one can calculate the reduction in the error sum of squares (see Exhibit 8.2) between the simpler model and the more complex model as a measure of the effect of adding the specific terms to the larger model. With complete factorial experiments in balanced designs these reductions in error sums of squares are exactly equal to the sums of squares for the main effects and interactions (e.g., Table 6.2). When the designs are unbalanced or covariates are included, the reduction in error sums of squares provides the appropriate measure of the effects of the various model factors and covariates on the response.

EXHIBIT 8.2 REDUCTION IN ERROR SUMS OF SQUARES

Symbolically denote two statistical models by M_1 and M_2. Assume that the models are subset hierarchical models with the terms in M_2 a subset of those in M_1. Denote the respective error sums of squares by SSE_1 and SSE_2. The reduction in error sums of squares, $R(M_1|M_2)$, is defined as

$$R(M_1|M_2) = \text{SSE}_2 - \text{SSE}_1.$$

The number of degrees of freedom for this reduction equals the difference in the numbers of degrees of freedom for the two error sums of squares.

The reason for introducing the concept of reduction in error sums of squares is that even for very complicated models and/or experimental designs this procedure can be used to assess the effects of model factors and covariates. Statistical analyses of such models will ordinarily be performed with the aid of suitable computer software, not computing formulas. Thus, the ability to correctly formulate models and identify the appropriate sums of squares on computer printout can enable one to investigate many of these complicated designs and models. Examples of the use of reductions in error sums of squares are given throughout this text as needed. The procedures are general and are not restricted to any specific design or model.

As general as this method is for the calculation of sums of squares for model effects, one must be careful to adhere to the hierarchy principle. For example, with models that have both main effects and interactions, only sums of squares for the highest-order interactions can be calculated using the reduction in error sums of squares principle. One cannot assess main effects using the reduction in error sums of squares calculations if any interactions involving the main effects are present in the model. Note that this issue does not arise with data from most balanced experiments, including all balanced complete factorials. The next section discusses an alternative method of analysis that can be used with data from unbalanced designs and models that have both main effects

TABLE 8.1 Reduction in Error Sums of Squares for an Unbalanced Pilot-Plant, Chemical-Yield Experiment

Temperature (°C)	Concentration (%)	Catalyst	Yield (g)
160	20	C_1	59
		C_2	50,54
	40	C_1	
		C_2	46,44
180	20	C_1	74,70
		C_2	81
	40	C_1	69
		C_2	79

Reduction in Error Sums of Squares

Source of Variation	df	Sum of Squares	Mean Square	F-Value
(a) With Temperature × Catalyst				
Model	6	1682.40	280.40	46.73
Error	3	18.00	6.00	
Total	9	1700.40		
(b) Without Temperature × Catalyst				
Model	5	1597.07	319.41	12.36
Error	4	103.33	25.83	
Total	9	1700.40		
(c) Temperature × Catalyst F-Statistic				
Temperature × catalyst	1	85.33	85.33	14.22
Error	3	18.00	6.00	

and interactions. That alternative does not diminish the importance and the general applicability of the reduction in error sums of squares principle.

To illustrate the principle of reduction in error sums of squares, the pilot-plant, chemical-yield data of Table 6.1 were altered as shown in Table 8.1. The formulas in Table 6.2 for balanced designs cannot be now used to obtain the sums of squares, because the design is unbalanced. Some factor–level combinations have one response, some have two response values, and one does not have any response values. The formulas required to calculate the sums of squares for unbalanced designs depend on the model being fitted and

the number of observations for the various factor–level combinations. Rather than discuss formulas that would be appropriate for only this example, we rely on computer software to fit models to these data.

Suppose the only models of interest are those that do not include the three-factor interaction. If one wishes to draw inferences on, say, the two-factor interaction between temperature and catalyst, one fits two models: one with temperature-by-catalyst interaction terms and one without them, both models also excluding the three-factor interaction terms. Table 8.1 lists the model and error sums of squares for these two fits, as well as the reduction in error sums of squares attributable to the effect of the temperature-by-catalyst interaction. The F-statistic for the temperature-by-catalyst interaction is formed by dividing the mean square due to this interaction by the error mean square from the larger model, the one that includes the interaction. Note that, as in Table 6.4, the temperature-by-catalyst interaction has a large F-ratio.

Means, linear combinations of means, and effects parameters for analysis-of-variance models for unbalanced design data can be estimated similarly to the corresponding quantities for data from balanced designs. For example, all the estimation formulas discussed in Section 6.2.2 remain valid so long as one properly accounts for the unequal numbers of repeats in the formulas for standard errors. Specifically, the confidence interval in Equation (6.13) is still appropriate if m is replaced by

$$m = \sum_i \sum_j \sum_k \frac{a_{ijk}^2}{n_{ijk}},$$

where n_{ijk} is the number of repeats for the average $\bar{y}_{ijk\bullet}$. Comparisons among factor–level means would need similar adjustments. In a one-factor model, a confidence interval for the difference of two factor–level means $\theta = \mu_i - \mu_j$ would also use (6.13) with $m = n_i^{-1} + n_j^{-1}$.

The hypothesis testing procedures detailed in Sections 6.3 through 6.5 also are applicable to data from unbalanced designs provided that the formulas for standard errors properly account for the unequal numbers of repeats. As an example, consider the least significant interval plots described in Exhibit 6.13. When calculating the least significant difference (LSD), use $(n_i^{-1} + n_j^{-1})$ instead of 2/n (see Exhibit 6.9) and when calculating the Bonferroni significant difference (BSD), use n_i instead of n (see Exhibit 6.10).

When the design is unbalanced and the model is saturated, estimation of the error variance still requires that repeat tests be available. If there are repeat tests, the error mean square is still a pooled estimator of the common variance. The sample variances are now based on different numbers of degrees of freedom for each factor–level combination, and some factor–level combinations that do not have repeat tests will not contribute to

the estimate of the variance. The general form of the error mean square can be shown to equal

$$s_e^2 = \mathrm{MS}_E = \frac{1}{\nu} \sum_i \sum_j \sum_k (n_{ijk} - 1) s_{ijk}^2, \quad \nu = \sum_i \sum_j \sum_k (n_{ijk} - 1), \quad (8.1)$$

where the summations are only over those factor–level combinations for which there are at least two repeat tests ($n_{ijk} > 1$).

If there are no repeat tests and the model is saturated, there is no estimate of experimental error available. Equation (8.1) is not defined if none of the n_{ijk} are greater than 1, because none of the sample variances in the expression can be calculated. In this situation an estimate of experimental-error variation is only obtainable if some of the parameters in the saturated model can be assumed to be zero (see Section 6.2.1).

8.2 ANALYSIS OF MARGINAL MEANS FOR DATA FROM UNBALANCED DESIGNS

Computationally, a disadvantage to the use of reductions in error sums of squares for the analysis of general fractional factorial experiments is that a number of hierarchical models must be fit in order to comprehensively analyze main effects and interactions. A further complication for data from most unbalanced designs is that the precise form of hypotheses being tested depends on the numbers of repeats for various factor–level combinations and not solely on the model parameters themselves. For these reasons, much research has been conducted on alternative computing algorithms and on the most appropriate hypotheses to be tested when designs are not balanced. Of the many alternatives that are available, those based on *population marginal means* (Searle, Speed, and Milliken 1980) appear to be the most useful for data from unbalanced designs.

Marginal means analyses are based on the simple fact that a response mean (μ) for any factor–level combination can be estimated by the corresponding average ($\bar{y}$) of the responses for that combination, regardless of whether a design is balanced or not, so long as there is at least one response value for that combination. Because main effect and low-order interaction means are averages of higher-order interaction means, the former can be estimated unbiasedly so long as the latter can be, again regardless of whether a design is balanced or not. Hence, the analysis of main effects and interactions focuses solely on averages of averages.

For illustration purposes, consider a two-factor complete factorial experiment that is not balanced but for which there is at least one response value

for every combination of the levels (a and b) of the two factors. Write the model as

$$y_{ijk} = \mu_{ij} + e_{ijk} \quad i = 1, 2, \ldots, a; \quad j = 1, 2, \ldots, b; \quad k = 1, 2, \ldots, n_{ij}$$

where

$$\mu_{ij} = \alpha_i + \beta_j + (\alpha\beta)_{ij}.$$

No uniqueness constraints are imposed on the effects parameters as was done for balanced experiments in Section 6.1. Hence, a main effect average for the first factor is

$$\begin{aligned}\overline{\mu}_{i\bullet} &= b^{-1} \sum_{j=1}^{b} \mu_{ij} \\ &= \alpha_i + \overline{\beta}_\bullet + \overline{(\alpha\beta)}_{i\bullet}\end{aligned}$$

where the bars over the means and the effects parameters again denote averages.

Main effect and interaction hypotheses involving population marginal means are the same as those tested with balanced complete factorial experiments. Table 8.2 shows the hypotheses for a two-factor model. Note that the hypotheses are expressed in terms of the model means and not the effects parameters. For balanced experimental designs, Equation (6.3) relates individual effects parameters to the means in Table 8.2. Main effect and interaction hypotheses expressed in terms of effects parameters can readily be seen to be the same as those in Table 8.2.

The effects parameters need not be explicitly estimated nor are they needed for the tests of these hypotheses. This is especially important when designs are unbalanced. Analyses with marginal means are not based on the sums of squares presented in Chapter 6 for balanced complete factorials, but they reduce to these sums of squares when a design is balanced. Specialized software is needed to correctly calculate the sums of squares for marginal means hypotheses. Many analysis-of-variance software programs calculate sums of squares that do not test the hypotheses in Table 8.2. Rather, the tests for data

TABLE 8.2 Marginal Means Hypotheses

Effect	Hypothesis Tested
Main Effect for A	$\overline{\mu}_{1\bullet} = \overline{\mu}_{2\bullet} = \ldots = \overline{\mu}_{a\bullet}$
Main Effect for B	$\overline{\mu}_{\bullet 1} = \overline{\mu}_{\bullet 2} = \ldots = \overline{\mu}_{\bullet b}$
AB Interaction	$\overline{\mu}_{ij} - \overline{\mu}_{il} - \overline{\mu}_{kj} + \overline{\mu}_{kl} = 0$ for all i, j, k, l

from unbalanced designs in many software programs are functions of the numbers of data values in the averages used to calculate the sums of squares. Analyses of fractional factorial experiments using Type-III sums of squares from the General Linear Models Procedure, PROC GLM, in the SAS Software (SAS 2000) do test marginal means hypotheses. Examples using this software will be presented in the next two sections.

An important by-product of the analysis of marginal means is the comparison of individual factor–level averages. Multiple comparisons similar to those presented in Section 6.4 can be made, so long as the correct standard errors of the averages are used. This is straightforward for most analyses. The basic difference with analyses for balanced designs is that averages used to estimate means are calculated hierarchically as averages of averages. For example, for a two-factor design, the marginal mean average for the ith level of the first factor would be calculated as

$$\hat{\mu}_{i\bullet} = b^{-1} \sum_{j=1}^{b} \hat{\mu}_{ij} = b^{-1} \sum_{j=1}^{b} \bar{y}_{ij\bullet}. \tag{8.2}$$

Note that averages are first taken over repeats, then over the levels of the second factor. Standard errors (SE) must simply take into account the averaging process. For the estimator in (8.2),

$$\mathrm{SE}(\hat{\mu}_{i\bullet}) = \frac{\hat{\sigma}}{b} \sqrt{\sum_{j=1}^{b} n_{ij}^{-1}}. \tag{8.3}$$

Marginal means analyses are appropriate for fractional factorial experiments for which at least one response value is available for all levels of main effects and all combinations of factor levels of interactions included in the model to be analyzed. With balanced designs, these averages equal ordinary averages taken across repeats and the levels of the factors. With unbalanced designs, depending on the numbers of repeats for each factor–level combination, marginal mean averages (referred to as *least squares means* and abbreviated as LSMEANS in SAS) can be substantially different from ordinary averages. In the next section analyses using marginal means will be illustrated on fractions of two-level, factorial experiments.

8.3 ANALYSIS OF DATA FROM TWO-LEVEL, FRACTIONAL FACTORIAL EXPERIMENTS

Depending on the design used, including the inclusion or exclusion of repeat tests, fractional factorial experiments (Chapter 7) can be analyzed in a variety

TABLE 8.3 Factors for Color Test Experiment

Factors		Levels	
Number	Name	Low (−1)	High (+1)
1	Solvent	Low	High
2	Catalyst	0.025	0.035
3	Temperature (°C)	150	160
4	Reactant Purity	92	96
5	Reactant pH	8.0	8.7

of ways. To illustrate some of the most important alternative analyses, this section presents the analyses of data from three fractional factorial experiments. The first example does not contain repeats, the second has an equal number of repeats, and the third has different numbers of repeats for some of the factor–level combinations. All three examples have at least one data value for all combinations of the interaction factor levels included in the respective analyses.

The first experiment was conducted to determine which factors were affecting the color of a product produced by a chemical process. A team of employees identified five factors that were considered to be possible candidates in reducing variation in the product color. These factors and the levels chosen for consideration are given in Table 8.3.

A half-fraction of a 2^5 factorial experiment was selected, but no repeat test runs could be included. The resulting unreplicated R_V completely randomized design is shown in Table 8.4. Because there are no repeat test runs, an estimate of the error standard deviation cannot be obtained when all main effects and two-factor interactions are included in the fitted model. Thus, an analysis-of-variance table cannot be used to determine the significance of the factor effects. With no repeats, one can use a normal quantile plot (Section 5.4) of the effects to help determine which are dominant. Figure 8.1 contains a plot of the fifteen factor effects obtained using the data given in Table 8.4. The effects are calculated using the formulas in Chapter 5. It is clear that the main effect of factor 4 is very large, while the main effects of factors 1, 2, and 5 are moderate in size. All the other main and interaction effects are negligible; thus, future studies should concentrate primarily on the reactant purity and, to a lesser extent, on solvent, catalyst, and reactant pH.

A second example of the use of fractional factorial experiments concerns a study to evaluate the effects of six factors on the corrosion rate of copper. Such a study is important to minimize the potential hazards of copper tubing corrosion in natural gas distribution systems. The six factors and their levels are defined in Table 8.5. Copper temper refers to whether the copper tubing

TABLE 8.4 Responses and Coded Factor Levels for Color Test Experiment

Factors					Color Measurement
Solvent (1)	Catalyst (2)	Temperature (3)	Purity (4)	pH (5)	(coded)
1	−1	1	1	−1	24.76
1	1	−1	1	−1	17.36
1	−1	−1	1	1	18.94
−1	1	−1	−1	−1	0.64
1	1	−1	−1	1	2.68
−1	−1	1	1	1	16.44
1	1	1	−1	−1	9.86
−1	1	1	1	−1	14.60
−1	−1	−1	1	−1	19.58
−1	−1	−1	−1	1	4.74
1	−1	−1	−1	−1	11.02
−1	−1	1	−1	−1	10.12
−1	1	−1	1	1	12.90
−1	1	1	−1	1	1.82
1	1	1	1	1	14.10
1	−1	1	−1	1	8.44

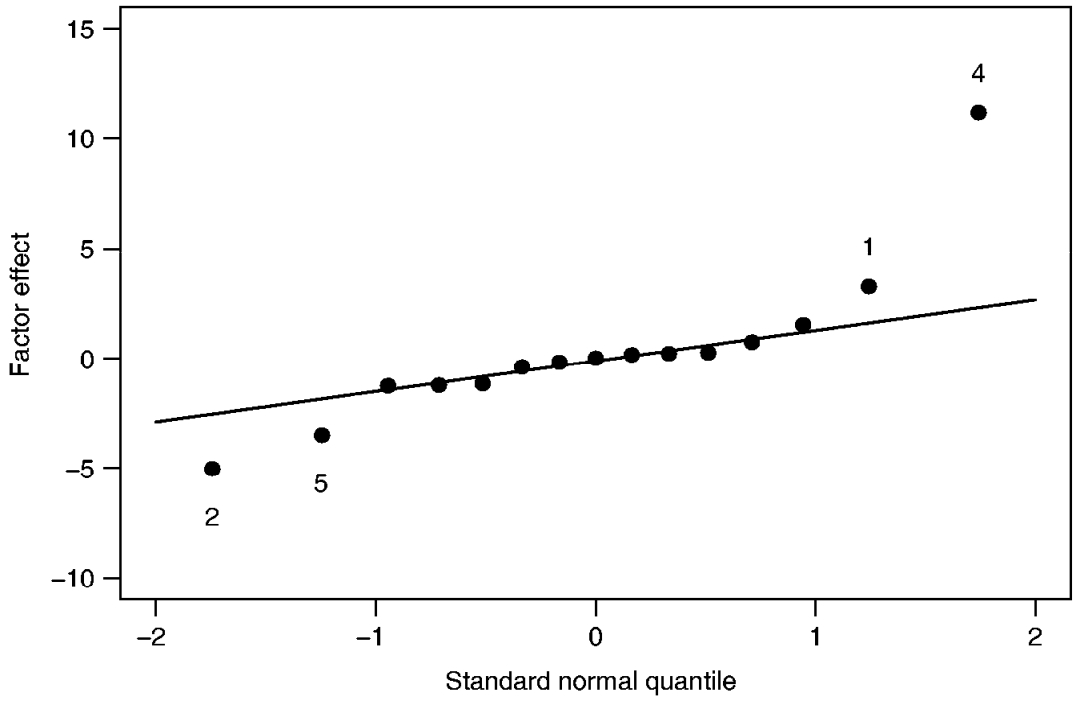


Figure 8.1 Normal quantile plot of color test factor effects.

TABLE 8.5 Factors for the Copper Corrosion Experiment

Factor Number	Factor Name	Coded Levels	
		−1	1
1	Copper temper	Soft (O)	Hard (H)
2	Metal	Base	Silver solder
3	H_2S	16 ppm	55 ppm
4	Mercaptan	20 ppm	58 ppm
5	Oxygen	0.5 vol %	1.5 vol %
6	CO_2	1.5 vol %	4.5 vol %

samples were hard-temper (i.e., drawn) or soft-temper (i.e., annealed). Two types of copper material were used, one being a base metal and the other including sections of base metal joined to couplings using silver solders. Four gaseous contaminants of natural gas were included; some were measured in parts per million (ppm) and the others in volume percent (vol %). Oxygen was added because of its ability to oxidize copper and CO_2 because of its possible influence on copper corrosion. Mercaptan (RHS) is an added odorant to gas and H_2S is a naturally occurring contaminant.

Test specimens, consisting of short pieces of 3/8-inch copper tubing, were drilled with small holes, hung on glass support trees, and placed in 1-liter glass containers. The containers were filled with liquid water saturated with the gaseous contaminants. All tests were run using a constant temperature and pressure, and the exposure time of the copper tubing was set at 5,000 hours. At the conclusion of each experiment, the copper specimens were cleaned and weighed to determine a corrosion rate measured in mils per year (mpy).

Because it was suspected that some two-factor and possibly three-factor interactions might be important, a half fraction of a complete factorial experiment with defining equation $I = ABCDEF$ was selected and run in a completely randomized design. Duplicate tests were run for each combination in the half fraction. This provided a Resolution-VI design with two repeats per factor–level combination.

The analysis-of-variance (ANOVA) model for this experiment is saturated (i.e., all available degrees of freedom for main effects and interactions are used for the included effects; see Exhibit 7.7) if all main effects, all two-factor interactions, and the ten possible (aliased) three-factor interactions are included in the model. An ANOVA table for the saturated model is shown in Table 8.6. The aliased three-factor interactions are shown parenthetically in the table. Because there are an equal number of repeats for each of the highest-order interactions included in the analysis, calculation of the sums of squares by the reduction in error sums of squares method of Section 8.1 and

TABLE 8.6 Analysis of Variance Table for the Copper Corrosion Experiment

Source	df	SS	MS	F	p-Value
Temper (A)	1	1.4862	1.4862	17.18	0.000
Metal (B)	1	0.0234	0.0234	0.27	0.606
H_2S (C)	1	12.9910	12.9910	150.17	0.000
Mercaptan (D)	1	0.0325	0.0325	0.38	0.544
Oxygen (E)	1	15.5135	15.5135	179.33	0.000
CO_2(F)	1	0.8388	0.8388	9.70	0.004
AB	1	0.0220	0.0220	0.25	0.618
AC	1	0.0426	0.0426	0.49	0.488
AD	1	0.5420	0.5420	6.27	0.018
AE	1	0.0065	0.0065	0.07	0.787
AF	1	0.1140	0.1140	1.32	0.260
BC	1	0.2448	0.2448	2.83	0.102
BD	1	0.0030	0.0030	0.04	0.853
BE	1	0.9542	0.9542	11.03	0.002
BF	1	1.4236	1.4236	16.46	0.000
CD	1	0.0502	0.0502	0.58	0.452
CE	1	1.0250	1.0250	11.85	0.002
CF	1	0.6625	0.6625	7.66	0.009
DE	1	0.7377	0.7377	8.53	0.006
DF	1	0.2073	0.2073	2.40	0.132
EF	1	0.2861	0.2861	3.31	0.078
$ABC(+DEF)$	1	0.1575	0.1575	1.82	0.187
$ABD(+CEF)$	1	0.0028	0.0028	0.03	0.859
$ABE(+CDF)$	1	1.3275	1.3275	15.35	0.000
$ABF(+CDE)$	1	0.0050	0.0050	0.06	0.811
$ACD(+BEF)$	1	0.1366	0.1366	1.58	0.218
$ACE(+BDF)$	1	0.3149	0.3149	3.64	0.065
$ACF(+BDE)$	1	0.3761	0.3761	4.35	0.045
$ADE(+BCF)$	1	0.1232	0.1232	1.42	0.241
$ADF(+BCE)$	1	0.0468	0.0468	0.54	0.467
$AEF(+BCD)$	1	0.4222	0.4222	4.88	0.034
Error	32	2.7683	0.0865		
Total	63	42.8877			

the marginal means analysis of Section 8.2 are identical. By assuming that the higher-order interactions that are not included in Table 8.6 are negligible, the F-ratios and corresponding p-values shown in the table can be used to indicate the significant effects. The normal quantile plot shown in Figure 8.2 suggests that the most influential of the significant factor effects are the main effects

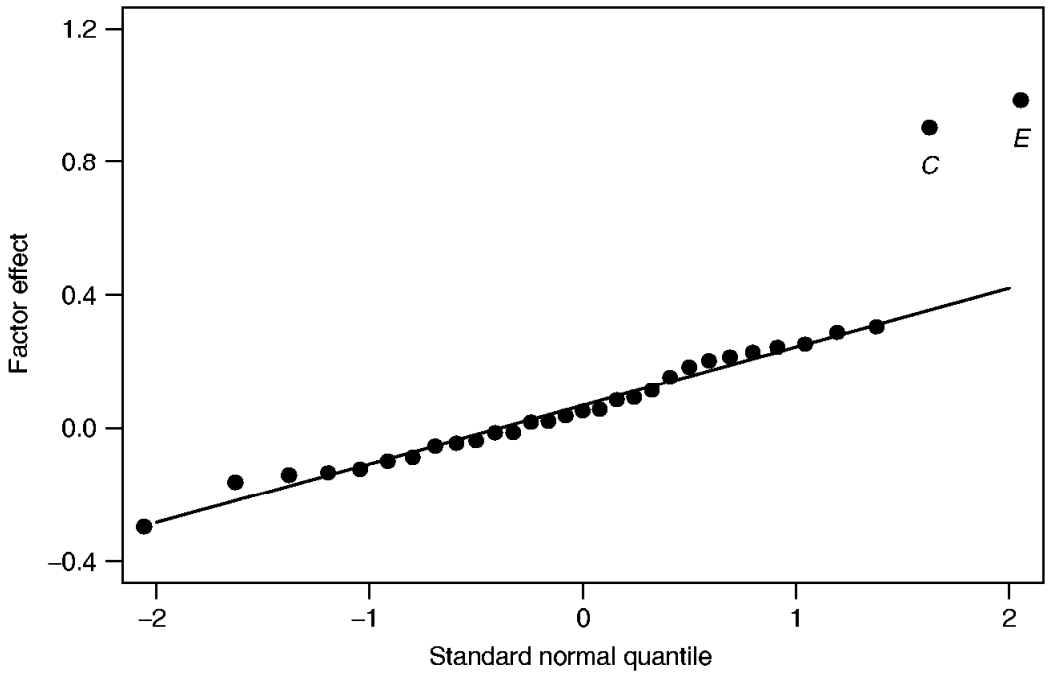


Figure 8.2 Normal quantile plot of corrosion rate factor effects.

for hydrogen sulfide and oxygen. This interpretation is confirmed by the main effects plots in Figure 8.3 and the interaction plots in Figure 8.4. The greatest change in the average corrosion rate occurs when hydrogen sulfide and oxygen are changed from their lower to their higher levels.

The statistically significant three-factor interactions are $ABE(+CDF)$, $ACF(+BDE)$, and $AEF(+BCD)$; however, the inability to separate the three-factor interactions from their aliases hinders their interpretation. One cannot unequivocally state the source of the significance, but one can only assign the effects to the sum of the two aliased interactions in each case. Nevertheless, the significant main effects and two-factor interactions provide extremely valuable information about the importance of these factors on the corrosion rate. Clearly, the corrosion rate of copper depends on the amounts of hydrogen sulfide and oxygen. Increasing the concentration of either gas caused large increases in the copper corrosion rate. In addition, eight of the nine significant interaction effects included at least one of these two factors. Controlling corrosion in natural gas systems requires an ability to control these two contaminants.

The third example of the analysis of a fractional factorial experiment is a variant of the previous example. The same factors, levels, and fractional factorial is used. Rather than repeat all 32 combinations, however, a coin toss was used with each combination to determine whether a repeat would be included. Only 14 of the 32 combinations were selected to be repeated. The analysis of variance table is shown in Table 8.7, with the sums

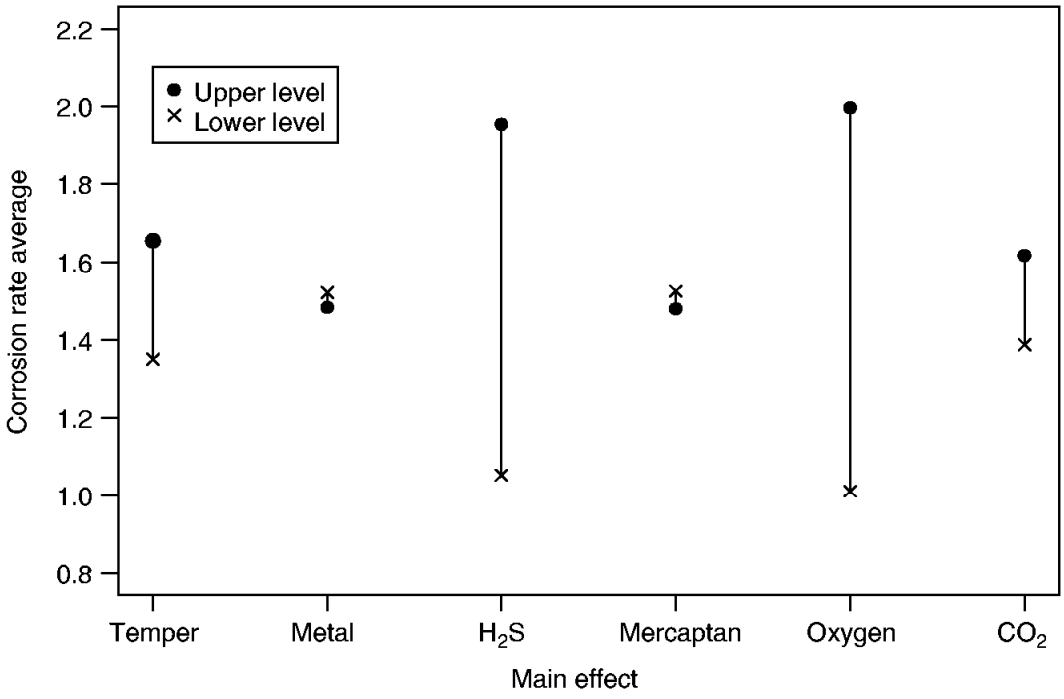


Figure 8.3 Main effects for corrosion rate study.

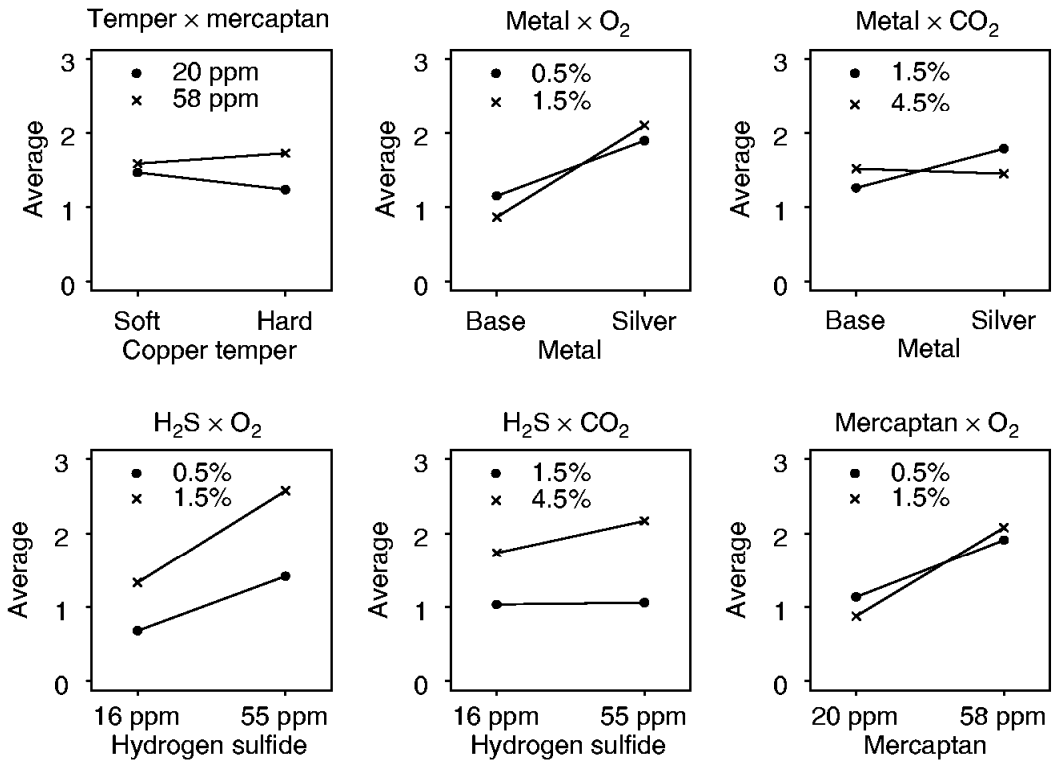


Figure 8.4 Interaction effects for corrosion rate study.

TABLE 8.7 Analysis of Variance Table for the Copper Corrosion Experiment, Unbalanced Experimental Design

Source	df	SS	MS	F	p-Value
Temper (A)	1	0.8425	0.8425	8.09	0.013
Metal (B)	1	0.0304	0.0304	0.29	0.598
H_2S (C)	1	8.0374	8.0374	77.21	0.000
Mercaptan (D)	1	0.1613	0.1613	1.55	0.234
Oxygen (E)	1	10.4384	10.4384	100.27	0.000
CO_2 (F)	1	0.5848	0.5848	5.62	0.033
AB	1	0.0178	0.0178	0.17	0.686
AC	1	0.0187	0.0187	0.18	0.678
AD	1	0.1200	0.1200	1.15	0.301
AE	1	0.0288	0.0288	0.28	0.607
AF	1	0.0457	0.0457	0.44	0.519
BC	1	0.2274	0.2274	2.18	0.162
BD	1	0.0511	0.0511	0.49	0.495
BE	1	0.5435	0.5435	5.22	0.038
BF	1	1.0837	1.0837	10.41	0.006
CD	1	0.0360	0.0360	0.35	0.566
CE	1	0.8214	0.8214	7.89	0.014
CF	1	0.4822	0.4822	4.63	0.049
DE	1	0.4797	0.4797	4.61	0.050
DF	1	0.0189	0.0189	0.18	0.677
EF	1	0.0680	0.0680	0.65	0.433
ABC (+ DEF)	1	0.1633	0.1633	1.57	0.231
ABD (+ CEF)	1	0.0393	0.0393	0.38	0.549
ABE (+ CDF)	1	0.5684	0.5684	5.46	0.035
ABF (+ CDE)	1	0.0016	0.0016	0.01	0.905
ACD (+ BEF)	1	0.0029	0.0029	0.03	0.869
ACE (+ BDF)	1	0.0556	0.0556	0.53	0.477
ACF (+ BDE)	1	0.3151	0.3151	3.03	0.104
ADE (+ BCF)	1	0.1825	0.1825	1.75	0.207
ADF (+ BCE)	1	0.2140	0.2140	2.06	0.174
AEF (+ BCD)	1	0.1031	0.1031	0.99	0.337
Error	14	1.4575	0.1041		
Total	45				

of squares calculated using the analysis of marginal means method described in Section 8.2.

The dominant main effects of hydrogen sulfide and oxygen that were noted in Table 8.6 and the subsequent analyses are again apparent from the p-values in Table 8.7 and the main effects and interaction plots in Figures 8.5 and 8.6.

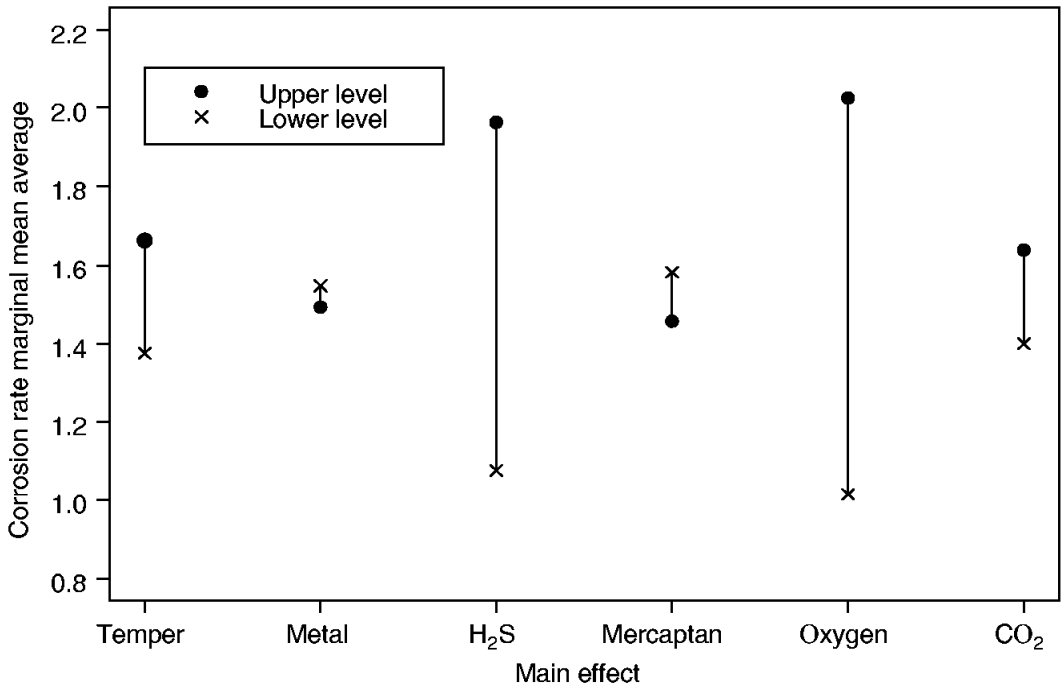


Figure 8.5 Main effects for corrosion rate study, unbalanced data set.

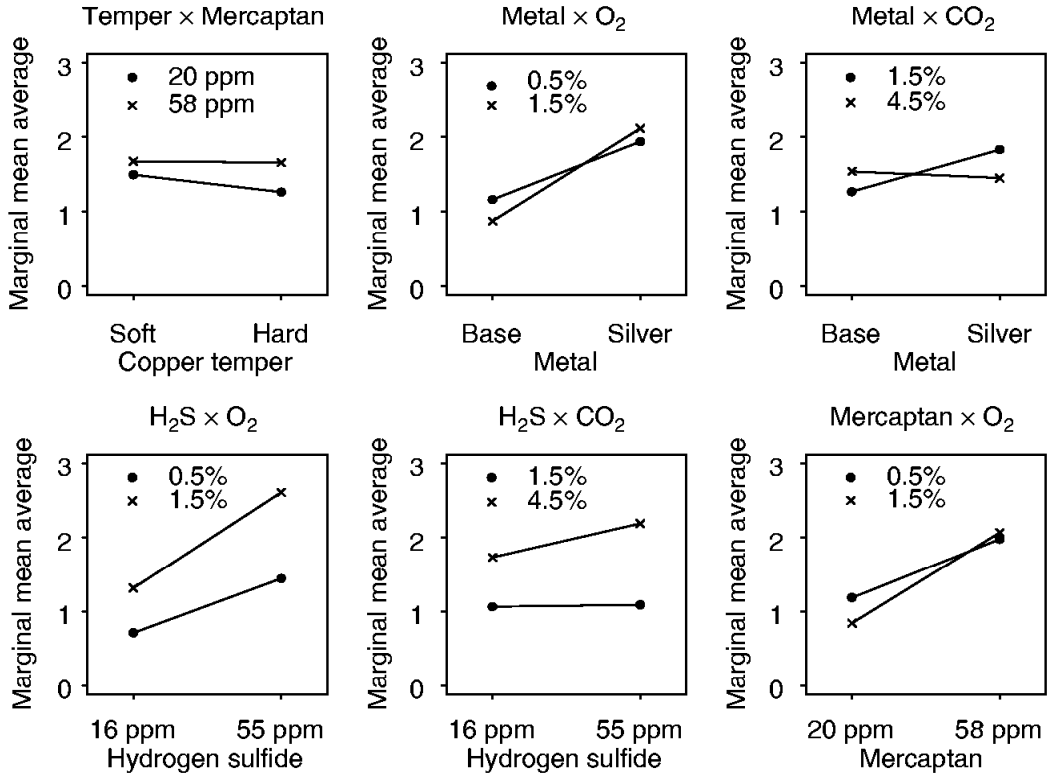


Figure 8.6 Interaction effects for corrosion rate study, unbalanced data set.

The only notable differences in Tables 8.6 and 8.7 are the lack of significant interactions *AD, ABE, ACF,* and *AEF* in Table 8.7. It is clear from the marginal mean averages (least squares means) plotted in Figures 8.5 and 8.6 that the lack of significance of the above four interactions does not materially affect conclusions that would be drawn from comparisons of the key interaction and main effect averages. The consistency of conclusions between analyses of data from balanced and unbalanced experimental designs like these two versions of the corrosion rate study is characteristic of statistically designed complete and fractional factorial experiments.

8.4 ANALYSIS OF DATA FROM THREE-LEVEL, FRACTIONAL FACTORIAL EXPERIMENTS

Analysis of data from three-level, fractional factorial experiments differs from the corresponding analyses for two-level, fractional factorials primarily in that aliasing patterns are based on main effects that have two degrees of freedom rather than one degree of freedom as with two-level fractional factorials. To clarify the similarities and the differences between the two types of analyses, consider an experiment (Ostle and Malone 1988) that was conducted to study the effects of three factors on the density of small bricks. The factors and their levels are listed in Table 8.8, and the corresponding test runs and data are given in Table 8.9. Each factor–level combination was run in duplicate so the total number of test runs is 54.

To illustrate the analysis of a three-level, fractional factorial experiment, a portion of the data in Table 8.9 will be analyzed. Suppose this experiment had been run using only the nine factor–level combinations listed in Table 8.10. The experiment would then be a 3^{3-1} (i.e., one-third) fractional factorial experiment of Resolution III. The defining relationship and aliasing pattern is:

$$I = AB^2\ C^2$$

$$A = ABC$$

$$A^2 = BC$$

$$B = AC^2$$

$$B^2 = ABC^2$$

$$C = AB^2$$

$$C^2 = AB^2C$$

$$AB = AC$$

$$A^2B^2 = BC^2$$

TABLE 8.8 Factor Levels for the Brick Density Experiment

Factors		Levels		
Number	Name	Low (−1)	Medium (0)	High (+1)
1	Particle Size	5–10	10–15	15–20
2	Pressure	5	12.5	20
3	Temperature	1,900	2,000	2,100

TABLE 8.9 Responses and Coded Factor Levels for Brick Density Experiment

Combination Number	Particle Size	Pressure	Temperature	Brick Density	
1	−1	−1	−1	340	375
2	−1	−1	0	316	386
3	−1	−1	1	374	350
4	−1	0	−1	388	370
5	−1	0	0	338	214
6	−1	0	1	334	366
7	−1	1	−1	378	378
8	−1	1	0	348	378
9	−1	1	1	380	398
10	0	−1	−1	260	244
11	0	−1	0	388	304
12	0	−1	1	266	234
13	0	0	−1	322	342
14	0	0	0	300	420
15	0	0	1	234	258
16	0	1	−1	330	298
17	0	1	0	260	366
18	0	1	1	350	284
19	1	−1	−1	134	140
20	1	−1	0	146	194
21	1	−1	1	152	212
22	1	0	−1	186	30
23	1	0	0	412	428
24	1	0	1	194	208
25	1	1	−1	40	210
26	1	1	0	436	490
27	1	1	1	230	254

Thus, the experiment is able to estimate the main effects due to the three factors and one of the two-way interactions, provided all other interactions are assumed to be negligible. Abstracting these particular runs from Table 8.9 yields the eighteen observations listed in Table 8.10. Note that the duplicate runs have been retained to provide an estimate of the experimental error standard deviation.

An ANOVA table for the data in this fractional factorial experiment is given in Table 8.11. As with the second example of the previous section, because the factor–level combinations included in the design have the same number

TABLE 8.10 Responses and Coded Factor Levels for a One-Third Fraction of the Brick Density Experiment

Combination Number	Particle Size	Pressure	Temperature	Brick Density	
1	−1	−1	−1	340	375
6	−1	0	1	334	366
8	−1	1	0	348	378
11	0	−1	0	388	304
13	0	0	−1	322	342
18	0	1	1	350	284
21	1	−1	1	152	212
23	1	0	0	412	428
25	1	1	−1	40	210

TABLE 8.11 ANOVA Table for Brick Density Data

Source	df	SS	MS	F	p-Value
Pressure	2	43,448	21,724	8.19	0.009
Linear	1	39,331	39,331	14.84	0.004
Quadratic	1	4,117	4,117	1.55	0.244
Particle Size	2	31,458	15,729	5.93	0.023
Linear	1	2,160	2,160	0.81	0.390
Quadratic	1	29,298	29,298	11.05	0.009
Temperature	2	39,667	19,834	7.48	0.012
Linear	1	397	397	0.15	0.708
Quadratic	1	39,270	39,270	14.81	0.004
Press × Size	2	27,832	13,916	5.25	0.031
Error	9	23,858	2,651		
Total	17	166,264			

TABLE 8.12 p-values for the Complete Brick Density Factorial Experiment

Effect	Pressure (A)	Size (B)	Temperature (C)	AB	AC	BC	ABC
p-Value	0.008	0.000	0.000	0.077	0.346	0.000	0.005

of repeats, the reduction in error sums of squares calculations and the analysis of marginal means calculations result in the same sums of squares, those shown in Table 8.11. The aliasing pattern indicated that once each main effect was included, only two of the degrees of freedom for the AB interaction remained unaliased with main effects. Hence, the only interaction included in the analysis of variance table is the AB (pressure $\times$ size and its aliases) interaction.

Observe that each of the three main effects and the two-factor interaction has two degrees of freedom, and each is significant at a 0.05 significance level. Because each factor is quantitative and consists of two degrees of freedom, it is informative to decompose each main effect into orthogonal linear and quadratic effects (see Section 6.2.3). As indicated in the ANOVA table, the significant quantitative effects include the linear effect of pressure and the quadratic effects of temperature and size.

There also is a significant two-factor interaction between pressure and size. An interaction plot for this set of factor–level combinations reveals that density is a nonlinear function of pressure for each particle size. The interaction occurs because the nonlinearity with pressure is different for each of the three particle sizes.

This example also illustrates the importance of clearly appreciating the effects of aliasing in fractional factorial experiments. Because the complete 3^3 factorial data with replicates is given in Table 8.9, one can compare the results of the analysis from the fractionated experiment to that of the complete factorial experiment. Table 8.12 contains p-values for the complete experiment. It is clear that the three main effects remain significant, as they were with the fractional factorial. However, in the complete experiment the BC (size $\times$ temperature) interaction, not the AB (pressure $\times$ size) interaction, is statistically significant. The apparent significant pressure $\times$ size interaction in Table 8.6 is likely because of its aliasing with the size $\times$ temperature interaction and not because of the joint effects of pressure and size.

8.5 ANALYSIS OF FRACTIONAL FACTORIAL EXPERIMENTS WITH COMBINATIONS OF FACTORS HAVING TWO AND THREE LEVELS

In many industrial settings, experiments are designed that consist of combinations of two-level and three-level factors. These designs can be impractical

because of their potentially large sizes. In such cases, fractionating the design can reduce the experimental size and often improve its efficiency. A general outline of how such fractionating can be accomplished was discussed in Section 7.5. This section describes how analyses of these designs can be performed.

An experiment was conducted to study the effects of four properties of diesel fuels on the nitrous oxide (NO_x) emissions from a heavy-duty diesel engine (Matheaus, et al. 1999). The factors and their levels are listed in Table 8.13. The delta cetane factor has three levels, each level indicating the increase in the total cetane number of the fuel from a base value of 42. The increase results from adding an ignition-improver additive to a base fuel. The other three factors each have two levels, and they measure the density and aromatic content of the fuels.

If a complete factorial experiment were used to study these factors, a total of $3 \times 2^3 = 24$ different diesel fuels would need to be blended and then tested in a laboratory on an engine dynamometer. Because such tests are very costly to run, it was decided to fractionate the experiment. Each level of delta cetane was crossed with a one-half fraction of the other three factors. This reduced the size of the experiment to $3 \times 2^{3-1} = 12$ combinations of the four factors. The fractionating removed the ability to evaluate all interactions except the two-factor interactions of each of the two-level factors with delta cetane.

The target factor levels of the diesel fuels in the design are given in Table 8.14. Duplicate test runs were made on each fuel and the resulting NO_x values were measured and recorded. Note that although the number of test runs for this fractional factorial experiment equals the number of test runs in the unreplicated complete factorial experiment, substantial savings in costs are achieved because repeat test runs are much less expensive than single-test runs on different blends because of the cost of blending the fuels. An 8-mode test procedure was utilized for these tests; consequently, the measured NO_x emissions from the various modes were combined to form a single weighted value, labeled $Wt(NO_x)$.

TABLE 8.13 Factor Levels for Diesel Fuel NO_x Experiment

Factors		Levels		
Number	Name	Low (−1)	Medium (0)	High (+1)
1	Delta cetane	0	5	10
2	Density (kg/m^3)	830		860
3	Monoaromatic content (%m)	10		25
4	Polyaromatic content (%m)	3		11

TABLE 8.14 Responses and Targeted Factor Levels for the Diesel Fuel NO_x Experiment

Combination	Delta Cetane	Density	Monoaromatics	Polyaromatics	Wt(NO_x)
1	0	830	10	11	2.338
2	5	830	10	11	2.383
3	10	830	10	11	2.373
4	0	860	10	3	2.446
5	5	860	10	3	2.441
6	10	860	10	3	2.498
7	0	830	25	3	2.423
8	5	830	25	3	2.408
9	10	830	25	3	2.455
10	0	860	25	11	2.643
11	5	860	25	11	2.634
12	10	860	25	11	2.659
1	0	830	10	11	2.401
2	5	830	10	11	2.399
3	10	830	10	11	2.443
4	0	860	10	3	2.448
5	5	860	10	3	2.495
6	10	860	10	3	2.462
7	0	830	25	3	2.440
8	5	830	25	3	2.411
9	10	830	25	3	2.507
10	0	860	25	11	2.594
11	5	860	25	11	2.618
12	10	860	25	11	2.626

TABLE 8.15 ANOVA Table for the Diesel Fuel NO_x Experiment

Source	df	SS	MS	F	p-Value
Cetane (A)	2	0.0059	0.0030	3.53	0.062
Density (B)	1	0.1044	0.1044	124.74	0.000
Monoaromatics (C)	1	0.0694	0.0694	82.96	0.000
Polyaromatics (D)	1	0.0191	0.0191	22.81	0.001
AB	2	0.0009	0.0005	0.54	0.598
AC	2	0.0011	0.0006	0.67	0.530
AD	2	0.0006	0.0003	0.38	0.693
Error	12	0.0100	0.0008		
Total	23	0.2116			

An ANOVA table was constructed for the diesel fuel NO_x experiment. The results are summarized in Table 8.15. By assuming that the higher-order interactions are negligible, we can use the F-ratios to indicate the significant effects at the 0.05 significance level. The main effects of density and both aromatics are statistically significant. None of the two-factor interactions are significant. A main effects plot confirms that the most influential of the significant factors is density, followed by monoaromatics and polyaromatics.

8.6 ANALYSIS OF SCREENING EXPERIMENTS

In Sections 7.4 and 7.6, different types of screening designs were introduced. These included classical fractional factorials and special fractions; for example, both saturated and supersaturated designs. These designs are intended for the identification of a small number of active factors, often with the intent of conducting further experimentation using the active factors. In this section, we present the analysis of three of these types of fractional factorial experiments. The first example is a screening experiment that uses one of the main-effects fractional factorials popularized by Plackett and Burman.

Certain types of plated materials (for example, a connector in an electronic assembly) are soldered for strength. The manufacturer of the plated material tests it for solderability prior to releasing it for shipment to the customer. In an experiment conducted to assess the effects of various factors on the soldering process, the written procedure for the solderability test method was reviewed in detail to identify variables that might affect the response — percentage solder coverage. Ten factors were identified and are listed, along with the levels chosen for consideration, in Table 8.16. A sixteen-run (randomized) Plackett–Burman screening design was used, with the results shown in Table 8.17.

Each of the factors listed in Table 8.16 consists of fixed effects; that is, these are the only factor levels of interest in the experiment. A fixed-effects model containing only main effects was fitted, and the resulting ANOVA table is shown in Table 8.18. F-ratios are formed by comparing each of the factor-effect mean squares with the error mean square.

Several of the factors in Table 8.18 are statistically significant. One's philosophy in analyzing screening designs often is to allow a larger significance level in testing hypotheses in order to (a) select several potentially important factors that will be studied further in future experiments, or (b) identify factors in a process that must be more tightly controlled so a more uniform product will result. If one uses a 10% significance level in Table 8.18, more consistent solderability test results can be obtained by better control of the surface areas, dip device, flux time, drain time, and solder time.

A second example of the use of screening designs in a ruggedness test concerns a laboratory viscosity method. Seven potentially sensitive steps in the viscosity method are shown in Table 8.19 with the levels chosen for

TABLE 8.16 Factors for Solderability Ruggedness Test

Factor		Levels	
Number	Name	−1	+1
1	Solvent dip	No	Yes
2	Surface area	Small	Large
3	Dip device	Manual	Mechanical
4	Magnification	10×	30×
5	Solder age	Fresh	Used
6	Flux time	8 sec	30 sec
7	Drain time	10 sec	60 sec
8	Stir	No	Yes
9	Solder time	2 sec	8 sec
10	Residual flux	Not clean	Clean

TABLE 8.17 Responses and Coded Factor Levels for Solder-Coverage Ruggedness Testing Data

Solder	Level									
Coverage (%)	Factor 1	2	3	4	5	6	7	8	9	10
91	1	1	1	1	−1	1	−1	1	1	−1
97	1	−1	1	−1	1	1	−1	−1	1	−1
89	1	−1	1	1	−1	−1	1	−1	−1	−1
82	1	−1	−1	−1	1	1	1	1	−1	1
82	1	1	1	−1	1	−1	1	1	−1	−1
74	1	−1	−1	1	−1	−1	−1	1	1	1
54	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
66	−1	1	1	−1	−1	1	−1	−1	−1	1
79	−1	1	−1	−1	−1	1	1	1	1	−1
25	−1	1	−1	1	1	−1	−1	1	−1	−1
77	−1	−1	−1	1	1	1	1	−1	1	−1
44	1	1	−1	−1	1	−1	−1	−1	1	1
86	−1	−1	1	1	1	1	−1	1	−1	1
97	−1	−1	1	−1	−1	−1	1	1	1	1
84	1	1	−1	1	−1	1	1	−1	−1	1
97	−1	1	1	1	1	−1	1	−1	1	1

TABLE 8.18 ANOVA Table for Solder-Coverage Ruggedness Tests

Source	df	SS	MS	F	p-Value
Solvent dip	1	240.25	240.25	2.25	0.194
Surface area	1	484.00	484.00	4.53	0.087
Dip device	1	2162.25	2162.25	20.25	0.007
Magnification	1	30.25	30.25	0.28	0.619
Solder age	1	121.00	121.00	1.13	0.336
Flux time	1	625.00	625.00	5.85	0.061
Drain time	1	1406.25	1406.25	13.17	0.015
Stir	1	4.00	4.00	0.04	0.849
Solder time	1	484.00	484.00	4.53	0.087
Residual flux	1	81.00	81.00	0.76	0.423
Error	5	534.00	106.80		
Total	15	6172.00			

TABLE 8.19 Factors for Viscosity Ruggedness Tests

Factor		Levels	
No.	Name	−1	+1
1	Sample preparation	Method 1	Method 2
2	Moisture measure	Volume	Weight
3	Speed	800	1600
4	Mixing time	0.5	3.0
5	Equilibrium time	1	2
6	Spindle	Type 1	Type 2
7	Lid	Absent	Present

investigation. Sixteen experiments, corresponding to a one-eighth fraction of a complete factorial experiment, were run. The factor–level combinations, along with the viscosity responses, are shown in Table 8.20. Because this completely randomized design is of Resolution IV (see Table 7A.1), the main effects are not confounded with any two-factor interactions.

The ANOVA table for a main-effects model fitted to these data is given in Table 8.21. Again assuming negligible interactions, the appropriate F-ratios are obtained by comparing the factor-effect mean squares with the error mean square. Speed, mixing time, and type of spindle are statistically significant effects.

In each of the above analyses the assumption of negligible interaction effects is critical. In both analyses the presence of interaction effects would

TABLE 8.20 Viscosity Data for Ruggedness Tests

	Factor Level						
Viscosity	Factor 1	2	3	4	5	6	7
2220	−1	−1	−1	1	1	1	−1
2460	1	−1	−1	−1	−1	1	1
2904	−1	1	−1	−1	1	−1	1
2464	1	1	−1	1	−1	−1	−1
3216	−1	−1	1	1	−1	−1	1
3772	1	−1	1	−1	1	−1	−1
2420	−1	1	1	−1	−1	1	−1
2340	1	1	1	1	1	1	1
3376	1	1	1	−1	−1	−1	1
3196	−1	1	1	1	1	−1	−1
2380	1	−1	1	1	−1	1	−1
2800	−1	−1	1	−1	1	1	1
2320	1	1	−1	−1	1	1	−1
2080	−1	1	−1	1	−1	1	1
2548	1	−1	−1	1	1	−1	1
2796	−1	−1	−1	−1	−1	−1	−1

TABLE 8.21 ANOVA Table for Viscosity Ruggedness Tests

Source	df	SS	MS	F	p-Value
Sample preparation	1	49	49	0.00	0.973
Moisture measure	1	74,529	74,529	1.91	0.204
Speed	1	859,329	859,329	21.99	0.002
Mixing time	1	361,201	361,201	9.24	0.016
Equilibrium time	1	51,529	51,529	1.32	0.284
Spindle	1	1,723,969	1,723,969	44.12	0.000
Lid	1	1,521	1,521	0.04	0.846
Error	8	312,608	39,076		
Total	15	3,384,735			

inflate the estimate of experimental error, MS_E. Likewise, the presence of interaction effects would bias the factor effects and could either inflate or reduce the size of the individual factor-effect sums of squares, depending on the nature of the biases. With the fractional factorial experiment for the viscosity ruggedness tests, only the three-factor or higher interaction effects would

bias the main effects. Using these designs for screening purposes, however, indicates that interest is only in very large effects of the factors, effects that are expected to be detectable with main-effects models.

A third example of the use of screening designs is an application of a supersaturated design (Section 7.6.2). Prior to discussing this example, we again caution against routine use of supersaturated designs. The potential for severe bias in these designs is very real and can be far more serious than the misinterpretation of the AB interaction in the fractional factorial for the brick density example in Section 8.4.

A 12-run Plackett-Burman design was used by Hunter, Hodi, and Eager (1982) to study the effects of seven factors on the fatigue life of weld-repaired castings. These factors and their levels are listed in Table 8.22, and the corresponding test runs and data are given in Table 8.23. Note that the first seven columns of the design matrix were assigned to the seven factors.

As discussed in Section 7.6, one can create a supersaturated design from this design by splitting it into two fractions. Using this approach, suppose the twelve Plackett–Burman test runs listed in Table 8.23 are split by choosing column 8 as the branching column and selecting all test runs with a +1 in that column. With this choice, the test runs, which are boxed in the table, would include only rows 3, 5, 6, 8, 9, and 10, as well as columns 1 to 7. Thus, a six-run supersaturated design for the study of seven factors has been created. Any of the other unused columns could have been selected to split the Plackett–Burman design, and doing so would probably produce slightly different results.

In the absence of repeat runs, a normal quantile plot of the effects can be used to identify active effects. The dominant factor in the effects plot is factor 6 (polish), with factors 4 (heat treat) and 7 (final treat) being less dominant. This result agrees with the findings from the original analysis of the complete 12-run data set shown in Table 8.23. When analyzed, the factors identified as

TABLE 8.22 Factors for Cast Fatigue Experiment

Factors		Levels	
Number	Name	Low (−1)	High (+1)
1	Initial structure	Beta treatment	As received
2	Bead size	Large	Small
3	Pressure treatment	Hot isostatic pressing	None
4	Heat treatment	Solution treatment	Anneal
5	Cooling rate	Rapid	Slow
6	Polish	Mechanical	Chemical
7	Final treatment	Peen	None

TABLE 8.23 Responses and Coded Factor Levels for Cast Fatigue Experiment

Log Fatigue	Factor Number										
	1	2	3	4	5	6	7	8	9	10	11
6.058	1	1	−1	1	1	1	−1	−1	−1	1	−1
4.733	1	−1	1	1	1	−1	−1	−1	1	−1	1
4.625	−1	1	1	1	−1	−1	−1	1	−1	1	1
5.899	1	1	1	−1	−1	−1	1	−1	1	1	−1
7.000	1	1	−1	−1	−1	1	−1	1	1	−1	1
5.752	1	−1	−1	−1	1	−1	1	1	−1	1	1
5.682	−1	−1	−1	1	−1	1	1	−1	1	1	1
6.607	−1	−1	1	−1	1	1	−1	1	1	1	−1
5.818	−1	1	−1	1	1	−1	1	1	1	−1	−1
5.917	1	−1	1	1	−1	1	1	1	−1	−1	−1
5.863	−1	1	1	−1	1	1	1	−1	−1	−1	1
4.809	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1

significant included factor 6, and, to a lesser degree, factor 4. It is encouraging to see that this supersaturated design obtained similar conclusions about the main effects as the original screening design. Hamada and Wu (1992) also analyzed the complete data set but used a half-normal quantile plot. In their analysis, they concluded that factor 6 was significant. However, these authors also considered all the two-way interactions terms involving factor 6, and found that the polish × final treatment interaction was significant and that factor 4, in the presence of this interaction, became nonsignificant.

REFERENCES

Text References

The following references discuss the analysis of a wide variety of designs for fractional factorial experiments. Other useful references appear at the end of Chapters 4, 6, and 7.

Federer, W. T. (1955). *Experimental Design*, New York: Macmillan Co. *Many examples of split-plot, crossover, and similar designs.*

Kirk, R. E. (1968). *Experimental Design: Procedures for the Behavioral Sciences*, Belmont, CA: Brooks/Cole Publishing Co.

Ostle, B. and Malone, L. C. (1988). *Statistics in Research*, Fourth Edition, Ames, IA: Iowa State University Press.

SAS Institute, Inc (2000). SAS Online Doc, Version 8, Cary, NC.

Searle, S. R., Speed, F. M., and Milliken, G. A. (1980). "Population Marginal Means in the Linear Model: An Alternative to Least Squares Means," *The American Statistician*, **34**, 216–221.

Data References

The data for the analyses reported in Sections 8.5 and 8.6 were taken from

Hamada, M. and Wu, C. F. J. (1992). "Analysis of Designed Experiments with Complex Aliasing," *Journal of Quality Technology*, **24**, 130–137.

Hunter, G. B., Hodi, F. S., and Eager, T. W. (1982). "High-Cycle Fatigue of Weld Repair Cast Ti-6A1-4V," *Metallurgical Transactions*, **13A**, 1589–1594.

Matheaus, A. C., Ryan, T. W., Mason, R. L., Neeley, G., and Sobotowski, R. (1999). "Gaseous Emissions from a Caterpillar 3176 (with EGR) Using a Matrix of Diesel Fuels (Phase 2)," EPA Contract No. 68-C-98-169. San Antonio, TX: Southwest Research Institute. (Data reported to two decimal places; the analysis in the text used data reported to three decimal places.)

EXERCISES

1 Use the data from the solderability ruggedness test in Tables 8.17 and 8.18 to determine which factor levels in Table 8.16 should be used to provide a high percentage of solder coverage. Just as importantly, which factors do not seem to affect the coverage? Why is this latter finding important?

2 Answer the same questions as in Exercise 1 for the viscosity ruggedness test data presented in Tables 8.19 and 8.20.

3 Six factors, each with two levels, were studied in a screening experiment designed to determine the uniformity of a new etching tool used in the etching process of wafers in a semiconductor industry. A measure of etch uniformity is measured at several sites on a wafer. The coded data from a Resolution-IV, quarter-fraction design are given below, where smaller measurement values represent greater uniformity. Four repeats runs were made in order to obtain an estimate of the error variation.

(a) Construct an analysis of variance table for this experiment and determine the significant factor effects (use a 0.05 significance level).

(b) Construct a normal quantile plot of the factor effects and verify your results from (a).

(c) What assumptions were used in answering part (a)?

(d) What recommendations would you make to the experimenters at the conclusion of this analysis?

Run	Factors						Coded Etch Uniformity
	A	*B*	*C*	*D*	*E*	*F*	
1	−1	−1	−1	−1	−1	−1	4.31
2	−1	1	−1	−1	1	1	4.02
3	−1	−1	1	−1	1	1	3.88
4	−1	1	1	−1	−1	−1	3.94
5	−1	−1	−1	1	−1	1	2.88
6	−1	1	−1	1	1	−1	3.31
7	−1	−1	1	1	1	−1	3.48
8	−1	1	1	1	−1	1	2.78
9	1	−1	−1	−1	1	−1	2.85
10	1	1	−1	−1	−1	1	3.00
11	1	−1	1	−1	−1	1	2.98
12	1	1	1	−1	1	−1	2.75
13	1	−1	−1	1	1	1	1.82
14	1	1	−1	1	−1	−1	4.42
15	1	−1	1	1	−1	−1	5.98
16	1	1	1	1	1	1	2.35
6	−1	1	−1	1	1	−1	2.08
15	1	−1	1	1	−1	−1	5.02
3	−1	−1	1	−1	1	1	4.12
13	1	−1	−1	1	1	1	2.99

4 A Plackett–Burman screening experiment was run to determine the effects of seven factors associated with a proposed measurement system on the flash point of kerosene. Flash point is defined as the lowest temperature at which application of a test flame causes the kerosene to ignite. Duplicate tests were run to obtain an adequate measure of the error variation.

(a) Construct an ANOVA table for this experiment and determine the significant factor effects (use a 0.05 significance level).

(b) Construct a normal quantile plot of the factor effects and verify your results from (a).

(c) What assumptions were used in answering part (a)?

(d) What conclusions would you make to the experimenters concerning the effects of the factors on the flash point of kerosene?

Run	Factors							Flash Point (°C)
	A	B	C	D	E	F	G	
1	1	1	1	−1	1	−1	−1	52
2	−1	1	1	1	−1	1	−1	50
3	−1	−1	1	1	1	−1	1	51
4	1	−1	−1	1	1	1	−1	46
5	−1	1	−1	−1	1	1	1	49
6	1	−1	1	−1	−1	1	1	51
7	1	1	−1	1	−1	−1	1	46
8	−1	−1	−1	−1	−1	−1	−1	50
9	1	1	1	−1	1	−1	−1	53
10	−1	1	1	1	−1	1	−1	44
11	−1	−1	1	1	1	−1	1	51
12	1	−1	−1	1	1	1	−1	45
13	−1	1	−1	−1	1	1	1	50
14	1	−1	1	−1	−1	1	1	52
15	1	1	−1	1	−1	−1	1	47
16	−1	−1	−1	−1	−1	−1	−1	51

5 A 2^{5-2} fractional factorial experiment was run to evaluate the effects of five factors on the production of wire harnesses. The response variable of interest is a coded variable representing the number of harnesses produced per hour in a given workday. The production process is automated and uses a series of robotic cells. The five factors, where MTBF denotes the mean time between failures, include the following: unit cycle time (A), minimum MTBF (B), robot cycle time (C), robot MTBF (D), and number of robots (E). The resulting responses and coded factor levels are given below, and are based on use of a defining contrast given by $I = ABD = ACE$. Because no repeat test runs were available in this experiment, use a normal quantile plot to identify the active factors.

Run	A	B	C	D	E	Response
1	−1	1	1	−1	−1	2.481
2	−1	−1	1	1	−1	1.730
3	1	1	−1	1	−1	1.875
4	1	−1	−1	−1	−1	1.939
5	−1	−1	−1	1	1	1.708
6	−1	1	−1	−1	1	1.491
7	1	−1	1	−1	1	2.593
8	1	1	1	1	1	2.195

6 Suppose three repeat tests were run for the design given in Exercise 5. Test runs 1, 4, and 7 were repeated, with the resulting values of the response variable being 2.513, 1.961, and 2.570, respectively. Re-analyze the data and determine the factor effects on the coded response variable.

7 Nine factors are studied in an 18-run screening design including two centerpoints inserted at the midpoint of the design. Each factor has two levels, and it assumed that all interactions are negligible. Analyze the data from this experiment and determine which factors influence the response (use a 0.05 significance level). Obtain an estimate of the error variation by assuming that the interactions among the factors are negligible. Analyze and interpret the results of the experiment.

Run	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	Response
1	−1	−1	−1	−1	−1	−1	−1	−1	−1	13.5
2	1	1	1	−1	−1	1	1	−1	−1	3.1
3	−1	−1	1	−1	1	−1	1	−1	1	12.7
4	1	1	−1	−1	1	1	−1	−1	1	10.1
5	0	0	−1	0	0	0	0	−1	0	11.3
6	1	−1	1	1	−1	1	−1	−1	1	7.6
7	−1	1	−1	1	−1	−1	1	−1	1	5.9
8	1	1	1	1	1	−1	−1	−1	−1	14.2
9	−1	−1	−1	1	1	1	1	−1	−1	6.5
10	1	1	1	−1	−1	−1	−1	1	1	14.1
11	−1	−1	−1	−1	−1	1	1	1	1	6.0
12	1	−1	−1	−1	1	−1	1	1	−1	17.8
13	−1	1	1	−1	1	1	−1	1	−1	22.6
14	0	0	1	0	0	0	0	1	0	14.7
15	1	1	−1	1	−1	1	−1	1	−1	10.1
16	−1	−1	1	1	−1	−1	1	1	−1	13.6
17	1	1	1	1	1	−1	1	1	1	5.8
18	−1	−1	−1	1	1	1	−1	1	1	19.6

8 Consider the two-factor experiment described in Exercise 6 in Chapter 6 on ballistic limit velocity. Suppose the three observations for the factor combination consisting of the conical nose shape and the 45° firing angle was not available.

(a) Re-analyze the data to determine the effects of nose shape and firing angle on the ballistic velocities, and state your conclusions.

(b) Show that the marginal mean averages associated with the nose shape factor differ from the ordinary averages.

9 Consider the three-factor experiment described in Exercise 20 in Chapter 6 on the fragmentation of an explosive device. Suppose the design had been unbalanced and the data given below were obtained. Analyze these data, and perform appropriate tests on the main effects and interactions.

		Pipe Diameter			
		0.75 in.		1.75 in.	
	Material:	Sand	Earth	Sand	Earth
Air Gap	0.50 in.	0.0698	0.0659	0.0625	0.0699
		—	—	0.0615	0.0620
		—	0.0676	0.0619	—
	0.75 in.	—	—	—	—
		0.0613	0.0635	0.0601	0.0612
		0.0620	—	—	0.0594

10 In Exercise 9, delete the observation for air gap = 0.50, material = sand, and pipe diameter = 0.75. Reanalyze the data and compare your results with those obtained in Exercise 9. State your conclusions at the 0.05 significance level.

11 Fuel consumption data were collected on an engine as four factors were varied: fuel economy improvement device (A), engine speed (B), engine load (C), and engine timing (D). There were three different devices to be compared, and each of the three engine factors was tested using three levels. Because of cost constraints, the chosen design was a 3^{4-2} fractional factorial experiment. Each of the nine test combinations was run in duplicate. Analyze the data, using a 0.05 significance level, and determine which factors influence fuel consumption. If $A1$ denotes the device tested at the low level, $A2$ denotes the device tested at the middle level, and $A3$ denotes the device tested at the high level of factor A, which device was best in terms of providing the lowest fuel consumption? What assumptions were used in the analysis of these data?

Test	A	B	C	D	Coded Fuel Consumption
1–2	−1	−1	−1	−1	0.421, 0.419
3–4	−1	0	0	1	0.424, 0.433
5–6	−1	1	1	0	0.355, 0.369
7–8	0	−1	0	0	0.385, 0.391

Test	A	B	C	D	Coded Fuel Consumption
9–10	0	0	1	−1	0.426, 0.429
11–12	0	1	−1	1	0.410, 0.397
13–14	1	−1	1	1	0.405, 0.411
15–16	1	0	−1	0	0.392, 0.395
17–18	1	1	0	−1	0.356, 0.372

12 An experiment is run to determine the effects of four factors on the particulate emissions of a two-cycle outboard motor used in recreational motor boats. The factors, each of which has two levels, include the following: A = exhaust depth, B = water injection rate, C = propeller speed, and D = engine speed. Because of cost concerns, a half-fraction design was chosen using as the defining contrast $I = ABCD$. In addition, three repeat test runs were conducted. Analyze the data from this experiment and determine which factors influence the response (use a 0.05 significance level). Include plots of the main effects and significant interactions. Also, include a normal plot of the data. Interpret the results of the experiment.

Test	A	B	C	D	Particulates
1	−1	1	1	−1	6.4
2	1	−1	−1	1	2.1
3	1	1	−1	−1	2.6
4	−1	−1	−1	−1	3.1
5	−1	1	−1	1	4.7
6	1	−1	1	−1	5.2
7	−1	−1	1	1	5.5
8	1	1	1	1	4.6
2	1	−1	−1	1	2.6
4	−1	−1	−1	−1	3.4
8	1	1	1	1	5.0

13 A study was conducted to determine the factors that are associated with the failures of rotors in a disc brake system used on automobiles. The response variable was a measure of the thermal stress on a rotor because of braking, with a larger value indicating higher stress. The four factors of interest include A = rotor type, B = rotor thickness (from the tip of the rotor to its hub), C = type of braking, and D = rotor material. There were only two rotor materials to be compared, but the other three factors each had three levels. A nine-run fraction of a 2×3^3 experiment was chosen

for the design. The data, including three repeats, are given below. Analyze the data from this experiment and determine which factors influence the response (use a 0.05 significance level). Interpret the results of the experiment, with special emphasis on inferences to determine which factor–level combinations produce the least mean stress.

Test	*A*	*B*	*C*	*D*	Stress
1	−1	−1	−1	−1	49
2	−1	0	0	−1	50
3	−1	1	1	0	54
4	0	−1	0	0	53
5	0	0	1	−1	54
6	0	1	−1	−1	56
7	1	−1	1	−1	41
8	1	0	−1	0	44
9	1	1	0	−1	47
3	−1	1	1	0	56
5	0	0	1	−1	51
8	1	0	−1	0	42

14 A sixteen-run experiment was devised to determine the effects of five factors on a response variable. Each of the five factors has three levels, and three repeat tests are included. The factor–level combinations and the corresponding responses are given below. Analyze the data from this experiment and determine which factors influence the response (use a 0.05 significance level). Draw appropriate inferences to determine which combinations of the factor levels produce the greatest and the least values of the mean response.

Test	*A*	*B*	*C*	*D*	*E*	*Y*
1	−1	−1	−1	−1	−1	133
2	−1	0	0	1	0	137
3	−1	1	1	0	0	139
4	−1	0	0	0	1	142
5	0	−1	0	0	0	135
6	0	0	−1	0	1	148
7	0	1	0	1	−1	126
8	0	0	1	−1	0	132
9	1	−1	1	1	1	150
10	1	0	0	−1	0	146
11	1	1	−1	0	0	136

Test	A	B	C	D	E	Y
12	1	0	0	0	−1	148
13	0	−1	0	0	0	122
14	0	0	1	0	−1	122
15	0	1	0	−1	1	145
16	0	0	−1	1	0	129
3	−1	1	1	0	0	135
7	0	1	0	1	−1	120
14	0	0	1	0	−1	128

15 Consider a test run to determine the effects of five factors on the start time of a diesel engine. The five factors include ambient temperature (A), ambient humidity (B), engine load (C), engine speed (D), and type of fuel (E). Each factor has two levels. The experiment was run as a one-half fraction of a 2^5experiment using as the defining contrast $I = ABCDE$.

(a) Analyze the data from this experiment and determine which factors influence the response using a normal quantile plot.

(b) Repeat the analysis in (a) but assume all the two-way interactions with A and B are negligible and use analysis-of-variance techniques.

(c) Compare your conclusions using approaches (a) and (b).

Test	A	B	C	D	E	Start Time
1	−1	−1	−1	−1	1	9.8
2	−1	−1	−1	1	−1	7.0
3	−1	−1	1	−1	−1	3.6
4	−1	−1	1	1	1	2.6
5	−1	1	−1	−1	−1	6.9
6	−1	1	−1	1	1	8.2
7	−1	1	1	−1	1	12.1
8	−1	1	1	1	−1	10.1
9	1	−1	−1	−1	−1	7.6
10	1	−1	−1	1	1	5.8
11	1	−1	1	−1	1	13.4
12	1	−1	1	1	−1	12.9
13	1	1	−1	−1	1	9.8
14	1	1	−1	1	−1	7.5
15	1	1	1	−1	−1	11.6
16	1	1	1	1	1	9.2

16 Suppose the start-time experiment in Exercise 15 were modified to include eight factors, each at two levels. The factors include ambient temperature (A), ambient humidity (B), engine load (C), engine speed (D), and fuel properties $E, F, G,$ and H. A twelve-run Plackett–Burman design is utilized. Analyze the data from this experiment and determine which factors influence the response (use a 0.05 significance level). Determine the factor levels that produce the shortest mean start times.

Test	A	B	C	D	E	F	G	H	Start Time
1	1	1	−1	1	1	1	−1	−1	15.3
2	1	−1	1	1	1	−1	−1	−1	12.7
3	−1	1	1	1	−1	−1	−1	1	7.3
4	1	1	1	−1	−1	−1	1	−1	9.2
5	1	1	−1	−1	−1	1	−1	1	15.0
6	1	−1	−1	−1	1	−1	1	1	12.4
7	−1	−1	−1	1	−1	1	1	−1	7.5
8	−1	−1	1	−1	1	1	−1	1	8.3
9	−1	1	−1	1	1	−1	1	1	8.5
10	1	−1	1	1	−1	1	1	1	10.9
11	−1	1	1	−1	1	1	1	−1	7.7
12	−1	−1	−1	−1	−1	−1	−1	−1	6.6

17 Construct a normal quantile plot for the data in Exercise 16. Compare your results with those obtained by using ANOVA techniques. Explain any differences.

18 A chemical experiment is run to determine the effects of six factors on the yield of a chemical reactor. The factors each have two levels and are designated as A–F. The defining contrast is $I = -ABCE = BCDF$. Analyze the data from this experiment and determine which factors influence the response (use a 0.05 significance level).

Test	A	B	C	D	E	F	Yield
1	−1	−1	−1	−1	1	−1	3.21
2	−1	1	−1	−1	−1	1	7.87
3	−1	−1	1	−1	−1	1	4.67
4	−1	1	1	−1	1	−1	6.49
5	−1	−1	−1	1	1	1	6.32
6	−1	1	−1	1	−1	−1	8.31
7	−1	−1	1	1	−1	−1	4.84
8	−1	1	1	1	1	1	7.78

Test	A	B	C	D	E	F	Yield
9	1	−1	−1	−1	−1	−1	13.66
10	1	1	−1	−1	1	1	14.50
11	1	−1	1	−1	1	1	13.99
12	1	1	1	−1	−1	−1	14.57
13	1	−1	−1	1	−1	1	12.28
14	1	1	−1	1	1	−1	16.24
15	1	−1	1	1	1	−1	15.89
16	1	1	1	1	−1	1	17.56
5	−1	−1	−1	1	1	1	6.53
10	1	1	−1	−1	1	1	13.85
2	−1	1	−1	−1	−1	1	7.20
14	1	1	−1	1	1	−1	16.02

19 Reanalyze the data in Exercise 18 but do not use the repeat data. Compare your results with those obtained in Exercise 18.

PART III

Design and Analysis with Random Effects

CHAPTER 9

Experiments in Randomized Block Designs

In this chapter we introduce block designs as an experimental technique for controlling variability, thereby allowing more precision in the estimation of factor effects. Of special relevance to this discussion are the following topics:

- *the use of replication and local control in reducing the variability of estimators of factor effects,*
- *the application of blocking principles to factorial experiments,*
- *the construction and use of randomized complete block designs when there are sufficiently many homogeneous experimental units so that all factor combinations of interest can be run in each block,*
- *the use of incomplete block designs when the number of homogeneous experimental units in each block is less than the number of factor–level combinations of interest, and*
- *the use of latin-square, Graeco-latin-square, and crossover designs to control variability when special restrictions are placed on the experimentation.*

In Section 4.2 the presence of excessive experimental variability was shown to contribute to the imprecision of response variables and of estimators of factor effects. This imprecision affects statistical analyses in that real factor effects can be hidden because of the variation of individual responses. Excessive variability can occur from the measurement process itself, the test conditions at the time observations are made, or a lack of homogeneity of the experimental units on which measurements are made.

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

Chief among the statistical design techniques for coping with variability in an experiment are repeat tests, replication, and local control. Evolving from these techniques will be the usefulness of block designs, particularly randomized complete and incomplete block designs. These designs can be used with either complete or fractional factorial experiments. When restrictions are placed on the experimentation in addition to the number of test runs that can be conducted under homogeneous test conditions or on homogeneous experimental units, alternative block designs such as latin squares and crossover designs can be effective in controlling variability. Each of these topics is discussed in the following sections of this chapter.

9.1 CONTROLLING EXPERIMENTAL VARIABILITY

There are several ways that an experimenter can, through the choice of a statistical design, control or compensate for excessive variability in experimental results. Three of the most important are the use of repeat tests, replication, and local control. Although these techniques neither reduce nor eliminate the inherent variability of test results, their proper application can greatly increase the precision of statistics used in the calculation of factor effects.

As mentioned earlier in Table 4.1, we distinguish between repeat tests and replications (see Exhibit 9.1). Repeat tests enable one to estimate experimental measurement-error variability, whereas replications generally provide estimates of error variability for the factors or conditions that are varied with the replication. Not only do repeat tests and replication provide estimates of each type of variability, but increasing the number of repeat tests or replications also increases the precision of estimators of factor effects that are subject to each type of error.

As an illustration of the distinction between the use of repeat tests and replications, consider an experiment designed to investigate the cutoff times of automatic safety switches on lawnmowers. One version of this type of safety mechanism requires that a lever be pressed by the operator while the mower is being used. Release of this lever by the operator results in the mower engine automatically stopping. The cutoff time of interest is the length of time between the release of the lever and the activation of the engine cutoff.

EXHIBIT 9.1

Repeat Tests. Repeat tests are two or more observations in which all factor–level combinations are held constant and the observations are taken under identical experimental conditions.

Replication. Replication occurs when an entire experiment or portion of an experiment is repeated under two or more different sets of conditions.

TABLE 9.1 Experimental Design for Lawnmower Automatic Cutoff Times (Nonrandomized)

Test Run	Manufacturer	Lawnmower	Speed
1	A	1	Low
2	A	1	Low
3	A	1	High
4	A	1	High
5	A	2	Low
6	A	2	Low
7	A	2	High
8	A	2	High
9	A	3	Low
10	A	3	Low
11	A	3	High
12	A	3	High
13	B	1	Low
14	B	1	Low
15	B	1	High
16	B	1	High
17	B	2	Low
18	B	2	Low
19	B	2	High
20	B	2	High
21	B	3	Low
22	B	3	Low
23	B	3	High
24	B	3	High

Table 9.1 displays a nonrandomized test sequence for an experiment that could be used to evaluate the cutoff times of lawnmowers produced by two manufacturers. Three lawnmowers from each manufacturer are included in the design. The cutoff times are to be evaluated at two engine operating conditions, low and high speed.

In this design there are two repeat tests for each manufacturer, lawnmower, and speed combination. Note too that the four test runs for each manufacturer (the two high-speed and the two low-speed test runs) are replicated by using three lawnmowers from each manufacturer. There are then two sources of experimental-error variability: error due to using different lawnmowers and error due to uncontrolled variation in the repeat tests.

Figure 9.1 depicts the combining of these two sources of error variation into the overall variability for an observation. In this figure the repeat tests have an uncontrolled error which is depicted as normally distributed around zero. The

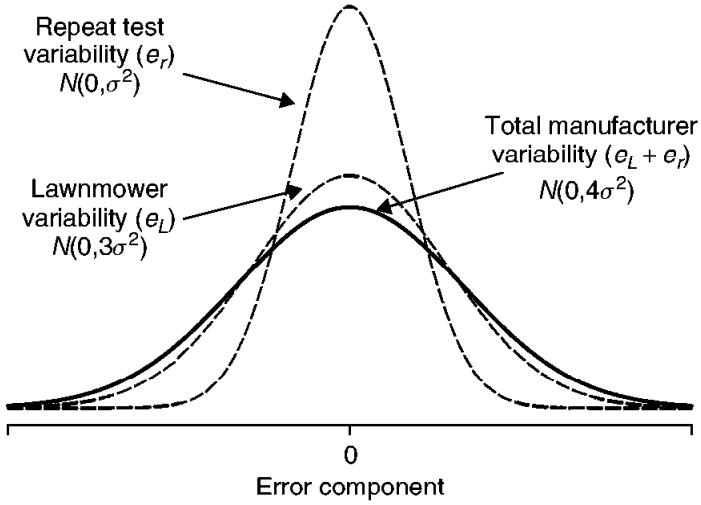


Figure 9.1 Components of variability for lawnmower cutoff times.

variability of the lawnmowers is also depicted as normally distributed around zero, but with a variance that is three times larger than that of the repeat test errors. The total manufacturer variability shows the combining of these errors in the distribution of responses from one of the manufacturers. Assuming that these errors are statistically independent (see Section 2.3), the distribution of the total error is normal about zero with variance equal to the sum of the component variances.

Any comparisons between the cutoff times for the two manufacturers must take account of the variability due to both sources of error. This is because the effect of the manufacturers is calculated as the difference of the average cutoff times for the two manufacturers (Section 5.3), and each average is based on observations from all three lawnmowers made by that manufacturer and on the two repeat tests for each of the speeds. In particular, any comparison of manufacturer A with manufacturer B requires averages over different lawnmowers (and repeat tests); no comparison can be made on a single lawnmower.

Comparisons between the two operating speeds are ordinarily more precise than those between the manufacturers. This is because comparisons between the two speeds are only affected by the repeat-test variability. To appreciate this fact note that speed effects can be calculated for each individual lawnmower by taking differences of averages across repeat tests. Each of these six individual speed effects is subject to only the repeat-test variability, because each is based on averages for one lawnmower from one of the manufacturers. The overall effect of the two speeds (the difference of the overall averages for the high and the low speeds) can be calculated by taking the average of the six individual speed effects. Consequently, the overall effect of the speeds on the

cutoff times, like the speed effects for the individual lawnmowers, is subject to only the uncontrolled variability due to the repeat tests.

As will be seen in Chapter 11, estimates of variability suitable for comparing manufacturers and for comparing speeds can be obtained from this experiment. The variability estimate needed for a comparison of the speeds is available from the repeat tests. The variability estimate needed for a comparison of the manufacturers is available from the replications. The greater precision in the comparisons of the speeds occurs because the estimate of error variability due to the repeat tests is ordinarily less than that of the replications.

The importance of this example is that the use of replication and repeat tests in an experiment design may be necessary to adequately account for different sources of response variability. Had no repeat tests been conducted, any estimate proposed for use as an estimate of the uncontrolled experimental error might be inflated by effects due to the lawnmowers or to one or more of the design factors (manufacturers, speed conditions). If the replicate observations on the six lawnmowers were not available, comparisons between the manufacturers would ordinarily be made relative to the estimate of the uncontrolled measurement error. If so, an erroneous conclusion that real differences exist between the manufacturers could be drawn when the actual reason was large variation in cutoff times of the lawnmowers made by each manufacturer. With no replication (only one lawnmower from each manufacturer), the effect due to the manufacturers would be confounded with the variation due to lawnmowers.

An important application of the use of repeat tests and replication is in the design of interlaboratory testing studies. In these studies experiments are conducted to evaluate a measurement process, or *test method.* The primary goal of such experiments is to estimate components of variation of the test method. Two key components of the total variation of the test method are variability due to repeat testing at a single laboratory and that due to replicate testing at different laboratories. The variability associated with the difference of two measurements taken at a single laboratory by the same operator or technician is referred to as the *repeatability* of the measurement process. The variability associated with the difference of two measurements taken at two different laboratories by two different operators or technicians is referred to as the *reproducibility* of the measurement process. Repeatability and reproducibility studies, often termed Gage R&R studies, are important scientific experiments used to assess the adequacy of test methods.

Another major technique for controlling or compensating for variability in an experiment is termed *local control* (see Exhibit 9.2). It consists of the planned grouping of experimental units.

EXHIBIT 9.2 LOCAL CONTROL

Local control involves the grouping, blocking, and balancing of the experimental units.

Grouping simply refers to the placement of experimental units or of test runs into groups.

Blocking refers to the grouping of experimental units so that units within each block are as homogeneous as possible.

Balancing refers to the allocation of factor combinations to experimental units so that every factor–level combination occurs an equal number of times in a block or in the experiment.

The grouping of experimental units or of test runs often is undertaken because it is believed that observations obtained from tests within a group will be more uniform than those between groups. For example, one might choose to conduct certain test runs during one day because of known day-to-day variation. Repeat tests are sometimes conducted back to back in order to remove all sources of variability other than those associated with measurement error. While this type of grouping of test runs can result in unrealistically low estimates of run-to-run uncertainty, it is sometimes used to provide an estimate of the minimum uncertainty that can be expected in repeat testing. The grouping of a sequence of tests with one or more factors fixed is sometimes necessitated because of economic or experimental constraints.

Blocking is a special type of grouping whose origin is found in agricultural experiments. Agricultural land often is separated into contiguous blocks of homogeneous plots. Within each block the plots of land are experimental units which receive levels of the experimental factors (fertilizers, watering regimens, etc.). Thus, the experimental errors among plots of ground within a block are relatively small because the plots are in close physical proximity. Errors between plots in two different blocks are usually much larger than those between plots in the same block, because of varying soil or environmental conditions.

Blocking or grouping can occur in a wide variety of experimental settings. Because the experimental design concepts relating to blocking and grouping are ordinarily identical, we shall henceforth use the term *blocking* to refer to any planned grouping of either test runs or experimental units. In some experiments blocks may be batches of raw materials. In others, blocks may be different laboratories or groups of subjects. In still others, blocking may be the grouping of test runs under similar test conditions or according to the levels of one or more of the design factors. Blocking is, thus, an experimental design technique that removes excess variation by grouping experimental units or test runs so that those units or test runs within a block are more homogeneous than those in different blocks. Comparisons of factor effects are then achieved by comparing responses or averages of observations within each block, thereby removing variability due to units or test conditions in different blocks.

The lawnmowers in the cutoff-time example are blocks, because the four observations for each lawnmower are subject only to repeat-test variability whereas observations on different lawnmowers are also subject to variations among the lawnmowers. The agricultural experiment depicted in Figure 4.2 also illustrates a block design with the three fields constituting blocks.

Balancing (see also Section 6.1) is a property of a design that assures that each factor–level combination occurs in the experiment an equal number of times. One often seeks to have balance in a design to ensure that the effects of all factor levels are measured with equal precision. Designs that are not balanced necessarily have certain factor effects that are measured with less precision than others. For ease of interpretation and analysis, we generally desire that designs be balanced.

9.2 COMPLETE BLOCK DESIGNS

One of the most often used designs is the randomized complete block (RCB) design. The steps in constructing a RCB design are given in Exhibit 9.3. It is assumed that no interaction effects exist between the blocks and the experimental factors. If this is not the case, designs that allow for interaction effects must be used (see Chapter 5).

EXHIBIT 9.3 RANDOMIZED COMPLETE BLOCK DESIGN

1. Group the experimental units or test runs into blocks so that those within a block are more homogeneous than those in different blocks.
2. Randomly assign all factor combinations of interest in the experiment to the units or test sequence in a block; use a separate randomization for each block. Complete or fractional factorial experiments can be included in each block.

A common application of the blocking concept occurs in experiments in which responses are collected on pairs of experimental units. Pairing may occur because only two experimental units can be produced from each shipment of raw materials, because experiments are performed on sets of nearly identical experimental units such as twins, or because experimental units in close geographical proximity (such as plots of ground) are more uniform than widely separated ones.

In each of the above illustrations, the blocks represented physical units. Closely related to observations taken on pairs of experimental units are pairs of observations that are collected under similar experimental conditions. Repeat observations taken at several different locations, pre- and post-testing of the same unit or individual, and experimental conditions that limit the experimenter to two test runs per day are examples of pairs of observations that are blocked.

RCB designs are particularly useful in the physical and engineering sciences. This is because experimentation often is fragmented across several time periods (such as work shifts), different laboratories, separate groups of materials, or different test subjects. Blocking (grouping) on the sources of uncontrolled variability allows more sensitive comparisons of factor effects.

Despite the many advantages in using a RCB design, there are experimental conditions that necessitate the use of alternative blocking designs. For instance, the RCB designs described above only control for one extraneous source of variability. If two or more sources of extraneous variability exist, one can sometimes block on all the sources simultaneously using latin-square (Section 9.4) or other alternative block designs.

RCB designs require replicate experiments in which all factor–level combinations called for by a complete or a fractional factorial experiment are contained in each block. Thus, even when there is only one source of extraneous variability that is of concern, there may not be a sufficient number of homogeneous experimental units available to include all factor–level combinations of interest in each block. In such situations one can often utilize designs that are balanced but do not include all factor–level combinations in each block. A selection of some of the most important and most widely used incomplete block designs is presented in the next two sections.

9.3 INCOMPLETE BLOCK DESIGNS

Situations can occur in constructing a block design in which it is not possible to include all factor combinations in every block. The result is an incomplete block design. There are several different types of incomplete block designs available. This section introduces two general classes of incomplete block designs: designs derived from confounding patterns of fractional factorials and a special type of balanced incomplete block designs.

9.3.1 Two-Level Factorial Experiments

The planned confounding of factor effects in blocking designs is an effective design strategy for factorial experiments in which all factor–level combinations of interest cannot be tested under uniform conditions. The lack of uniform conditions can occur either because there are not a sufficient number of homogeneous experimental units or because environmental or physical changes during the course of the experiment affect the responses.

Randomized complete block designs were introduced in the last section as an alternative to completely randomized designs when factor–level combinations must be blocked or grouped. However, randomized complete block designs require that the number of experimental units be at least as large as the total number of factor–level combinations of interest. This requirement is impractical in many experiments.

There are many situations in which the intentional aliasing of factor effects with block effects can satisfy the requirements of the experiment and still allow the evaluation of all the factor effects of interest. Only specifically chosen factor effects are confounded with block effects when factorial experiments are partitioned into blocks. The goal of conducting factorial experiments in blocks is to plan the aliasing so that only those effects that are known to be zero or that are of little interest are confounded with the block effects.

The manufacturing-plant example in Section 7.1 illustrates the key concepts in constructing block designs for complete factorial experiments. Care must be taken to ensure that effects of interest are not aliased with block effects. Table 9A.1 in the appendix to this chapter lists the defining equations for some block designs involving three to seven factors. In addition to the specified aliased effects, other effects are implicitly confounded as with fractional factorial experiments. The additional aliased effects are identified by taking the algebraic products of the defining contrasts, as described in Section 7.3.

The designing of a randomized incomplete block design in which one or more effects from a complete factorial experiment are intentionally aliased with blocks proceeds as described in Exhibit 9.4. The construction of randomized incomplete block designs requires 2^p blocks of experimental units if p effects are to be confounded. If there are k factors in the experiment, each block consists of at least 2^{k-p} experimental units. If more than 2^{k-p} experimental units are available in one or more of the blocks, repeat observations can be taken. This will result in the design being unbalanced, but it will enable estimation of the uncontrolled experimental-error variation without the possible influences of factor or block effects.

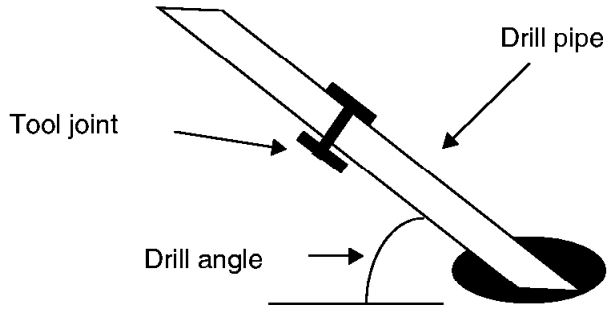
EXHIBIT 9.4 CONSTRUCTION OF RANDOMIZED INCOMPLETE BLOCK DESIGNS FOR TWO-LEVEL COMPLETE FACTORIAL EXPERIMENTS

1. Determine a block size that can be used in the experiment. If there are to be no repeats, the block size should be a power of 2.
2. Choose the effects that are to be confounded with the block effects. Table 9A.1 can be used to choose defining contrasts for experiments with three to seven factors.
3. List the effects representation of the defining contrasts, and randomly select the sign to be used with each.
4. Randomly select a block, and place in that block all factor–level combinations that have a specific sign combination in the defining contrasts.
5. Repeat step 4 for each of the groups of factor–level combinations.
6. Randomize the test sequence or the assignment to experimental units separately for each block.

In some experiments there are an ample number of blocks. In such circumstances the construction of incomplete block designs as described in Exhibit 9.4 can be modified to include replicates of the complete factorial experiment. In each replicate different effects can be aliased with blocks so that every effect can be estimated from at least one of the replicates. This type of design is referred to as a *partially* confounded design, and the effects are referred to as *partially confounded.*

To illustrate the construction of complete factorial experiments in randomized incomplete block designs, consider the experiment schematically outlined in Figure 9.2. This experiment is to be conducted to investigate the wear on drill bits used for oil exploration. A scale model of a drilling platform is to be built in order to study five factors believed to affect drill-bit wear: the rotational (factor A) and longitudinal (factor B) velocities of the drill pipe, the length of the drill pipe (factor C), the drilling angle (factor D), and the geometry (factor E) of the tool joint used to connect sections of the pipe.

A factorial experiment is feasible for this study, but only eight test runs, one per hour, can be run each day. Daily changes in the equipment setup may influence the drill-bit wear so the test runs will be blocked with eight test runs per block. Two weeks are set aside for the testing. A complete factorial experiment can be conducted during each week by running a quarter fraction on each of four days. The other day can be reserved, if needed, for setting up and testing equipment. Because two replicates of the complete factorial experiment can be conducted, different defining contrasts will be used



Response: drill bit wear

Factors	Levels
Rotational drill speed	60 rpm, 75 rpm
Longitudinal velocity	50 fpm, 100 fpm
Drill pipe length	200 ft, 400 ft
Drilling angle	30, 60 degrees
Tool joint geometry	Straight, ellipsoidal edges

Figure 9.2 Drill-bit wear experiment.

each week so that all factor effects can be estimated when the experiment is concluded.

In Table 9A.1, the defining equations for blocking are $I = ABC = CDE$ $(= ABDE)$. The effects representations for these interactions are shown in Table 9.2. To conduct the experiment, the factor–level combinations that have negative signs on the effects representations for both defining contrasts would be randomly assigned to one of the four test days in the first week. Those with a positive sign on ABC and a negative sign on CDE would be randomly assigned to one of the remaining three test days in the first week. The other two blocks of eight factor–level combinations would similarly be randomly assigned to the remaining two test days in the first week. Prior to conducting the test runs, each of the eight runs assigned to a particular test day would be randomly assigned a run number.

For the second week of testing, a similar procedure would be followed except that two different three-factor interactions would be aliased with blocks. One might choose as the defining equations $I = BCD = ADE (= ABCE)$. It is important to emphasize that the aliasing pattern for this complete factorial experiment conducted in four blocks is different from that of a quarter fraction of the complete factorial. In the latter situation, each effect is aliased with three other effects, with the confounding pattern being determined from the defining contrasts. In this incomplete block design, however, it is only the defining contrasts that are aliased with blocks. All other main effects and interactions are unaliased because this is a complete factorial experiment.

A further adaptation of randomized incomplete block designs is their use with fractional factorial experiments. With such designs there is the confounding of factor effects as with any fractional factorial experiment, and there is additional confounding with blocks because all of the factor–level combinations in the fractional factorial cannot be run in each block.

Table 9A.2 lists sets of defining contrasts for fractional factorial experiments conducted in randomized incomplete block designs. All of the designs resulting from the use of Table 9A.2 are at least of Resolution V. To conduct these experiments, one first selects the factor–level combinations to be included in the experiments from the defining contrasts for the factors in Table 9A.2. Once these factor–level combinations are identified, they are blocked according to their signs on the defining contrasts for the blocks in Table 9A.2.

For example, to conduct a half fraction of the acid-plant corrosion-rate study (introduced in Chapter 7) in two blocks of 16 test runs each, one would first select the factor–level combinations to be tested using the defining equation $I = \pm ABCDEF$, with the sign randomly chosen. Suppose the negative signs are selected so that the combinations are as listed in Table 7.5. Next these 32 factor–level combinations would be blocked according to whether the effects representation for the ABC interaction is positive or negative. Thus,

TABLE 9.2 Effects Representation for Blocking in the Drill-Bit Wear Experiment

Factor–Level Combination	Effects Representation						
	A	*B*	*C*	*D*	*E*	*ABC*	*CDE*
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
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
17	1	−1	−1	−1	−1	1	−1
18	1	−1	−1	−1	1	1	1
19	1	−1	−1	1	−1	1	1
20	1	−1	−1	1	1	1	−1
21	1	−1	1	−1	−1	−1	1
22	1	−1	1	−1	1	−1	−1
23	1	−1	1	1	−1	−1	−1
24	1	−1	1	1	1	−1	1
25	1	1	−1	−1	−1	−1	−1
26	1	1	−1	−1	1	−1	1
27	1	1	−1	1	−1	−1	1
28	1	1	−1	1	1	−1	−1
29	1	1	1	−1	−1	1	1
30	1	1	1	−1	1	1	−1
31	1	1	1	1	−1	1	−1
32	1	1	1	1	1	1	1

Factors	Levels	
	−1	+1
A = Rotational drill speed (rpm)	60	75
B = Longitudinal drill speed (fpm)	50	100
C = Drill pipe length (ft)	200	400
D = Drilling angle (deg)	30	60
E = Tool joint geometry	Straight	Ellipse

factor–level combinations 1–4, 13–16, and 21–29 would be assigned to one of the blocks, and the remaining combinations to the other block.

If a large experiment is required to satisfactorily investigate all the factor effects of interest, consideration should be given to blocking the test runs so that information can be gained about the factors as the experiment progresses. For fractional factorial experiments, Table 9A.2 shows blocking patterns that ensure that no main effects or two-factor interactions are confounded. These blocking patterns can be exploited to allow sequential testing in which information about key factors can be evaluated as the experiment proceeds, not simply when it is completed.

As noted in Section 7.2, a feature of two-level, fractional factorial experiments is that a fractional factorial of Resolution R is a complete factorial experiment in any $R - 1$ factors. Because of this property, one can often block the experiment so that preliminary information on main effects and two-factor interactions can be obtained from each of the blocks. This information is considered preliminary because information on all the main effects and two-factor interactions may not be obtainable from each of the blocks and because each of the effects will be estimated with greater precision at the completion of the experiment.

Recall the discussion about conducting a half fraction of the complete factorial for the acid-plant corrosion-rate study in two blocks of 16 test runs. The defining contrast for the half fraction is $I = -ABCDEF$. The defining contrast for the blocking is $I = ABC$. After the first block of test runs is completed, a quarter fraction of the complete factorial experiment will have been conducted. This quarter fraction has as its defining equations either $I = -ABCDEF = ABC$ or $I = -ABCDEF = -ABC$, depending on which block of factor–level combinations was run first. Note that this quarter fraction is itself a Resolution-III design. Thus, from the first block of test runs one can estimate main effects that are unconfounded with one another. Based on this information, decisions may be made about continuing the experimentation or terminating it, depending on whether the main effects show that one or more of the factors is having a strong effect on the corrosion rates.

This example illustrates how blocking designs can be used to exploit the planned confounding of factor effects in sequential experimentation. Care must be exercised in the interpretation of results from highly fractionated designs so that one does not draw erroneous conclusions because of the confounding pattern of the effects. Nevertheless, conducting experiments in blocks can alert one early to dominant effects, which in turn can influence one's decision to continue the experimentation.

9.3.2 Three-Level Factorial Experiments

Because three-level factorial experiments require such a large number of experimental units or homogeneous test run conditions, they are often blocked.

Using the same strategy as with two-level factorials, intentional aliasing of selected factor effects with blocks enables three-level factorials to be executed in incomplete block designs with no aliasing of important factor effects. This section summarizes the general approach to incomplete block designs for three-level factorials. The same notation for effects that was used in Section 7.4.1 is used in this section.

Three-level factorials are most efficiently blocked when one-third fractions of complete factorials can be included in each block. Other blocking schemes usually result in very complicated confounding patterns. In general, one attempts to alias the high-order interactions with blocks. Recall from Section 7.4.1 that a three-factor complete factorial has interactions that are labeled ABC, ABC^2, AB^2C, and AB^2C^2, each having two degrees of freedom. The steps listed in Exhibit 9.5 should use one of these interactions as the defining contrast denoted T.

EXHIBIT 9.5 CONSTRUCTION OF RANDOMIZED INCOMPLETE BLOCK DESIGNS IN THREE BLOCKS FOR THREE-LEVEL, COMPLETE FACTORIAL EXPERIMENTS

1. Choose a defining contrast; e.g., for five-factor factorial $I = AB^2C^2D^2E^2$. This contrast is aliased with the two degrees of freedom for block effects.
2. Denote by x_j the coded levels for factor j, where the coded levels are 0, 1, and 2.
3. Symbolically denote the defining contrast as

$$T = a_1x_1 + a_2x_2 + \ldots a_kx_k$$

   where x_j denotes the coded levels of factor j and $a_j = 0$, 1, or 2 is the power on the jth factor in the defining contrast ($a_j = 0$ indicates the factor is not in the defining contrast); e.g., for the contrast in step 1, $T = x_1 + 2x_2 + 2x_3 + 2x_4 + 2x_5$.
4. For each factor–level combination in the complete factorial, determine the value of $X = T \bmod(3)$. To compute X, determine the value of r (0, 1, or 2) for which $(T - r)/3$ is an integer.
5. Randomly select one of the three blocks and one of the integers $X = 0$, 1, or 2. Assign all factor–level combinations to the block that have the selected value of X. Do the same for the other two blocks and the other two values of X.
6. Randomize the test sequence or the assignment of factor–level combinations to experimental units separately for each block.

Table 9.3 shows the construction of an incomplete block design for a complete factorial experiment with three three-level factors in three blocks. The fifth column of the table shows the calculations needed for step 4 in

TABLE 9.3 Nonrandomized Incomplete Block Design for a 3^3 Complete Factorial Experiment in 3 Blocks: Defining Contrast $I = ABC$

Factor–Level Combination	Factor			X	
	A (x_1)	B (x_2)	C (x_3)	$(x_1 + x_2 + x_3)$ mod(3)	Block
1	0	0	0	0	1
2	0	0	1	1	2
3	0	0	2	2	3
4	0	1	0	1	2
5	0	1	1	2	3
6	0	1	2	0	1
7	0	2	0	2	3
8	0	2	1	0	1
9	0	2	2	1	2
10	1	0	0	1	2
11	1	0	1	2	3
12	1	0	2	0	1
13	1	1	0	2	3
14	1	1	1	0	1
15	1	1	2	1	2
16	1	2	0	0	1
17	1	2	1	1	2
18	1	2	2	2	3
19	2	0	0	2	3
20	2	0	1	0	1
21	2	0	2	1	2
22	2	1	0	0	1
23	2	1	1	1	2
24	2	1	2	2	3
25	2	2	0	1	2
26	2	2	1	2	3
27	2	2	2	0	1

Exhibit 9.5. The three one-third fractions and the blocks to which they are assigned are shown in the last column of the table.

9.3.3 Balanced Incomplete Block Designs

By definition, incomplete block designs cannot be balanced within a block because not all factor–level combinations occur in each block. Balance can be achieved across blocks that contain a complete replicate of the experiment. The specially designed balanced incomplete block (BIB) designs described in this

section achieve balance in the number of times each factor–level combination occurs in the design and in the number of blocks in which each pair of factor–level combinations occur.

BIB designs can be found in Table 9A.3 in the appendix to this chapter. Table 9A.3 lists 17 BIB designs according to several design parameters. The design parameters are listed in Exhibit 9.6. Although Table 9A.3 only lists designs for eight or fewer factor–level combinations, BIB designs are available for experiments having more than eight combinations. However, BIB designs do not exist for all possible choices of the design parameters. This is why only select designs are presented in Table 9A.3.

EXHIBIT 9.6 KEY PARAMETERS IN A BIB DESIGN

f = number of factor–level combinations
r = number of times each factor–level combination occurs in the design
b = number of blocks
k = number of experimental units in a block
p = number of blocks in which each pair of factor–level combinations occurs together

---

Balance is achieved in BIB designs because every factor–level combination occurs an equal number of times (r) in the design and each pair of factors occur together in an equal number (p) of blocks in the design. Because of the desire for balance in these designs, certain restrictions are placed on them. These restrictions prevent BIB designs from existing for every possible choice of the above design parameters. The nature of some of the restrictions is indicated by the properties of BIB designs shown in Exhibit 9.7. Note in particular that p must be an integer and that b must be at least as great as f.

EXHIBIT 9.7 PROPERTIES OF A BALANCED INCOMPLETE BLOCK DESIGN

- Total number of observations:

$$N = fr = bk.$$

- Number of blocks in which a pair of factor–level combinations occur:

$$p = r\frac{k-1}{f-1}.$$

- Relation between the number of blocks and the number of factor–level combinations:

$$b \geqslant f.$$

The factor–level combinations referred to in BIB designs can be all the levels of a single factor or combinations of levels of two or more factors. Both complete and fractional experiments can be conducted in BIB designs, subject to the existence of a design for the chosen design parameters. A BIB design is chosen following the procedure in Exhibit 9.8.

EXHIBIT 9.8 BALANCED INCOMPLETE BLOCK DESIGN

1. Group the experimental units or test runs into blocks so that those within a block are more homogeneous than those in different blocks.
2. Check to be certain that the number of factor–level combinations to be used in the experiment is greater than the number of units or test runs available in each block.
3. Refer to Table 9A.3 to select a design for the chosen values of the design parameters.
4. Assign the factor–level combinations to the blocks as dictated by the BIB design selected.
5. Randomly assign the factor–level combinations to the experimental units or to the test sequence in a block; use a separate randomization for each block.

To illustrate the use of BIB designs, consider an experiment intended to measure the fuel consumption (gal/mi) of four engine oils. A single test engine is to be used in the experiment in order to eliminate as many extraneous sources of variation as possible. The engine is run on a dynamometer test stand, which is recalibrated after every second test run because each test run is the equivalent of several thousand road miles. Three replicates with each oil are considered necessary.

In this example a block consists of the two measurements taken between calibrations. In terms of the design parameters, f is the number of engine oils (4), r is the number of replicate tests on each oil (3), and k is the number of tests between calibrations (2). Using the first of the above BIB design equations, the testing must be conducted using $b = 4(3/2) = 6$ blocks. Using the second of the three design equations, each pair of oils will appear together in exactly $p = 3(1/3) = 1$ block. These quantities dictate that the dynamometer will need to be recalibrated six times during the course of the experiment, and that between calibrations any pair of oils being tested will appear together exactly once. Using BIB design 1 from Table 9A.3, the layout of the experiment is as shown in Table 9.4. Prior to actually running this experiment one would randomize both the order of the blocks and the order of the testing of the two oils in each block.

The layout in Table 9.4 illustrates the key features of a BIB design. There are fewer units (2) available in each block than there are factor levels (4).

TABLE 9.4 Balanced Incomplete Block Design for Oil-Consumption Experiment

Block (Recalibration)	Oils Tested
1	A, B
2	C, D
3	A, C
4	B, D
5	A, D
6	B, C

Every oil appears in exactly three blocks and appears with every other oil in exactly one block. Finally, it is reasonable to assume that there is no interaction effect between oil and calibration.

The analysis of experiments conducted in BIB designs is presented in Chapter 10. Because of the restrictions introduced in the construction of BIB designs, it is not appropriate to calculate factor effects using differences of ordinary averages. The averages must be adjusted to take account of the fact that each factor level (or factor–level combination) does not occur in the same blocks as each of the other levels (combinations). Details of this adjustment are presented in Chapter 10.

9.4 LATIN-SQUARE AND CROSSOVER DESIGNS

The only type of restriction considered thus far in the discussion of blocking designs is a restriction on the number of experimental units or test runs that are available in each block. In this section we discuss latin-square designs and crossover designs to illustrate the rich variety of blocking designs that are available for implementation in experimental work. Latin-square designs are appropriate when two sources of extraneous variation are to be controlled. An extension to Graeco-latin-square designs allows the simultaneous control of three sources of extraneous variation. Crossover designs are useful when two or more factor levels are to be applied in a specified sequence in each block.

9.4.1 Latin-Square Designs

Latin-square designs are ordinarily used when there is one factor of interest in the experiment and the experimenter desires to control two sources of variation. These designs can also be used to control two sources of variability with two or more factors of interest. In the latter setting, the factor levels

referred to in the following discussion would be replaced by the factor–level combinations of interest.

Latin-square designs can be used when the number of levels of the factor of interest and the numbers of levels of the two blocking factors are all equal, each number being at least three. It is assumed that no interactions exist between the factor(s) of interest and the two blocking factors. The assumption of no interactions between the factor and blocks is critical to the successful implementation of latin-square designs.

Although either of the blocking factors in a latin-square design could be replaced by another design factor, this implementation is strongly discouraged because of the possibility of interactions between the design factors. If two or all three of the factors in a latin-square design are design factors, the design is actually a form of a fractional factorial experiment. In such experiments it is preferable to conduct fractional factorial experiments in completely randomized, randomized block, or balanced incomplete block designs, as appropriate.

The basic construction of latin-square designs is straightforward. The design is laid out in rows and columns; the number of each equals the number of levels of the factor. Denote the levels of the factor of interest by capital letters: $A, B, C, \ldots, K$. The letters, in the order indicated, constitute the first row of the design:

	Col. 1	**Col. 2**	**Col. 3**	$\cdots$	**Col. *k***
Row 1	A	B	C	$\cdots$	K

Each succeeding row of the design is obtained by taking the first letter in the previous row and placing it last, shifting all other letters forward one position. Examples of latin-square designs for factors having from three to seven levels are given in Table 9A.4 of the appendix to this chapter. The procedure for constructing and using latin-square designs, including the randomization procedure, is given in Exhibit 9.9.

EXHIBIT 9.9 CONSTRUCTION OF LATIN-SQUARE DESIGNS

1. Set up a table having k rows and k columns.
2. Assign the letters, $A, B, C, \ldots, K$ to the cells in the first row of the table.
3. For the second through kth rows of the table, place the first letter of the previous row in the last position and shift all other letters forward one position.
4. Randomly assign one of the blocking factors to the rows, the other one to the columns, and assign the design factor to the body of the table.
5. Randomly assign the levels of the three factors to the row positions, column positions, and letters, respectively.

A latin-square experimental design was used in the road testing of four tire brands for a tire wholesaler. The purpose of the study was to assess whether tires of different brands would result in demonstrably different fuel efficiencies of heavy-duty commercial trucks. To obtain reasonably reliable measures of fuel efficiency, a single test run would have to be of several hundred miles. It was, therefore, decided to use several test trucks and conduct the test so that each type of tire was tested on each truck. It was also necessary to conduct the experiment over several days. The experimenters then required that each tire be tested on each day of the test program.

A latin-square design was chosen to meet the requirements of the test program. Four trucks of one make were selected, and a four-day test period was identified. Randomly assigning the two blocking factors (trucks, days) to the design resulted in the rows of the design corresponding to the trucks, the columns to the days, and the letters to the tires. The layout of the design is given in Table 9.5. Note that each tire is tested on each truck and on each day of the test program.

In the above experiment the factor of interest was the tire brands. The truck and day factors were included because they add extraneous variation to the response. By including these factors, one is blocking (or grouping) on sources of variation and thereby allowing for more precise measurement of the effects due to the tire brands. Failure to include these factors in the design would result in an inflated experimental-error estimate. True differences due to the tire brands might then be masked by the variation due to trucks and days.

As mentioned above, the levels of factors used in latin-square designs can be factor combinations from two or more factors. For example, the four brands could represent four combinations of two tire types, each with two tread designs. Effects on the response due to both tire type and tread design can be analyzed. One can also assess whether an interaction due to tires and treads exists. One cannot, however, assess any interactions between tires or treads and the two other factors.

Graeco-latin-square designs permit three extraneous sources of variation to be controlled. These designs are constructed by superimposing two latin-square designs in such a way that each combination of letters in the two designs

TABLE 9.5 Latin-Square Design for Tire-Test Study

		Day			
		4	3	1	2
Truck	3	Tire 4	Tire 2	Tire 1	Tire 3
	4	Tire 2	Tire 1	Tire 3	Tire 4
	2	Tire 1	Tire 3	Tire 4	Tire 2
	1	Tire 3	Tire 4	Tire 2	Tire 1

appears once in each row and once in each column, similar to latin-square designs. The interested reader should see the references for more detail.

9.4.2 Crossover Designs

Crossover designs are used in experiments in which experimental units receive more than one factor–level combination in a preassigned time sequence. Many biological and medical experiments use crossover designs so that each subject or patient can act as its own control. This type of design is a block design in which each experimental unit (subject, patient) is a block.

Crossover designs are especially useful when there is considerable variation among experimental units. If excessive variation exists among experimental units, comparisons of factor effects could be confounded by the effects of the groups of experimental units that receive the respective factor levels.

There is an extensive literature on crossover designs. We discuss only very simple crossover designs in this section and refer the interested reader to the references for more detailed information. Part of the reason for the extensive literature is the problem of *carryover*, or *residual*, effects. This problem occurs when a previous factor–level combination influences the response for the next factor–level combination applied to the experimental unit. These carryover effects are common in drug testing, where the influence of a particular drug cannot always be completely removed prior to the administering of the next drug. We use the simplifying assumptions in Exhibit 9.10 in the remaining discussion of crossover designs.

EXHIBIT 9.10 CROSSOVER DESIGN ASSUMPTIONS

- Excessive uncontrollable variation exists among the experimental units to be included in the experiment.
- Time periods are of sufficient duration for the factor effects, if any, to be imparted to the response.
- No residual or carryover effects remain from one time period to the next.
- Factor effects, if any, do not change over time (equivalently, no time–factor interactions).

For very small numbers of factor–level combinations, latin-square designs are frequently used as crossover designs. The rows of the latin square represent experimental units, the columns represent time periods, and the symbols represent the factor levels or factor–level combinations. For example, the three-level latin square in Table 9A.4 could be used for a single factor that has three levels. The design can be replicated (with a separate randomization for each) to accommodate more than three experimental units. The four-level

TABLE 9.6 Crossover Design Layout for a Clinical Study of Three Experimental Drugs

	Time Period			
Sequence Number	1	2	3	Subject Numbers
1	Drug 1	Drug 2	Drug 3	20, 4, 2
2	Drug 1	Drug 3	Drug 2	17, 9, 16, 3
3	Drug 2	Drug 1	Drug 3	13, 19, 5
4	Drug 2	Drug 3	Drug 1	12, 7, 1
5	Drug 3	Drug 1	Drug 2	14, 11, 6
6	Drug 3	Drug 2	Drug 1	8, 18, 10, 15

design can be used for a single factor that has four levels or for all four factor–level combinations of a 2^2 factorial experiment.

Larger numbers of factor–level combinations can be accommodated by listing all possible time-order sequences for the factor–level combinations and randomly assigning the sequences to the experimental units. For illustration, Table 9.6 lists all six time orderings for the testing of three experimental drugs. There are twenty subjects who are available for the experimental program. They have been randomly assigned to the six time orderings. Similar designs can be constructed when there are many more than the three factor levels used in this illustration.

APPENDIX: INCOMPLETE BLOCK DESIGN GENERATORS

TABLE 9A.1 Selected Randomized Incomplete Block Designs for Complete Factorial Experiments

No. of Factors	Block Size	Fraction	Defining Equations*
3	2	$\frac{1}{4}$	$I = AB = AC$
	4	$\frac{1}{2}$	$I = ABC$
4	2	$\frac{1}{8}$	$I = AB = BC = CD$
	4	$\frac{1}{4}$	$I = ABD = ACD$
	8	$\frac{1}{2}$	$I = ABCD$
5	2	$\frac{1}{16}$	$I = AB = AC = CD = DE$
	4	$\frac{1}{8}$	$I = ABE = BCE = CDE$

TABLE 9A.1 (*continued*)

No. of Factors	Block Size	Fraction	Defining Equations*
	8	$\frac{1}{4}$	$I = ABC = CDE$
	16	$\frac{1}{2}$	$I = ABCDE$
6	2	$\frac{1}{32}$	$I = AB = BC = CD = DE = EF$
	4	$\frac{1}{16}$	$I = ABF = ACF = CDF = DEF$
	8	$\frac{1}{8}$	$I = ABCD = ABEF = ACE$
	16	$\frac{1}{4}$	$I = ABCF = CDEF$
	32	$\frac{1}{2}$	$I = ABCDEF$
7	2	$\frac{1}{64}$	$I = AB = BC = CD = DE = EF = FG$
	4	$\frac{1}{32}$	$I = ABG = BCG = CDG = DEG = EFG$
	8	$\frac{1}{16}$	$I = ABCD = EFG = CDE = ADG$
	16	$\frac{1}{8}$	$I = ABC = DEF = AFG$
	32	$\frac{1}{4}$	$I = ABCFG = CDEFG$
	64	$\frac{1}{2}$	$I = ABCDEFG$

*Each of the defining contrasts is confounded with a block effect.

TABLE 9A.2 Selected Randomized Incomplete Block Designs for Fractional Factorial Experiments

Number of Factors	Block Size	Resolution	Defining Equations*	
			Factors	Blocks
5	None	V	$I = ABCDE$	
6	16	VI	$I = ABCDEF$	$I = ABC$
7	8	VII	$I = ABCDEFG$	$I = ABCD = ABEF = ACEG$
8	16	V	$I = ABCDG = ABEFH$	$I = ACE = CDH$

(*continued overleaf*)

TABLE 9A.2 (*continued*)

Number of Factors	Block Size	Resolution	Defining Equations*	
			Factors	Blocks
9	16	VI	$I = ACDFGH = BCEFGJ$	$I = ACH = ABJ = GHJ$
10	16	V	$I = ABCGH = BCDEJ = ACDFK$	$I = ADJ = ABK = HJK$
11	16	V	$I = ABCGH = BCDEJ = ACDFK = ABCDEFGL$	$I = ADJ = ABK = HJK$

*No main effects or two-factor interactions are confounded with one another or with block effects.

TABLE 9A.3 Selected Balanced Incomplete Block Designs for Eight or Fewer Factor–Level Combinations

f	k	r	b	p	Design No.
4	2	3	6	1	1
	3	3	4	2	2
5	2	4	10	1	3
	3	6	10	3	4
	4	4	5	3	5
6	2	5	15	1	6
	3	5	10	2	7
	3	10	20	4	8
	4	10	15	6	9
	5	5	6	4	10
7	2	6	21	1	11
	3	3	7	1	12
	4	4	7	2	13
	6	6	7	5	14
8	2	7	28	1	15
	4	7	14	3	16
	7	7	8	6	17

Note: In the designs that appear below, "Rep." indicates a complete replication of the f factor levels.

DESIGN 1: $f = 4, k = 2, r = 3, b = 6, p = 1$

Block	Rep. I		Block	Rep. II	
(1)	1	2	(3)	1	3
(2)	3	4	(4)	2	4

Block	Rep. III	
(5)	1	4
(6)	2	3

DESIGN 2: $f = 4, k = 3, r = 3, b = 4, p = 2$

Block				Block			
(1)	1	2	3	(3)	1	3	4
(2)	1	2	4	(4)	2	3	4

DESIGN 3: $f = 5, k = 2, r = 4, b = 10, p = 1$

Block	Reps. I, II		Block	Reps. III, IV	
(1)	1	2	(6)	1	4
(2)	3	4	(7)	2	3
(3)	2	5	(8)	3	5
(4)	1	3	(9)	1	5
(5)	4	5	(10)	2	4

DESIGN 4: $f = 5, k = 3, r = 6, b = 10, p = 3$

Block	Reps. I, II, III			Block	Reps. IV, V, VI		
(1)	1	2	3	(6)	1	2	4
(2)	1	2	5	(7)	1	3	4
(3)	1	4	5	(8)	1	3	5
(4)	2	3	4	(9)	2	3	5
(5)	3	4	5	(10)	2	4	5

DESIGN 5: $f = 5, k = 4, r = 4, b = 5, p = 3$

Block	Reps. I–IV			
(1)	1	2	3	4
(2)	1	2	3	5
(3)	1	2	4	5
(4)	1	3	4	5
(5)	2	3	4	5

DESIGN 6: $f = 6, k = 2, r = 5, b = 15, p = 1$

Block	Rep. I	
(1)	1	2
(2)	3	4
(3)	5	6

Block	Rep. II	
(4)	1	3
(5)	2	5
(6)	4	6

Block	Rep. III	
(7)	1	4
(8)	2	6
(9)	3	5

Block	Rep. IV	
(10)	1	5
(11)	2	4
(12)	3	6

Block	Rep. V	
(13)	1	6
(14)	2	3
(15)	4	5

DESIGN 7: $f = 6, k = 3, r = 5, b = 10, p = 2$

Block			
(1)	1	2	5
(2)	1	2	6
(3)	1	3	4
(4)	1	3	6
(5)	1	4	5

Block			
(6)	2	3	4
(7)	2	3	5
(8)	2	4	6
(9)	3	5	6
(10)	4	5	6

DESIGN 8: $f = 6, k = 3, r = 10, b = 20, p = 4$

Block	Rep. I		
(1)	1	2	3
(2)	4	5	6

Block	Rep. II		
(3)	1	2	4
(4)	3	5	6

Block	Rep. III		
(5)	1	2	5
(6)	3	4	6

Block	Rep. IV		
(7)	1	2	6
(8)	3	4	5

Block	Rep. V		
(9)	1	3	4
(10)	2	5	6

Block	Rep. VI		
(11)	1	3	5
(12)	2	4	6

Block	Rep. VII		
(13)	1	3	6
(14)	2	4	5

Block	Rep. VIII		
(15)	1	4	5
(16)	2	3	6

Block	Rep. IX		
(17)	1	4	6
(18)	2	3	5

Block	Rep. X		
(19)	1	5	6
(20)	2	3	4

DESIGN 9: $f = 6, k = 4, r = 10, b = 15, p = 6$

Block	Reps. I, II			
(1)	1	2	3	4
(2)	1	4	5	6
(3)	2	3	5	6

Block	Reps. III, IV			
(4)	1	2	3	5
(5)	1	2	4	6
(6)	3	4	5	6

Block	Reps. V, VI			
(7)	1	2	3	6
(8)	1	3	4	5
(9)	2	4	5	6

Block	Reps. VII, VIII			
(10)	1	2	4	5
(11)	1	3	5	6
(12)	2	3	4	6

Block	Reps. IX, X			
(13)	1	2	5	6
(14)	1	3	4	6
(15)	2	3	4	5

DESIGN 10: $f = 6, k = 5, r = 5, b = 6, p = 4$

Block					
(1)	1	2	3	4	5
(2)	1	2	3	4	6
(3)	1	2	3	5	6
(4)	1	2	4	5	6
(5)	1	3	4	5	6
(6)	2	3	4	5	6

DESIGN 11: $f = 7, k = 2, r = 6, b = 21, p = 1$

Block	Reps. I, II		Block	Reps. III, IV	
(1)	1	2	(8)	1	3
(2)	2	6	(9)	2	4
(3)	3	4	(10)	3	5
(4)	4	7	(11)	4	6
(5)	1	5	(12)	5	7
(6)	5	6	(13)	1	6
(7)	3	7	(14)	2	7

Block	Reps. V, VI	
(15)	1	4
(16)	2	3
(17)	3	6
(18)	4	5
(19)	2	5
(20)	6	7
(21)	1	7

DESIGN 12: $f = 7, k = 3, r = 3, b = 7, p = 1$

Block				Block			
(1)	1	2	4	(5)	5	6	1
(2)	2	3	5	(6)	6	7	2
(3)	3	4	6	(7)	7	1	3
(4)	4	5	7				

DESIGN 13: $f = 7, k = 4, r = 4, b = 7, p = 2$

Block					Block				
(1)	3	5	6	7	(5)	2	3	4	7
(2)	1	4	6	7	(6)	1	3	4	5
(3)	1	2	5	7	(7)	2	4	5	6
(4)	1	2	3	6					

DESIGN 14: $f = 7, k = 6, r = 6, b = 7, p = 5$

Block							Block						
(1)	1	2	3	4	5	6	(5)	1	2	4	5	6	7
(2)	1	2	3	4	5	7	(6)	1	3	4	5	6	7
(3)	1	2	3	4	6	7	(7)	2	3	4	5	6	7
(4)	1	2	3	5	6	7							

DESIGN 15: $f = 8, k = 2, r = 7, b = 28, p = 1$

Block	Rep. I		Block	Rep. II	
(1)	1	2	(5)	1	3
(2)	3	4	(6)	2	8
(3)	5	6	(7)	4	5
(4)	7	8	(8)	6	7
Block	**Rep. III**		**Block**	**Rep. IV**	
(9)	1	4	(13)	1	5
(10)	2	7	(14)	2	3
(11)	3	6	(15)	4	7
(12)	5	8	(16)	6	8
Block	**Rep. V**		**Block**	**Rep. VI**	
(17)	1	6	(21)	1	7
(18)	2	4	(22)	2	6
(19)	3	8	(23)	3	5
(20)	5	7	(24)	4	8
Block	**Rep. VII**				
(25)	1	8			
(26)	2	5			
(27)	3	7			
(28)	4	6			

DESIGN 16: $f = 8, k = 4, r = 7, b = 14, p = 3$

Block	Rep. I			
(1)	1	2	3	4
(2)	5	6	7	8

Block	Rep. II			
(3)	1	2	7	8
(4)	3	4	5	6

Block	Rep. III			
(5)	1	3	6	8
(6)	2	4	5	7

Block	Rep. IV			
(7)	1	4	6	7
(8)	2	3	5	8

Block	Rep. V			
(9)	1	2	5	6
(10)	3	4	7	8

Block	Rep. VI			
(11)	1	3	5	7
(12)	2	4	6	8

Block	Rep. VII			
(13)	1	4	5	8
(14)	2	3	6	7

DESIGN 17: $f = 8, k = 7, r = 7, b = 8, p = 6$

Block							
(1)	1	2	3	4	5	6	7
(2)	1	2	3	4	5	6	8
(3)	1	2	3	4	5	7	8
(4)	1	2	3	4	6	7	8
(5)	1	2	3	5	6	7	8
(6)	1	2	4	5	6	7	8
(7)	1	3	4	5	6	7	8
(8)	2	3	4	5	6	7	8

TABLE 9A.4 Latin-Square Designs for Three Through Eight Levels of Each Factor

(3)

A	B	C
B	C	A
C	A	B

(4)

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

(5)

A	B	C	D	E
B	C	D	E	A
C	D	E	A	B
D	E	A	B	C
E	A	B	C	D

(6)

A	B	C	D	E	F
B	C	D	E	F	A
C	D	E	F	A	B
D	E	F	A	B	C
E	F	A	B	C	D
F	A	B	C	D	E

(7)

A	B	C	D	E	F	G
B	C	D	E	F	G	A
C	D	E	F	G	A	B
D	E	F	G	A	B	C
E	F	G	A	B	C	D
F	G	A	B	C	D	E
G	A	B	C	D	E	F

(8)

A	B	C	D	E	F	G	H
B	C	D	E	F	G	H	A
C	D	E	F	G	H	A	B
D	E	F	G	H	A	B	C
E	F	G	H	A	B	C	D
F	G	H	A	B	C	D	E
G	H	A	B	C	D	E	F
H	A	B	C	D	E	F	G

REFERENCES

Text References

The references in Chapters 4, 6, and 7 can be consulted for additional information on complete and incomplete block designs. The following references augment those of earlier chapters.

Anderson, V. L. and McLean, R. A. (1974). *Design of Experiments: A Realistic Approach*, New York: Marcel Dekker, Inc.

Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*, New York: John Wiley & Sons, Inc.

John, P. W. M. (1980). *Incomplete Block Designs*, New York: Marcel Dekker, Inc.

John, P. W. M. (1971). *Statistical Design and Analysis of Experiments*, New York: Macmillan Co.

The material in Tables 9A.1 and 9A.2 is compiled from [see also Box, Hunter, and Hunter (1978)]

Box, G. E. P. and Hunter, J. S. (1961). "The 2^{k-p} Fractional Factorial Designs, Part I, II," *Technometrics*, **3**, 311–351, 449–458.

EXERCISES

1 An industrial hygienist is concerned with the exposure of workers to chemical vapors during bulk liquid transfer operations on a chemical tanker. The downwind concentration of vapors displaced from loading tanks is being investigated. The industrial hygienist intends to measure the vapor concentrations (ppm) of three different chemicals (benzene, vinyl acetate, and xylene) being loaded into a tanker on five consecutive days. Design an experiment to control for the likely variation in days.

2 Three technicians in a chemical plant are given the task of investigating the chemical yield of a raw material by applying one of five experimental treatments. Three repeat tests are to be made with each of the five treatments. Design an experiment appropriate for this study. Is there a blocking factor in the design? If so, what is it?

3 Design an experiment to investigate deposits on a filter of a cooling system. The factors of interest are:

Flow rate: 5, 10 gps
Filter diameter: 0.5, 1, 2 cm
Fluid temperature: 75, 100, 125°F

A maximum of 20 test runs can be made with each batch of cooling fluid because of impurities that enter the cooling system. It is desired that several batches of cooling fluid be used, subject to a maximum of 75 total test runs.

4 Design an experiment involving three factors, each at two levels, in which no more than ten test runs can be conducted on each of several test days. It is desired that all main effects and interactions be estimated when the experiment is concluded. A maximum of ten test days are available for the experiment.

5 Design an experiment in which two factors are to be investigated, one at two levels and one at three levels. The experiment must be run in blocks, with no more than four test runs per block. Up to 20 blocks can be used in the experiment.

6 A study is to be conducted to assess the effects of kiln firing of clay sculptures on the color of the fired clay. Spectral measurements are to be taken on sculptures made from four different clay compositions. Location variations within the kiln and variations in temperature are also believed to influence the final color. Four locations in the kiln are to be investigated, as are four temperatures. The clay must be made in batches. Each batch can produce a maximum of 20 clay sculptures. No more than 50 clay sculptures can be included in the experiment. Construct a suitable design for this experiment.

7 A study is to be conducted to investigate seven computer algorithms for handling the scheduling of deliveries of parcels by a private delivery company. Because of the wide variety of scheduling scenarios that can arise, a collection of 50 commonly encountered problems have been carefully constructed on which to test the algorithms. A good algorithm is one that can solve the problems rapidly; consequently, the time needed to solve the problems is the key response variable. All the problems are complex and require intensive computing effort to solve. Because of this, it is prohibitively expensive for all the algorithms to solve all the problems. It is decided that approximately 40–50 problems should be used in the experiment, that no more than four of the algorithms should be given any one problem, and that each algorithm must be give at least ten of the problems. Design an experiment to satisfy all these requirements.

8 An experiment is to be designed to study the lift on airplane wings. Two wing models from each of five wing designs are to be included in the study. The wing models are to be placed in a wind tunnel chamber, in which two takeoff angles (5 and 10° from horizontal) can be simulated. The lift (lb) is to be measured twice on each wing at each angle. Design a suitable experiment to study the lift on these wing designs.

9 The U.S. National Weather Service wants to commission an experiment to study the performance of weather balloons made of different materials. Nylon, a polyester blend, and a rayon–nylon blend are the materials to be studied. Six balloons made from each of the materials are to be tested. Two balloons of each material will be loaded with instrumentation having one of three weights (15, 20, and 25 lb). Each balloon will be released at the same geographic location. The response of interest is the height the

balloons rise to before exploding due to atmospheric pressure. Design an appropriate experiment for this study.

10 An experiment is to be designed to determine the amount of corrosion for five different tube materials (stainless steel, iron, copper, aluminum, and brass) in a study similar to the fluidized-bed experiment in Exercise 5 of Chapter 4. Six repeat tests are to be run with each type of tube material. Only three tests can be run per month with the accelerated-aging machine before the equipment that pumps the water through the pipes must be overhauled. Design an experiment for this study.

11 The U.S. Army wishes to study the effects of two methanol fuel blends on the wear rate of iron in jeep engines. Nine jeeps are to be provided for this study. Both fuels are to be used in each jeep. Each fuel is to be used in a jeep for one month. Only three months are available for the study. Thus, three wear measurements will be available from each jeep. Design a study that will control for the sequencing of the fuels in the jeeps.

12 A clinical study is to be conducted on the effectiveness of three new drugs in reducing the buildup of fluid in the ears of infants during winter months. Twenty-five infants are to be included in the four-month study. Each infant is to receive each of the drugs at least once. Lay out an experimental design appropriate for this study.

13 Design an experiment involving three factors, each at two levels, in which no more than six test runs can be conducted on each of several test days. It is desired that all main effects and interactions be estimated when the experiment is concluded. A maximum of five test days are available for the experiment.

14 Consider redesigning the torque study using the factors and levels displayed in Figure 5.1 of Chapter 5. Suppose that only eight test runs can be made on each day of testing. Design an experiment to be conducted on five test-days. List the test runs to be made each day, the confounding, if any, and the resolution of the design.

15 Design an experiment in which a quarter-fraction of a 2^8 complete factorial is conducted in blocks of no more than 20 test runs. The design must be at least of Resolution-V.

16 Refer to the experiment described in Exercise 5 of Chapter 5. Suppose only one plant is to be included in the experiment and that operating restrictions dictate that only four test runs can be made on each of two days. If the experiment is conducted as indicated below, which effect is confounded with days?

DAY 1

Molten-Glass Temperature	Cooling Time	Glass Type
1300	15	A
1100	20	B
1300	20	B
1100	15	A

DAY 2

Molten-Glass Temperature	Cooling Time	Glass Type
1100	15	B
1100	20	A
1300	20	A
1300	15	B

17 In Exercise 10 of Chapter 7, block the constructed fractional factorial into three bocks using the defining contrast $I = A^2B^2C^2D$. Does this defining contrast yield a desirable aliasing pattern?

18 A furniture manufacturer wants to study the floor clearance of a sleeper sofa mechanism. Floor clearance is thought to be related to the type of TV pickup, the pilot arm length, and the type of front tube used. The levels to be tested for each of these three factors are:

	Level		
Factor	−1	0	+1
TV pickup	Rivet	Lever	Pin
Pilot arm (inches)	5.875	6.000	6.125
Front tube	Small	Regular	Large

Each sleeper mechanism in the test will be cycled 3,000 times with floor clearance measurements made on each side of the sleeper at the beginning of the study and then at each 1,000 cycles. Construct a block design with nine test runs per block using $I = ABC$ as the defining contrast. Construct a second block design using $I = AB^2C$ as the defining contrast. Compare the designs graphically using the format of Figure 7.6 denoting the within block runs with a common symbol. How are the two designs similar and how are they different? Which would you prefer, if either? Why?

19 Engineers in a wood bed plant decided to undertake an experiment to determine the dominant source of their quality problems: over spray, sags, and runs in the final lacquer coat. Four factors are of interest in the study, each occurring at three levels:

	Level		
Factor	−1	0	+1
Nozzle distance (inches)	6	9	12
Line speed (seconds/piece)	20	24	28
Lacquer temperature (degrees F)	120	145	170
Spray tip size (inches)	0.0014	0.0015	0.0016

Construct an incomplete block design with three 27-run blocks using the defining contrast $I = AB^2C^2D^2$.

20 A quality manager wants to study the effects of five factor variables on the output of a chemical manufacturing process. The raw material for this process is made in batches. Each batch contains only enough raw material to manufacture four storage tanks of the chemical (a storage tank is the experimental unit). To remove the effects due to differences in batches of raw material, the quality manager decides to conduct the experiment in blocks. Create a 16-run, two-level design in four blocks for this experiment.

CHAPTER 10

Analysis of Designs with Random Factor Levels

Random factor levels often occur in complete and fractional factorial experiments, block designs, nested designs, and designs for process improvement. In this chapter the analyses of statistically designed experiments in which one or more of the experimental factors have random levels are detailed. Because of the variety of analyses possible depending on both the nature of the design and the specific factors in the design that are fixed and random, a comprehensive discussion of parameter estimation methods is beyond the scope of this chapter. When special estimation methods are needed and not discussed, references are provided. The following topics are emphasized in this chapter:

- *the distinction between fixed and random factor effects,*
- *estimation of variance components for random factor effects,*
- *the analysis of data from complete and fractional factorial experiments,*
- *the analysis of data from randomized complete block and incomplete block designs, and*
- *the analysis of data from latin-square and crossover designs.*

In many experiments the levels of one or more of the factors under study are not selected because they are the only ones for which inferences are desired. Random factor levels are intended to be representative of a much larger population of levels. For example, specific vehicles used in an automobile mileage study are representative of a much larger population of vehicles having the same body style, power train, etc. We term factors that have random factor levels *random-effects* factors.

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

Experiments in which one or more of the factors have random effects are ordinarily conducted with the intent to draw inferences not on the specific factor levels included in the experiment, but on the variability of the population or process from which the levels are obtained. Inferences on fixed factor effects are made on means associated with the specific factor levels included in the experiment. Inferences on random factor effects are made on the variances of the factor–level effects.

Analysis-of-variance (ANOVA), parameter and interval estimation, and hypothesis testing procedures for fixed factor effects are detailed in Chapters 6 and 8. Although ANOVA procedures for random factor effects are emphasized in this chapter, we stress that a comprehensive analysis of the data from an experimental design requires the estimation of statistically significant fixed factor effects and, for random effects, of statistically significant variance components. Hence, in addition to the ANOVA procedure modifications needed for random factor effects, we highlight in this chapter modifications of fixed-effects modeling and estimation methods that are needed when one or more of the factors in an experiment are random.

10.1 RANDOM FACTOR EFFECTS

In many experimental programs, the levels of the factors under study are only a sample of all possible levels of interest. For example, in a process capability study, production lots are often one of the factors. The actual lots (levels) chosen for the study are of no particular interest to the engineer; rather, characteristics of all possible lots that might be produced by the process are of importance. In these situations the factors are said to have *random effects* (see Exhibit 10.1).

EXHIBIT 10.1

Random Effects. A factor effect is random if the levels included in the experiment are only a random subset of a much larger population of the possible factor levels. The model terms that represent random factor levels are themselves random variables.

Designs that include factors that have random effects can be identical to those with factors that have fixed effects. Blocking designs and nested designs frequently include one or more random-effects factors. The ANOVA models for random factor effects can be written in a form that appears to be identical to those for fixed factor effects. There are, however, important differences in the assumptions that accompany models having random factor effects (see Exhibit 10.2).

EXHIBIT 10.2 RANDOM-EFFECTS MODEL ASSUMPTIONS

1. The levels of all factors in the experiment represent only a random subset of the possible factor levels of interest.
2. The ANOVA model contains random variables representing main effects and interactions. These random variables are assumed to be statistically independent.
3. The experimental errors are statistically independent of each other and of the random-effects model terms.
4. The model terms for each main effect, each interaction, and the experimental errors are satisfactorily modeled by normal probability distributions, which have zero means but possibly different standard deviations.

Consider a two-factor model in which both factors are random and r repeat tests are conducted on each combination of the factor levels. The model, using Latin letters to denote random effects, would be expressed as follows:

$$y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + e_{ijk}, \qquad \begin{matrix} i = 1, 2, \ldots, a, \quad j = 1, 2, \ldots, b, \\ k = 1, 2, \ldots, r \end{matrix} \tag{10.1}$$

This model is similar in form to a two-factor version of the model (6.1)–(6.2). We now add the following assumptions to complete the definition of the model (10.1):

Model Terms	Distribution	Mean	Standard Deviation
a_i	Normal	0	σ_a
b_j	Normal	0	σ_b
$(ab)_{ij}$	Normal	0	σ_{ab}
e_{ijk}	Normal	0	σ

(10.2)

In addition, we assume that all of the above random variables are statistically independent.

The responses in random-effects models do not have the same distributional properties as those in fixed-effects models. If the factors in the above model were fixed, the responses y_{ijk} would be normally distributed with means

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

and common standard deviation σ. In contrast, when the factor effects are all random, the responses y_{ijk} have a common normal distribution with mean μ and standard deviation

$$\sigma_y = (\sigma^2 + \sigma_a^2 + \sigma_b^2 + \sigma_{ab}^2)^{1/2}. \tag{10.3}$$

It is the four *variance components* in equation (10.3) that contain the information about the populations of factor effects, since the main effects and interaction factor levels are all assumed to be normally distributed with zero means.

The variance terms in (10.3) are called variance components because of the variance of the responses, σ_y^2, is the sum of the individual variances for the main effects (σ_a^2, σ_b^2), interaction effects (σ_{ab}^2), and uncontrolled error (σ^2). Thus, in contrast to fixed factor levels, the random factor levels contribute to the overall variation observable in the response variable, not to differences in factor–level means. Random effects that have large standard deviations can be expected to influence the response more than effects that have small standard deviations.

To focus more clearly on these concepts, reconsider the data presented in Table 4.5 on skin color measurements. Measurements in this table were taken on seven participants during each of three weeks of the study. Note that it is not the specific participants who are of interest to the manufacturer of the skin-care product, but the population of possible purchasers of this product. By modeling the effects of the participants as a normal random variable, the standard deviation of the distribution measures the variability in the measurements that is attributable to different skin types among intended purchasers.

The week-to-week effect can also be modeled as a random effect. Because environmental conditions and other influences — such as daily hygiene, exposure to the sun, etc. — that can affect skin color measurements are not controllable, each individual does not have the same skin color measurement each week of the study. By modeling the weekly effects on skin color as random effects, the standard deviation of its assumed normal distribution measures the contribution of week-to-week variation to the overall variability of the skin color measurements.

One can also consider the inclusion of a random interaction effect between participants and weeks of the study. An interaction component is included if it is believed that weekly effects differ randomly from participant to participant. Thus, random main effects and interaction effects lead to the model (10.1) for the skin color measurements.

10.2 VARIANCE-COMPONENT ESTIMATION

The functions of the model parameters that are estimated by the mean squares in an ANOVA table are called *expected mean squares*. For factorial experiments in which all the factor effects are fixed, consideration of expected mean squares is not of great importance, because all the main effects and interactions in the ANOVA table are tested against the error mean square; that is, all the F-ratios have MS_E in the denominator. This is always not true for

TABLE 10.1 Expected Mean Squares for Typical Main Effects and Interactions: Fixed Factor Effects

Factor Effect	Source of Variation	Typical Mean Square	Expected Mean Square
Main effect	A	MS_A	$\sigma^2 + bcrQ_\alpha$
Interaction	AB	MS_{AB}	$\sigma^2 + crQ_{\alpha\beta}$
Interaction	ABC	MS_{ABC}	$\sigma^2 + rQ_{\alpha\beta\gamma}$
Error	Error	MS_E	σ^2

$$Q_\alpha = \sum_i \frac{\alpha_i^2}{(a-1)}, \quad Q_{\alpha\beta} = \sum_i \sum_j \frac{(\alpha\beta)_{ij}^2}{(a-1)(b-1)}$$

$$Q_{\alpha\beta\gamma} = \textstyle\sum_i \sum_j \sum_k \dfrac{(\alpha\beta\gamma)_{ijk}^2}{(a-1)(b-1)(c-1)}$$

experiments in which all the factors are random or for experiments in which some factor effects are fixed and some are random.

Expected mean squares for fixed-effects ANOVA models all have the same form. Using a three-factor balanced complete factorial experiment for illustration purposes, the general form is expressible as indicated in Table 10.1. Each expected mean square equals the error variance plus a sum of squares of the corresponding effects parameters. These expressions are valid when the model is parametrized as in (6.2) and the zero constraints are imposed.

The sums of squares in ANOVA tables for balanced complete factorial experiments are statistically independent regardless of whether the factors have fixed or random effects. For fixed-effects models, a hypothesis that main-effect or interaction model parameters are zero results in the corresponding expected mean squares simply equaling the error variance. One can show that under such hypotheses the sums of squares divided by the error variance follow chi-square distributions. Consequently, the ratio of the mean square for any of the main effects or interactions to the error mean square follows an F-distribution. This is the justification for the use of F-statistics in Section 6.3 to evaluate the statistical significance of main effects and interactions for fixed-effects models.

Expected mean squares play a fundamental role in the determination of appropriate test statistics for testing hypotheses about main effects and interactions. Expected mean squares determine which hypotheses are tested by each mean square. An F-statistic can only be formed when, under appropriate hypotheses, two expected mean squares have the same value. For fixed-effects

TABLE 10.2 Expected Mean Squares for Typical Main Effects and Interactions: Random Factor Effects

Factor Effect	Source of Variation	Typical Mean Square	Expected Mean Square
Main effect	A	MS_A	$\sigma^2 + r\sigma^2_{abc} + cr\sigma^2_{ab} + br\sigma^2_{ac} + bcr\sigma^2_a$
Interaction	AB	MS_{AB}	$\sigma^2 + r\sigma^2_{abc} + cr\sigma^2_{ab}$
Interaction	ABC	MS_{ABC}	$\sigma^2 + r\sigma^2_{abc}$
Error	Error	MS_E	σ^2

experiments, therefore, the expected mean squares identify the hypotheses being tested with each mean square and reveal that under each hypothesis the effects mean square should be tested using the error mean square in the denominator.

Expected mean squares for random factor effects differ from those of fixed effects in one important respect: they involve the variances of the effects rather than the sums of squares of the effects parameters. This follows because of the distributional assumptions [e.g., (10.2)] that accompany the specification of the random-effects model. Table 10.2 displays the expected mean squares for typical mean squares from a three-factor balanced complete factorial experiment. There are two major differences between Tables 10.1 and 10.2: (a) the replacement of the Q's in Table 10.1 with the variance components in Table 10.2, and (b) the hierarchical nature of the expected mean squares for the random effects in Table 10.2.

The pattern in Table 10.2 should be clear enough to adapt to models that have either fewer or more factors than three. Any effect has components in its expected mean square for its own variance component, all interactions included in the model involving the factor(s) contained in the mean square, and the error variance. The coefficients of the variance components equal the numbers of observations used in the calculation of an average for the effects represented by the variance components. General rules for determining these expected mean squares are provided in the appendix to this chapter.

The difference in expected mean squares between random and fixed-effects models suggests one important difference in the analysis of the two types of experiments. When analyzing experiments having fixed effects, main effects and low-order interactions are not ordinarily evaluated if they are included in a statistically significant higher-order interaction. This is because any

factor–level combinations of interest involving the main effects or low-order interactions can be evaluated from the averages for the higher-order interaction. There is no prohibition from analyzing the main effects or low-order interactions, but it is usually superfluous to do so.

With random-effects experiments, all the main effects and interactions are investigated. The expected mean square for each main effect and each interaction contains a unique variance component. Confidence interval estimation or hypothesis testing procedures should be used to assess whether each of the variance components differs significantly from zero.

The expected mean squares in Table 10.2 indicate that for three-factor random-effects models there are no exact F-tests for the main effects. This is because under the null hypothesis that a main-effect variance component is zero, the expected mean square for that main effect does not equal any of the other expected mean squares in the table. Each two-factor interaction mean square is tested against the three-factor interaction mean square because under a hypothesis that one of the two-factor interaction variance components is zero, the expected mean square for the corresponding interaction effect equals that of the three-factor interaction mean square. The three-factor interaction mean square is again tested against the error mean square.

Because the main effects in three-factor random-effects models and many other models of interest do not possess exact F-tests, approximate procedures have been developed that are similar to the two-sample t-test for means when the two variances are unequal and unknown. A direct extension of the approximate t-statistic (3.16) is Satterthwaite's approximate F-statistic. Details of this approximate F statistic can be found in the references to this chapter.

The expected mean squares for random factor effects provide an important source of information on the estimation of variance components for the factor effects. Estimates of the variance components are obtained by setting expected mean squares for an ANOVA table equal to their respective mean squares and solving for the variance component of interest (see Exhibit 10.3). Thus, estimates of the error variance and the three-factor interaction variance component in a three-factor balanced complete factorial experiment are, from Table 10.2,

$$s_e^2 = \mathrm{MS}_E,$$

$$s_{abc}^2 = \frac{\mathrm{MS}_{ABC} - \mathrm{MS}_E}{r},$$

where we denote the variance component estimate for σ^2 and σ_{abc}^2 by s_e^2 and s_{abc}^2, respectively. Other expected mean squares can be obtained in a similar fashion, by taking appropriate linear combinations of the mean squares.

EXHIBIT 10.3 ESTIMATION OF VARIANCE COMPONENTS

1. Calculate the mean squares for the main effects, interactions, and error terms in the ANOVA model.
2. If the design is a complete factorial with an equal number of repeat tests per factor–level combination, determine the expected mean squares as in Table 10.2 for random-effects models. If the design is nested or if some factors are fixed and others are random, refer to the rules presented in the appendix to this chapter.
3. Equate the mean squares to their expected mean squares and solve the resulting system of simultaneous equations.

Interval estimates for the ratio of two expected mean squares can be found using the methodology presented in Section 3.4. Denote one of the mean squares by s_1^2 and its corresponding expected mean square by σ_1^2. Denote the second mean square by s_2^2 and its expected mean square by σ_2^2. Then the inequality (3.19) provides an interval estimate for the ratio of the expected mean squares.

Interval estimates for ratios of expected mean squares are most often of interest as an intermediate step in the estimation of the ratio of one of the random-effects variance components to the error variance. Once an appropriate interval estimate for the ratio of the main effect or interaction expected mean square to the error variance is obtained, the interval can often be solved for the ratio of the variance component of interest to the error variance.

In Table 10.3, the ANOVA table for the skin color measurements from Table 4.5 is displayed. The F-statistics clearly indicate that the effect of the variability of the participants is statistically significant. The week-to-week variability is not significantly different from the uncontrolled variability of the measurements. Based on the significance of the main effect due to participants, it is now of interest to compare the variability due to the participants with that due to uncontrolled variability.

TABLE 10.3 Analysis-of-Variance Table for the Skin Color Measurements

Source	df	SS	MS	F	E(MS)
Subjects	6	1904.58	317.43	186.17	$\sigma^2 + 3\sigma_a^2$
Weeks	2	2.75	1.37	0.81	$\sigma^2 + 7\sigma_b^2$
Error	12	20.46	1.71		σ^2
Total	20	1927.78			

The estimates of the error and the participant standard deviations are

$$s_e = (1.71)^{1/2} = 1.31,$$

$$s_a = \left(\frac{317.43 - 1.71}{3}\right)^{1/2} = 10.26.$$

The participant standard deviation is almost eight times larger than the uncontrolled error standard deviation. A 95% confidence interval for the ratio of the standard deviations is obtainable from the interval estimate for the ratio of the two expected mean squares:

$$\frac{317.43}{(1.71)(3.73)} < \frac{\sigma^2 + 3\sigma_a^2}{\sigma^2} < \frac{(317.43)(5.37)}{1.71},$$

or

$$49.92 < \frac{\sigma^2 + 3\sigma_a^2}{\sigma^2} < 966.84.$$

Solving this inequality for the ratio of the estimated standard deviations yields

$$4.03 < \sigma_a/\sigma < 17.94.$$

Estimated variance components can be negative or very small in magnitude. The estimated variance component for the week-to-week variation in skin color measurements is negative. From Table 10.3, $s_b^2 = (1.37 - 1.71)/7 = -0.05$. When estimated variance components are negative, one should inspect the data to ensure that outliers are not present. A single outlier in a data set can cause estimates of variance components to be negative. Ordinarily when negative estimates of variance are obtained and the data have been examined to ensure that the negative estimate is not caused by an outlier, the estimated variance is set equal to zero.

Confidence intervals for individual variance components can be obtained, but the procedures for doing so are complicated and beyond the scope of this book. The above interval estimation procedures usually suffice for drawing inferences on variance components for most experimental settings.

Experiments often are conducted in which some of the factors are fixed and others are random. The corresponding ANOVA models are termed *mixed* models. Sometimes the random factors are the primary focus of interest. Other times the fixed factor effects are, but the random factors are included in order to generalize the range of applicability of the results. When this occurs, the variance components are termed *nuisance parameters*. That is, they must be accounted for to correctly test for fixed factor effects, but they are not the primary focus of the investigation. Variances of block effects are occasionally nuisance parameters.

Sums of squares and mean squares for mixed models are calculated using the same formulas as with models having either all fixed effects or all random

effects. As with random-effects models, the F-statistics used to assess main effects and interactions are not all formed with the error mean square in the denominator. Expected mean squares are critically important in the correct use of ANOVA tables for mixed models.

Expected mean squares for mixed models depend not only on the type of design but also on the structure of the ANOVA model that is used to relate the response variable to the design factors. For this reason, the formulation of expected mean squares for many unbalanced designs can only be determined once the design is specified. There are, however, general rules for determining expected mean squares when designs are balanced.

In the appendix to this chapter, rules are listed for the determination of expected mean squares for balanced experimental designs. These rules can be used for either crossed or nested (see Chapter 11) designs. They can be used for fixed, random, or mixed models.

10.3 ANALYSIS OF DATA FROM BLOCK DESIGNS

Block designs often include random block effects whose variances are nuisance parameters. This is because in these settings block designs are used to control known sources of experimental variability. The analysis of data from two general types of block designs is discussed in this section: complete and incomplete block designs. Randomized complete block designs (Section 9.2) consist of a sufficient number of test runs or homogeneous experimental units so that all factor–level combinations of interest can be included in each block. The incomplete block designs (Section 9.3) do not have all factor–level combinations occurring in every block.

10.3.1 Complete Blocks

Randomized complete block (RCB) designs are used as an experimental design technique to control known sources of variability. Variability of test conditions and of experimental units are two primary contributors to the imprecision of test results and, therefore, to the assessment of factor–level effects. RCB designs are effective in controlling such extraneous sources of variability when a block consists of a sufficient number of test runs or homogeneous experimental units so that all factor–level combinations of interest can be tested in each block.

A RCB design is analyzed like a factorial experiment in which one of the factors, the block factor, does not interact with any of the other factors. Thus, the analyses presented in Chapter 6 and Section 10.2, depending on whether the factors are all fixed or some are random, can be directly applied to RCB designs.

An ANOVA model for a randomized complete block design in which there are two fixed experimental factors and one random blocking factor can be written as follows:

$$y_{ijkl} = \mu + a_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} + e_{ijkl}, \tag{10.4}$$

where the first factor represents the random blocking factor. The ANOVA table for the model (10.4) would contain three main effects and one interaction. Because the block factor and the fixed experimental factors do not interact, all the effects are tested against the experimental-error mean square. This can be confirmed by determining the expected mean squares using the rules in the appendix.

10.3.2 Incomplete Blocks

Two types of incomplete block designs were introduced in Chapter 9, those for complete and fractional factorial experiments and those for balanced incomplete blocks. The incomplete block designs themselves are created from fractional factorials in which each fraction is placed in one block. Some or all of the fractions comprise the experiment. The only additional feature of the design is the block effect, which is generally a random effect and which is assumed to have no interactions with the experimental factors. Hence, the analysis of fractional factorials can be applied to this type of incomplete block design if an additional main effect for blocks is included.

This type of analysis would be appropriate for the drill-bit wear experiment introduced in Section 9.3.1. The five design factors shown in Figure 9.2 are of interest in the experiment; however, only eight test runs, one per hour, can be run each day. Because of expected daily changes, the test runs are blocked eight per day according to the signs on the two defining contrasts ABC and CDE shown in Table 9.2. An analysis of variance table for data from this experiment would include the following: main effects for blocks (3 degrees of freedom), main effects for the design factors (5 one-degree-of-freedom main effects), two-factor interactions (10 one-degree-of-freedom interactions), three-factor interactions, (8 one-degree-of-freedom interactions, because 2 three-factor interactions are aliased with blocks), and error (5 degrees of freedom). If three-factor interactions are not believed to exist and are not included in the model, the error sum of squares would have 13 degrees of freedom (5 for error + 8 for three-factor interactions).

Balanced incomplete block (BIB) designs are balanced in the sense that each factor–level combination occurs in exactly r blocks and each pair of factor–level combinations occurs together in exactly p blocks. Viewing the blocking factor as an additional factor in the design, however, results in the design not being balanced. This is because some pairs of blocks have one or more factor–level combinations in common, while others have none in

common. The principle of reduction in error sum of squares (Section 8.1) is used to determine appropriate sums of squares for assessing the block and factor effects.

Data from an experiment conducted using a BIB design are presented in Table 10.4. This experiment was conducted to develop a method of assessing the deterioration of asphalt pavement on a state highway. The proposed measurement scale, a quantitative measurement from 0 (no asphalt left) to 100 (excellent condition), needed to yield comparable results when used by different pavement experts on the same segment of the highway. In addition, the scale needed to be broad enough to cover a wide range of asphalt conditions. Sixteen state district engineers and sixteen 200-foot sections of highway were randomly selected. Each expert traveled six preassigned sites and rated the road segment using the proposed measurement scale. Each site was rated by six different experts.

The BIB design model used in this experiment is a main-effect model in which there are $f = 16$ levels of a random experimental factor and $b = 16$ levels of a random blocking factor:

$$y_{ij} = \mu + f_i + b_j + e_{ij}, \quad i = 1, 2, \ldots, 16, \quad j = 1, 2, \ldots, 16, \qquad (10.5)$$

TABLE 10.4 Balanced Incomplete Block Design for Asphalt-Pavement Rating Study*

		District Engineer															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	1				57	70		55	85	70							72
	2		65						60	55			55		74	70	
	3	65						72	60			85		80		78	
	4			52			66	50		62	48					50	
	5		68					60			63			68	75		61
	6	55		58	57					62				73	55		
	7			56			78		55				53	55			72
Road	8			56		55						50			38	74	56
Segment	9	57	59		50		64									70	58
	10	84				95	96	80					80		95		
	11		58			68	60			70		54		49			
	12	50	54	52		45			57		39						
	13		61	57	57			45				56	56				
	14	60								55	68	88	60				76
	15				76		68		60		58	55			75		
	16				35	40					32		35	38		30	

*Values are ratings for each road segment. Each engineer rated only six road segments. The data were provided by Professor Edward R. Mansfield, Department of Management Science and Statistics, The University of Alabama, Tuscaloosa, AL.

where f_i denotes the effect of the ith engineer and b_j denotes the effect of the jth road segment. Although the model (10.5) is similar to a two-factor random-effects model for a factorial experiment, only the $N = 96$ combinations shown in Table 10.4 are actually included in the model; in particular, the design is not a complete factorial experiment. Note that we again assume that there is no interaction between the block and experimental factors. We denote this model by M_1 in the calculation of the reduction in error sums of squares.

Denote the reduced models that only have the block (road segments) or experimental (engineers) factors by M_2 and M_3, respectively:

$$\begin{aligned} M_2: &\quad y_{ij} = \mu + b_j + e_{ij}, \\ M_3: &\quad y_{ij} = \mu + f_i + e_{ij}. \end{aligned} \tag{10.6}$$

The statistical significance of the engineer and road-segment effects is gauged by obtaining the reductions in sums of squares due to fitting the full model (10.5) instead of the reduced models in (10.6). Denote the error sums of squares for the three models by SSE_1, SSE_2, and SSE_3, respectively. Then the reductions in error sums of squares of interest are:

$$\begin{aligned} &\text{Road Segments}: && R(M_1|M_3) = \text{SSE}_3 - \text{SSE}_1, \\ &\text{Engineers}: && R(M_1|M_2) = \text{SSE}_2 - \text{SSE}_1. \end{aligned}$$

The numbers of degrees of freedom associated with each of these reductions in sums of squares are the differences in those for the error terms:

$$\begin{aligned} \text{df}\{R(M_1|M_3)\} &= \text{df}(\text{SSE}_3) - \text{df}(\text{SSE}_1) \\ &= (N - f) - (N - b - f + 1) = b - 1, \\ \text{df}\{R(M_1|M_2)\} &= \text{df}(\text{SSE}_2) - \text{df}(\text{SSE}_1) \\ &= (N - b) - (N - b - f + 1) = f - 1. \end{aligned}$$

The three ANOVA tables for the above fits are shown in Table 10.5. The "model" sum of squares for the complete model M_1 in Table 10.5(a) measures the combined effects of the engineers and the road segments. Table 10.5(b) and (c) show the fits to the reduced models M_2 and M_3. The above reductions and degrees of freedom are readily obtained from Table 10.5. The two reductions in sums of squares are often summarized in one ANOVA table such as the one shown in Table 10.6. Note that the reductions in sums of squares are not ordinarily additive; that is, the quantities in the sum-of-squares column do not usually add up to the total sum of squares. This is a consequence of

TABLE 10.5 ANOVA Tables for Three Fits to Asphalt Pavement Rating Data

Source	df	SS	MS	F	p-Value
		(a)Full Model: M_1			
Model (engineers and road segments)	30	13,422.13	447.40	7.10	0.000
Error	65	4,098.84	63.06		
Total	95	17,520.97			
		(b) Blocks (Road Segments) Only: M_2			
Road segments	15	11,786.96	785.80	10.96	0.000
Error	80	5,734.00	71.86		
Total	95	17,520.97			
		(c) Experimental Factor (Engineers) Only: M_3			
Engineers	15	2,416.96	161.13	0.85	0.621
Error	80	15,104.01	188.80		
Total	95	17,520.97			

TABLE 10.6 Analysis of Variance Table for BIB Analysis of Asphalt Pavement Rating Data

Source	df	SS	MS	F	p-Value
Road segments	15	11,005.17	733.68	11.63	0.000
Engineers	15	1,635.17	109.01	1.73	0.067
Error	65	4,098.84	63.06		
Total	95	17,520.97			

the imbalance in the design and the use of reductions in sums of squares to measure each of the factor effects.

Testing the significance of the experimental and block factor effects is accomplished through the procedure given in Exhibit 10.4. This procedure is valid regardless of whether the factors are fixed or random. The quantities needed are obtainable from the ANOVA table, Table 10.6.

EXHIBIT 10.4 TESTING THE SIGNIFICANCE OF EXPERIMENTAL AND BLOCK FACTOR EFFECTS

1. Obtain the reductions in error sums of squares for fitting the reduced models M_2 (blocks) and M_3 (experimental factor) by calculating $R(M_1|M_2)$ and $R(M_1|M_3)$.
2. Form the following F ratios:

$$\text{(blocks)} \quad F = \frac{R(M_1|M_3)/(b-1)}{\text{MSE}_1}$$

$$\text{(factor)} \quad F = \frac{R(M_1|M_2)/(f-1)}{\text{MSE}_1}$$

3. If either of the F-statistics exceeds an upper $100\alpha\%$ critical point of the F-distribution with numerator degrees of freedom $\nu_1 = b - 1$ or $f - 1$, respectively, and denominator degrees of freedom $\nu_2 = N - f - b + 1$, then reject the corresponding hypothesis of no block or factor effects.

From the F-statistics and the corresponding significance probabilities shown in Table 10.6, it is apparent that the road segments are highly significant, confirming that the asphalt conditions of the selected road surfaces varied widely. On the other hand, the engineer effects are not statistically significant, indicating that there is not significant disagreement among the engineers in their assessments of the condition of the asphalt in the various road segments.

Expected mean squares for BIB designs cannot be determined from the rules in the appendix. This is because the designs are not balanced in the way described in the appendix; i.e., each factor–level combination does not have an equal number of repeat tests. In this context, a factor–level combination is a block–factor combination. Since some factor levels do not appear in each block, some combinations have one repeat test and others have no repeat tests.

Table 10.7 displays the expected mean squares for random block and factor effects. The multipliers on the variance components are not equal because of the restriction $b \geqslant f$; if $b = f$—the case of *symmetric* BIB designs—the two multipliers are equal. Using these expected mean squares, the estimates of the model standard deviations for the asphalt-pavement rating data are:

$$s_e = 7.94, \quad s_f = 2.94, \quad s_b = 11.21.$$

Thus the estimated variability due to the different road surfaces is approximately 40% larger than that due to the uncontrolled measurement errors.

When the experimental factor and the block factor are fixed effects, the usual response averages, $\bar{y}_{i\bullet}$, are biased estimators of the average responses

TABLE 10.7 Expected Mean Squares for Random Effects in Balanced Incomplete Block Designs

Effect	Mean Square	Expected Mean Square*
Blocks (adj.)	$\frac{R(M_1\|M_3)}{b-1}$	$\sigma^2 + \frac{bk-f}{b-1}\sigma_b^2$
Factor (adj.)	$\frac{R(M_1\|M_2)}{f-1}$	$\sigma^2 + \frac{pf}{k}\sigma_f^2$

*For symmetric BIB designs, $b = f$ and $k = r$. Then $(bk - f)/(b - 1) = pf/k$.

for the factor levels. This is because each factor level (or factor–level combination) does not occur in every block. Thus, the responses for the factor levels contain the effects of some, but not all, of the blocks. Even if the block effects are random, it is desirable to adjust the averages for fixed-factor effects to eliminate differences due to block effects from comparisons of the factor levels. These adjusted factor–level averages are calculated as shown in Exhibit 10.5.

EXHIBIT 10.5 ADJUSTED FACTOR–LEVEL AVERAGES

$$m_i = \bar{y}_{\bullet\bullet} + c(\bar{y}_{i\bullet} - \bar{y}^{(i)}), \tag{10.7}$$

where

c = (N − r)/(N − b),
$\bar{y}_{i\bullet}$ = average response for factor–level combination i;
$\bar{y}^{(i)}$ = average of all observations from blocks that contain treatment i (this is an average of r block averages);
$\bar{y}_{\bullet\bullet}$ = overall average.

The multiplication by c in Equation (10.7) can be viewed as correcting for the number of times a factor–level combination appears (i.e., the number of blocks in which a factor–level combination occurs) relative to the total number of blocks in the experiment, b. Aside from this correction term, Equation (10.7) is the overall average plus the difference between the usual factor–level response average and the average response for the blocks in which the factor level occurs. If the raw response average for a factor level is greater (less) than the average for the blocks in which it occurs, the overall average is adjusted upward (downward). If each factor–level combination appears

TABLE 10.8 Adjusted Factor Averages for Asphalt Paving Scale Study

Blocking Factor		Experimental Factor		
Road Segment	Average Rating	Engineer	Average Rating	Adjusted Rating
1	68.17	1	61.83	56.21
2	63.17	2	60.83	63.58
3	73.33	3	55.17	60.49
4	54.67	4	55.33	59.24
5	65.83	5	62.17	64.65
6	60.00	6	72.00	69.39
7	61.50	7	60.33	53.21
8	54.83	8	62.83	60.65
9	59.67	9	62.33	61.40
10	88.33	10	51.33	55.74
11	59.83	11	64.67	63.55
12	49.50	12	56.50	55.36
13	55.33	13	60.50	62.80
14	67.83	14	68.67	64.11
15	65.33	15	62.00	67.27
16	35.00	16	65.83	64.61

in every block (that is, an RCB design) then $c = 1, \overline{y}^{(i)} = \overline{y}_{\bullet\bullet}$ for all i, and consequently $m_i = \overline{y}_{i\bullet}$, the usual response average.

Table 10.8 shows the values of $\overline{y}_{\bullet j}$, $\overline{y}_{i\bullet}$, and m_i for the data from the asphalt-measurement experiment. Although the experimental and block factors in this experiment are random effects, we present these summary statistics to illustrate the changes that can occur in some of the averages due to the imbalance of the BIB design. Note that some of the adjusted factor–level averages m_i differ by as much as 10% from the corresponding raw averages $\overline{y}_{i\bullet}$. Were these factor levels fixed, the adjusted averages would be the appropriate statistics to use to compare the factor–level effects.

The overall F-ratio using the reduction in sum of squares $R(M_1|M_3)$ can be used to ascertain that all the fixed-effect means are not equal. The experimenter will also be keenly interested in identifying which individual factor–level means are different from one another. The estimated standard errors of the adjusted averages can be used with the procedures discussed in Chapter 6 to make this determination. The standard-error estimates can be calculated according to the following formulas:

$$\mathrm{SE}(m_i) = s_e\left(N^{-1} + \frac{(k-1)c^2}{rk}\right)^{1/2}, \tag{10.8}$$

$$\mathrm{SE}(m_i - m_k) = s_e(2c/r)^{1/2},$$

where $s_e^2 = \mathrm{MSE}_1$ is the mean squared error for the full model fit and c is defined in (10.7). The standard errors given in (10.8) can also be used to construct confidence intervals.

10.4 LATIN-SQUARE AND CROSSOVER DESIGNS

Latin-square designs (Section 9.4) are block designs that control two extraneous sources of variability. A condition for the use of latin-square designs is that there must be no interactions between the design factor(s) and the two blocking factors. Because of this restriction, the analysis of data from latin-square designs utilizes main-effect ANOVA models.

The layout for a latin-square design used to road-test four different brands of commercial truck tires is given in Table 9.5. The mileage-efficiency data collected for this study are given in Table 10.9. An ANOVA model containing only main effects for these data can be expressed as follows:

$$y_{ijk} = \mu + a_i + b_j + \gamma_k + e_{ijk}, \qquad i = 1,2,3,4, \quad j = 1,2,3,4, \quad k = 1,2,3,4. \tag{10.9}$$

Note that although each of the subscripts in the above model ranges from 1 to 4, only the specific test runs shown in Tables 9.5 and 10.9 are included in the data set; in particular, this is not a complete factorial experiment.

In this example, the trucks (a_i) and days (b_j) are random effects. Neither the specific trucks nor the specific days on which the testing was conducted are of primary interest to the experimenters. Both sets of effects can be considered randomly selected from conceptually infinite populations of truck and day effects. Inferences are desired, however, on the specific tire brands (γ_k) used in the tests. For this reason, tire brands are considered fixed effects.

The ANOVA table for these data using the model (10.9) is given in Table 10.10. The sums of squares are calculated in the usual way for main effects; for example, $\mathrm{SS}_A = k^{-1}\sum y_{i\bullet\bullet}^2 - (k^2)^{-1}y_{\bullet\bullet\bullet}^2$, where k is the number of levels of each of the factors (see Table 6.2). Under either fixed or random-effects model assumptions, the expected mean squares for the main effects are the experimental-error variance plus a nonnegative function of the respective model parameters or variance components, that is, $\sigma^2 + k\sum\alpha_i^2/(k-1)$ or $\sigma^2 + k\sigma_a^2$. The expected mean square for error is the experimental-error variance. Therefore, the appropriate F-ratios for assessing factor effects for this

TABLE 10.9 Mileage Efficiencies for Commercial-Truck-Tire Experiment

Truck	Day	Tire Brand	Efficiency (mi/gal)
1	1	2	6.63
	2	1	6.26
	3	4	7.31
	4	3	6.57
2	1	4	7.06
	2	2	6.59
	3	3	6.64
	4	1	6.71
3	1	1	6.31
	2	3	6.73
	3	2	7.00
	4	4	7.38
4	1	3	6.55
	2	4	7.13
	3	1	6.54
	4	2	7.11

TABLE 10.10 ANOVA Table for Truck-Tire Experiment

Source	df	SS	MS	F	p-Value
Trucks	3	0.0676	0.0225	1.46	0.317
Days	3	0.2630	0.0877	5.67	0.034
Tires	3	1.3070	0.4357	28.17	0.000
Error	6	0.0928	0.0155		
Total	15	1.7305			

example, or for data from any latin-square design, are formed by taking the ratio of each of the main-effect mean squares to the error mean square.

From Table 10.10 it is seen that the tire-brand effects are statistically significant ($p < 0.001$), as are the day effects ($p = 0.034$). One would now proceed to estimate the standard deviations of the error and day effects (Section 10.2) and determine which specific tire brands are significantly affecting the mileage efficiency (Chapter 6). We leave the inferences to the exercises.

Crossover designs (Section 9.4) are appropriate when different levels of a factor are given or applied to an experimental unit in a time sequence. A

crossover design is in effect a blocking design. Each experimental unit is a block and receives every factor–level combination in a time sequence. In this manner each experimental unit acts as its own control, and variation among experimental units has no bearing on comparisons of factor effects.

Under the simplifying assumptions given in Section 9.4, the analysis of crossover designs involves fitting an ANOVA model with block terms, time-period terms, factor effect terms, and experimental error. If more than one block receives the same sequence of factor–level combinations over the time periods, then a sequence term is added to the model and blocks are nested within a sequence. If the same number of blocks are nested within each sequence, then the design is balanced; otherwise, it is unbalanced. When the design is balanced, the usual sum-of-squares calculations can be performed on the factors. When it is unbalanced, the principle of reduction in error sums of squares (Section 8.1) is used to obtain the sums of squares.

APPENDIX: DETERMINING EXPECTED MEAN SQUARES

Expected mean squares for most balanced experimental designs can be determined using the following rules. These rules apply to crossed factors when there are an equal number of repeat tests for each combination. They also apply to nested factors (see Chapter 11) when there are an equal number of levels at each stage of nesting. These rules can be applied to models that have fixed effects, random effects, or a combination of each.

1. Write out an ANOVA model for the experimental design. Impose constraints that the sum over any subscript of the fixed-effect parameters is zero. Interactions of fixed and random factors are considered random effects.
2. Label a two-way table with
   (a) A column for each of the subscripts in the ANOVA model and
   (b) A row for each term of the model (except the constant), with the last row being the error term expressed as a nested factor.
3. Proceed down each column labeled by a subscript that represents a fixed effect.
   (a) Enter a 0 if the column subscript appears in a fixed row effect (or a 1 if the column subscript appears in a random row effect) and if no other subscripts are nested within it.
   (b) Enter a 1 if the column subscript appears in the row effect and one or more other subscripts are nested within it.

(c) If the column subscript does not appear in the row effect, enter the number of levels of the factor corresponding to the subscript.

4. Proceed down each column labeled by a subscript that represents a random effect.
   (a) Enter a 1 if the column subscript appears in the row effect, regardless of whether any other effects are nested within it.
   (b) If the column subscript does not appear in the row effect, enter the number of levels of the factor corresponding to the subscript.
5. For a main effect or interaction consisting of only fixed effects factors, let $\phi = Q$, the mean square of the effects parameters. For a main effect or interaction consisting of one or more random effects factors, let $\phi = \sigma^2$, the variance component for the random effect. List the ϕ parameters in a separate column to the right of the two-way table, with each ϕ parameter on the same line as its corresponding model term.
6. Select a mean square from the ANOVA table. Let MS denote the mean square and C the set of subscripts on the corresponding model term.
   (a) Identify the ϕ parameters whose corresponding model terms contain all of the subscripts in C.
   (b) Determine the multiplier for each ϕ parameter by
      (i) eliminating the columns in the two-way table that correspond to each of the subscripts in C, and
      (ii) taking the product of the remaining elements in the row of the two-way table in which the ϕ parameter occurs.
   (c) The expected mean square for MS is the linear combination of ϕ values identified in (a) with coefficients of the linear combination determined by (b). The expected mean square for MS_E is σ^2.
7. Repeat Step 6 for each mean square in the ANOVA table.

To illustrate the application of the above rules, consider the following model

$$Y_{ijkl} = \mu + \alpha_i + b_{j(i)} + \gamma_k + (\alpha\gamma)_{ik} + e_{ijkl},$$
$$i = 1, \ldots, a \quad j = 1, \ldots, b \quad k = 1, \ldots, c \quad l = 1, \ldots, r.$$

This model has two fixed effects factors (A and C) and one random nested factor (B). The levels of factor B are nested within the levels of factor A. The two-way table of multipliers for the determination of the expected mean squares is as follows:

	i	j	k	l	ϕ
α_i	0	b	c	r	Q_α
$\beta_{j(i)}$	1	1	c	r	σ_b^2
γ_k	a	b	0	r	Q_γ
$(\alpha\gamma)_{ik}$	0	b	0	r	$Q_{\alpha\gamma}$
$e_{l(ijk)}$	1	1	1	1	σ^2

Using these multipliers, the expected mean squares are:

Effect	**Expected Mean Square**
A	$\sigma^2 + cr\sigma_b^2 + bcrQ_\alpha$
$B(A)$	$\sigma^2 + cr\sigma_b^2$
C	$\sigma^2 + abrQ_\gamma$
AC	$\sigma^2 + brQ_{\alpha\gamma}$
Error	σ^2

REFERENCES

Text References

Most of the references at the end of Chapters 4 and 6 provide discussions of random and mixed models. Expected mean squares are discussed in the texts by Ostle and Malone and by Johnson and Leone (cited at end of Chapters 4 and 6). The book by Milliken and Johnson cited in Chapter 6 provides a comprehensive discussion of the estimation of fixed factor effects and of variance components for unbalanced designs. The rules for expected mean squares are adapted from

Bennett, C. A. and Franklin, N. L. (1954). *Statistical Analysis in Chemistry and the Chemical Industry*, New York: John Wiley and Sons, Inc.

Cornfield, J. and Tukey, J. W. (1956). "Average Values of Mean Squares in Factorials," *Annals of Mathematical Statistics*, **27**, 907–949.

Satterthwaite approximations are discussed in most of the references listed at the end of Chapter 6. Satterthwaite's version of this approximation was first introduced in the following article:

Satterthwaite, F. E. (1946). "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin*, **2**, 110–114.

EXERCISES

1 Use the rules given in the appendix to this chapter to derive expected mean squares for the torque study described in Section 5.1 (see Figure 5.1 and Table 5.3).

2 Assume that the time effect for the experiment described in Exercise 9 of Chapter 6 is random. Derive the expected mean squares for the factor main effects. Reanalyze the data and compute a confidence interval for the ratio of the standard deviation of the time effects to that of the uncontrolled experimental error.

3 Examine the data on the chemical-analysis variation study in Table 1.1. Which factors are fixed and which are random? Write an ANOVA model for the responses. Derive the expected mean squares for all the effects listed in your model.

4 Analyze the data in Table 1.1 using the ANOVA model from Exercise 3. Perform mean comparisons on any fixed effects that are statistically significant. Estimate all the variance components in the model, regardless of whether they are statistically significant. Do the magnitudes of the estimated variance components conform to what should be expected from the ANOVA results? Explain your conclusion.

5 Analyze the fuel-economy data in Table 2.1. Include mean comparisons and variance-component estimation where appropriate. Reanalyze the data with the outlier on the Volkswagen discarded. How did the deletion of the outlier affect the results of the analysis?

6 A skin sensitivity study was conducted using two products, one a placebo (control) and the other a potentially allergenic product (treatment). Three animals were exposed to the control product and three to the treatment product. Each animal was administered the products on the ear and on the back. Duplicate tests were run. The data are tabulated below. Analyze these data by constructing an ANOVA table, including the expected mean squares. Perform mean comparisons and estimate variance components where appropriate.

		Measurements	
Group	**Animal**	**Ear**	**Back**
Control	1	73, 70	90, 92
	2	66, 66	83, 83
	3	71, 70	77, 74
Treatment	4	77, 74	89, 88
	5	67, 70	70, 78
	6	76, 77	96, 96

7 Plastic viscosity is an important characteristic of a lubricant used to reduce friction between a rotating shaft and the casing that surrounds it. Measurements of plastic viscosity were taken on two lubricants. Four batches of each fluid were used in the tests, with three samples of fluid taken from each batch. The data are given below. Analyze these data with specific reference to how the consistency of the plastic-viscosity measurements is affected by each of the design factors.

		Viscosity			
Lubricant	**Sample**	**Batch 1**	**2**	**3**	**4**
Xtra Smooth	1	4.86	5.87	6.07	4.97
	2	4.34	5.07	5.76	5.42
	3	4.64	5.38	5.22	5.77
EZ Glide	1	1.45	1.78	2.08	2.07
	2	1.26	1.69	2.41	2.19
	3	1.67	1.35	2.27	2.44

8 Treat the engineer effects for the asphalt-pavement rating study (Table 10.4) as fixed effects. Use Tukey's TSD procedure on the adjusted averages in Table 10.8 to determine which of the sixteen engineers have unusually high or low mean ratings.

9 Complete the analysis of the truck-tire data in Table 10.9 by determining which of the tire brands have significantly higher mean efficiencies than the others.

10 Place a 99% confidence interval on the difference of the mean response for the two drugs in the two-period crossover design data given below. Construct a 95% confidence interval on the ratio of the standard deviations for the subject and the uncontrolled error effects.

Subject	Sequence	Time Period	Drug	Response
1	1	1	1	7.2
		2	2	9.0
2	1	1	1	7.2
		2	2	8.0
3	2	1	2	10.2
		2	1	9.2
4	2	1	2	20.8
		2	1	15.6
5	2	1	2	11.2
		2	1	9.0
6	1	1	1	16.4
		2	2	20.9

11 A traffic engineer is interested in comparing the total unused red-light time for five different traffic-light signal sequences. The experiment was conducted with a latin-square design in which the two blocking factors are (a) five randomly selected intersections and (b) five time periods. Analyze this data set to determine if the signal sequences (denoted by Roman numerals) are statistically significant. Values in parentheses represent unused red-light time, the response variable, in minutes.

Intersection	Period (1)	(2)	(3)	(4)	(5)
1	I (15.2)	II (33.8)	III (13.5)	IV (27.4)	V (29.1)
2	II (16.5)	III (26.5)	IV (19.2)	V (25.8)	I (22.7)
3	III (12.1)	IV (31.4)	V (17.0)	I (31.5)	II (30.2)
4	IV (10.7)	V (34.2)	I (19.5)	II (27.2)	III (21.6)
5	V (14.6)	I (31.7)	II (16.7)	III (26.3)	IV (23.8)

12 A new technique was evaluated to determine if a simple low-cost process could prevent the generation of hazardous airborne asbestos fibers during the removal of friable insulation material from buildings. The new technique involved spraying an aqueous saturating solution in the room as the asbestos was being removed. It was necessary to determine if different lengths of spraying time had any effects on the fiber count in the room after the insulation had been removed. Ten buildings were chosen in which to test the spraying technique. Five spraying times were investigated: 90, 60, 50, 40, and 30 minutes. It was determined that only three rooms from each building could be used in the experiment. The results are given below, with the spraying times listed for each room and the fiber count (the response of interest) shown in brackets:

	Spraying Time [Fiber Count]		
Building	Room 1	Room 2	Room 3
1	90 [8]	40 [12]	30 [14]
2	60 [6]	50 [10]	40 [12]
3	30 [11]	60 [8]	90 [9]
4	40 [12]	50 [13]	30 [15]
5	50 [11]	90 [9]	60 [10]
6	50 [8]	40 [9]	90 [7]
7	60 [7]	30 [10]	40 [12]
8	50 [10]	90 [7]	30 [12]
9	60 [9]	40 [12]	90 [10]
10	60 [9]	50 [12]	30 [12]

Analyze this BIB design using the principle of reduction in error sum of squares. Are the spraying times statistically significant? Calculate the adjusted factor averages for the spraying times. Which spraying times have the lowest mean fiber counts?

13 Data were collected in the crossover design experiment described in Chapter 9, Exercise 11. Analyze the data and draw appropriate conclusions.

Jeep	Sequence	Month	Methanol	Iron Wear Rate
1	4	1	2	0.73
		2	1	0.30
		3	1	0.25
2	2	1	1	0.27
		2	2	0.56
		3	2	0.72
3	2	1	1	0.35
		2	2	0.43
		3	2	0.38
4	1	1	1	0.22
		2	2	0.56
		3	1	0.17
5	4	1	2	0.67
		2	1	0.15
		3	1	0.08
6	3	1	2	0.82
		2	1	0.26
		3	2	0.73
7	1	1	1	0.30
		2	2	0.64
		3	1	0.61
8	3	1	2	0.42
		2	1	0.20
		3	2	0.30
9	2	1	1	0.13
		2	2	0.56
		3	2	0.69

14 An experiment was run to help determine the laboratory protocol for the study of allergic reactions to environmental pollutants. Ten mice from a single strain were injected with a known pollutant in two locations, the back and shoulder. Repeat observations at both locations were taken immediately prior to and 45 minutes following the injection. The data obtained are shown below. What is the blocking factor? Analyze the data to determine the location and time effects. What location should be used for the laboratory protocol? Why?

	Back		Shoulder	
Animal	Initial	45 minutes	Initial	45 minutes
1	69	75	70	77
	68	80	86	76
2	66	77	66	74
	66	77	82	83
3	77	82	77	86
	77	87	76	87
4	76	78	70	74
	76	76	67	82
5	66	80	86	63
	66	80	62	79
6	82	96	82	95
	82	91	82	94
7	65	74	70	71
	62	77	80	88
8	61	73	63	72
	62	70	63	71
9	78	94	79	92
	77	90	80	88
10	72	78	66	82
	74	76	66	82

15 A study was undertaken to understand the effect of wiredrawing lubricant on the efficiency of wiredrawing machines. Eight different lubricants were included in the study. Two blocking factors were used—the steel rod supplier and the wiredrawing machine. A latin-square design was used and the following data were obtained (the efficiency values are in parentheses after the lubricant number):

Rod Supplier	Drawing Machine 1	2	3	4	5	6	7	8
1	1 (87.9)	2 (89.1)	3 (88.6)	4 (87.2)	5 (85.1)	6 (88.9)	7 (87.7)	8 (87.2)
2	2 (88.4)	3 (89.9)	4 (89.7)	5 (88.4)	6 (89.2)	7 (85.9)	8 (88.7)	1 (87.2)
3	3 (88.1)	4 (88.8)	5 (88.8)	6 (86.2)	7 (85.7)	8 (89.0)	1 (83.4)	2 (88.1)
4	4 (89.5)	5 (89.1)	6 (88.2)	7 (85.4)	8 (81.0)	1 (87.2)	2 (89.5)	3 (88.5)
5	5 (88.0)	6 (88.8)	7 (86.5)	8 (90.2)	1 (84.3)	2 (90.0)	3 (89.6)	4 (89.2)
6	6 (88.1)	7 (87.0)	8 (90.0)	1 (87.0)	2 (85.1)	3 (88.5)	4 (88.7)	5 (88.4)
7	7 (87.7)	8 (89.1)	1 (87.6)	2 (85.8)	3 (84.2)	4 (89.0)	5 (88.9)	6 (89.0)
8	8 (89.9)	1 (87.0)	2 (89.1)	3 (87.6)	4 (88.6)	5 (88.5)	6 (87.0)	7 (86.7)

Analyze the data to see if the lubricants are statistically different. Which lubricant is preferable (assume the costs of the eight lubricants are similar)?

16 Suppose the eight treatments in Exercise 15 correspond to the eight factor combinations from a 2^3 factorial experiment with factors A, B, and C where the factors are chemical properties of the lubricants. The treatment numbers correspond to the following factor combinations:

Treatment	A	B	C
1	−1	−1	+1
2	−1	+1	−1
3	+1	+1	+1
4	+1	+1	−1
5	+1	−1	+1
6	−1	−1	−1
7	−1	+1	+1
8	+1	−1	−1

Using this additional information, reanalyze the data in Exercise 15. Display any significant main or interaction effects graphically. What does your analysis say about lubricant design for wiredrawing machines?

17 An experiment was conducted to compare four different fermentation processes: F_1, F_2, F_3, and F_4. An organic raw material is common to each process and can be made in batches that are adequate for four runs. This raw material exhibits substantial variation batch-to-batch. A block design was used with the following results (the response is a measure of fermentation efficiency and is measured in percent):

	Process			
Batch	F_1	F_2	F_3	F_4
1	84	83	92	89
2	79	72	87	74
3	76	82	82	80
4	82	97	84	79
5	74	76	75	83

Analyze the data, and decide if there is statistical evidence that one of the processes is more efficient than the others.

18 Test results on yield (grams) from a chemical process are as follows:

Temperature	Concentration	Catalyst	Yield
55	10	C	30
55	10	M	26
55	5	C	27
55	5	M	22
80	10	C	36
80	10	M	41
80	5	C	34
80	5	M	40

Assess the main effects and two-factor interaction effects for this factorial experiment using numerical and graphical techniques. Now assume that the experiment was run in two incomplete blocks with the blocks defined by $I = ABC$. Reanalyze the experiment taking into account the blocking structure. Does this change the conclusions of your previous analysis?

19 The following data for the over spray response were obtained for the experiment described in Exercise 19 of Chapter 9:

Nozzle Distance	Line Speed	Lacquer Temp	Tip Size	Block	Over Spray
6	20	120	0.0014	1	1.5
6	20	145	0.0016	1	3.2
6	20	170	0.0015	1	5.5
6	24	120	0.0016	1	4.5
6	24	145	0.0015	1	5.4
6	24	170	0.0014	1	5.6
6	28	120	0.0015	1	3.0
6	28	145	0.0014	1	1.3
6	28	170	0.0016	1	3.3

Nozzle Distance	Line Speed	Lacquer Temp	Tip Size	Block	Over Spray
9	20	120	0.0015	1	2.4
9	20	145	0.0014	1	2.1
9	20	170	0.0016	1	2.8
9	24	120	0.0014	1	1.0
9	24	145	0.0016	1	4.4
9	24	170	0.0015	1	7.8
9	28	120	0.0016	1	1.4
9	28	145	0.0015	1	1.4
9	28	170	0.0014	1	5.3
12	20	120	0.0016	1	3.5
12	20	145	0.0015	1	2.1
12	20	170	0.0014	1	1.0
12	24	120	0.0015	1	5.5
12	24	145	0.0014	1	3.9
12	24	170	0.0016	1	2.7
12	28	120	0.0014	1	1.1
12	28	145	0.0016	1	1.0
12	28	170	0.0015	1	2.8
6	20	120	0.0016	2	1.0
6	20	145	0.0015	2	5.5
6	20	170	0.0014	2	5.8
6	24	120	0.0015	2	2.4
6	24	145	0.0014	2	4.4
6	24	170	0.0016	2	3.1
6	28	120	0.0014	2	1.0
6	28	145	0.0016	2	5.1
6	28	170	0.0015	2	6.9
9	20	120	0.0014	2	1.0
9	20	145	0.0016	2	4.0
9	20	170	0.0015	2	3.6
9	24	120	0.0016	2	1.5
9	24	145	0.0015	2	2.7
9	24	170	0.0014	2	4.5
9	28	120	0.0015	2	2.4
9	28	145	0.0014	2	4.4
9	28	170	0.0016	2	5.6
12	20	120	0.0015	2	4.8
12	20	145	0.0014	2	1.0
12	20	170	0.0016	2	6.0
12	24	120	0.0014	2	3.3
12	24	145	0.0016	2	4.0
12	24	170	0.0015	2	2.3

12	28	120	0.0016	2	1.0
12	28	145	0.0015	2	1.3
12	28	170	0.0014	2	3.2
6	20	120	0.0015	3	1.6
6	20	145	0.0014	3	1.0
6	20	170	0.0016	3	2.8
6	24	120	0.0014	3	1.0
6	24	145	0.0016	3	4.3
6	24	170	0.0015	3	3.4
6	28	120	0.0016	3	1.0
6	28	145	0.0015	3	2.3
6	28	170	0.0014	3	3.9
9	20	120	0.0016	3	3.5
9	20	145	0.0015	3	2.4
9	20	170	0.0014	3	4.7
9	24	120	0.0015	3	3.2
9	24	145	0.0014	3	2.2
9	24	170	0.0016	3	1.1
9	28	120	0.0014	3	1.0
9	28	145	0.0016	3	2.8
9	28	170	0.0015	3	6.8
12	20	120	0.0014	3	1.0
12	20	145	0.0016	3	3.8
12	20	170	0.0015	3	4.8
12	24	120	0.0016	3	1.5
12	24	145	0.0015	3	2.6
12	24	170	0.0014	3	1.9
12	28	120	0.0015	3	3.9
12	28	145	0.0014	3	4.1
12	28	170	0.0016	3	2.9

Analyze these data to determine the effects of the four factors on over spray. Because each of the factor levels are quantitative, explain any significant effects in terms of linearity and curvature. What would be the optimum process setting to minimize over spray?

20 Suppose in Exercise 19 that only a one-third fraction of the complete factorial was run. Use the fraction corresponding to the defining equation $I = AB^2C^2D^2$ and $T = 0 \bmod(3)$ [see Exhibit 7.5]. Now assume that this twenty-seven run fractional factorial was run in three incomplete blocks defined by $I = ABCD^2$. Analyze the data from this experiment using the regression analysis methods described in Chapter 15. Compare your conclusions to those of Exercise 19.

CHAPTER 11

Nested Designs

Frequently experimental situations require that unique levels of one factor occur within each level of a second factor. This nesting of factors can also arise when an experimental procedure restricts the randomization of factor–level combinations. In this chapter we discuss nested experimental designs, with special emphasis on:

- *the distinction between crossed and nested factors,*
- *designs for hierarchically nested factors,*
- *staggered nested designs in which the number of levels of a factor can vary within a stage of the nesting, and*
- *split-plot designs, which include both nested and crossed factors.*

Consider an experiment that is to be conducted using three samples of a raw material from each of two vendors. In such an experiment there is no physical or fundamental relationship between the samples labeled 1, 2, and 3 from each vendor. In experimental design terminology, the factor "samples" would be said to be nested within the factor "vendor." A design that includes nested factors is referred to as a nested design.

Experiments conducted to diagnose sources of variability in manufacturing processes or in a laboratory method often use nested designs. Experiments in which subjects, human or animal, are given different experimental treatments have nested factors if only one of the treatments is given to each subject. Nested designs also occur when restrictions of cost or experimental procedure require that some factor–level combinations be held fixed while combinations involving other factors are varied.

The first section of this chapter clarifies the main differences between nested and crossed factors. The second section is devoted to a discussion of hierarchically nested designs in which each factor is progressively nested within

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

all preceding factors. The third section discusses split-plot designs in which some of the factors are nested and others are crossed. The final section shows how certain types of restricted randomization result in split-plot designs. The statistical analysis of data from nested designs is discussed in Chapter 13.

11.1 CROSSED AND NESTED FACTORS

Crossed factors occur in an experiment when each level of the factor(s) has a physical or fundamental property that is the same for every level of the other factors in the experiment (see Exhibit 11.1). Complete factorial experiments, by definition, involve crossed factors because each level of each factor occurs in the experiment with each level of every other factor. Crossed factors are often used in an experiment when the levels of the factors are the specific values of the factor for which inferences are desired. In preceding chapters the levels of factors such as pipe angle (Table 4.2), soil treatment (Figure 4.2), catalyst (Table 4.8), alloy (Figure 5.1), reactor feed rate (Table 7.1), manufacturer (Table 9.1), oil (Table 9.4), and many others are crossed factors.

EXHIBIT 11.1

Crossed Factors. A crossed factor contains levels that have a physical or fundamental property that is the same for all levels of the other factors included in the experiment.

In contrast to crossed factors, nested factors (see Exhibit 11.2) have levels that differ within one or more of the other factors in the experiment. Nested factors often are included in experiments when it is desired to study components of response variation that can be attributable to the nested factors. Variation in chemical analyses (Table 1.1) and in lawnmowers (Table 9.1) are two examples of nested factors used in previous chapters. In each of these examples the factor level is simply an identifying label.

EXHIBIT 11.2

Nested Factor. A nested factor has unique levels within each level of one or more other factors.

As an example of an experiment where factors are nested, consider an investigation to determine the major sources of variability in a laboratory test for measuring the density of polyethylene. The test method requires that polyethylene pellets be heated, melted, and pressed into a plaque. A small disk

is then cut from the plaque, and the density of the disk is measured. The test is nondestructive, so that more than one measurement can be made on each disk. Data are to be collected for two shifts of technicians working in a laboratory. Each shift is to prepare two plaques and four disks from each plaque.

There are three factors of interest in this experiment: shift, plaque, and disk. Because the plaques made on one shift have no physical relationship with those made on another shift, plaques are nested within shifts. Similarly, because the disks cut from one plaque have no physical relationship with those cut from another plaque, disks are nested within plaques. Note that disk 1, 2, etc. and plaque 1, 2, etc. are only labels. It is the variation in the density measurements attributable to these plaques and disks that is of concern, not any inferences on the individual plaques and disks used in the experiment.

Figures 11.1 and 11.2 depict the difference between crossed and nested factors. In Figure 11.1 the factor levels for the torque study (Figure 5.1) are listed. The double-headed arrows from one factor to the next indicate that each level of each factor can be used in combination with any level of the other factors. One can obtain all factor–level combinations shown in Table 5.2 by following a particular sequence of arrows in Figure 11.1. For example, one arrow goes from steel to nonporous, another from nonporous to lubricant 4: this is factor–level combination 8 in Table 5.2. Note the crossing pattern of these arrows.

The arrows in Figure 11.2 for the density study do not cross one another. One cannot connect shift 2 with disk 1 cut from plaque 1 by following any of the arrows in Figure 11.2. This is because disk 1 of plaque 1 was produced during the first shift. It is a unique entity, different from disk 9 cut from plaque 3, even though each of these disks is the first one cut from the first plaque produced on a shift.

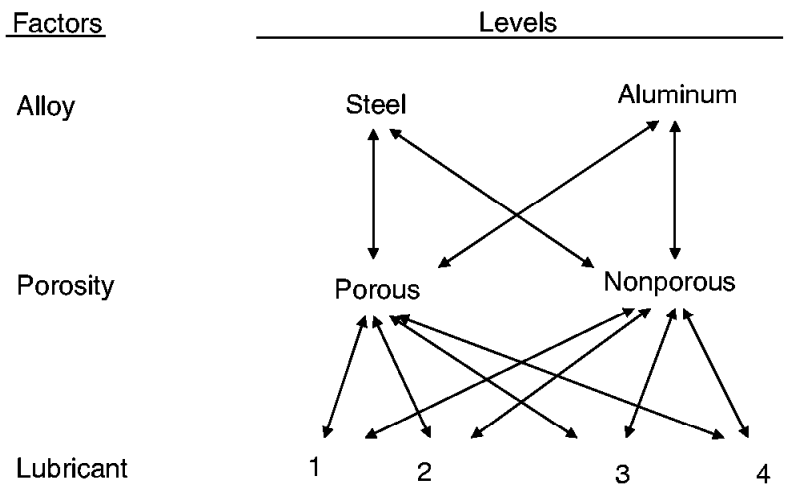


Figure 11.1 Crossed factors.

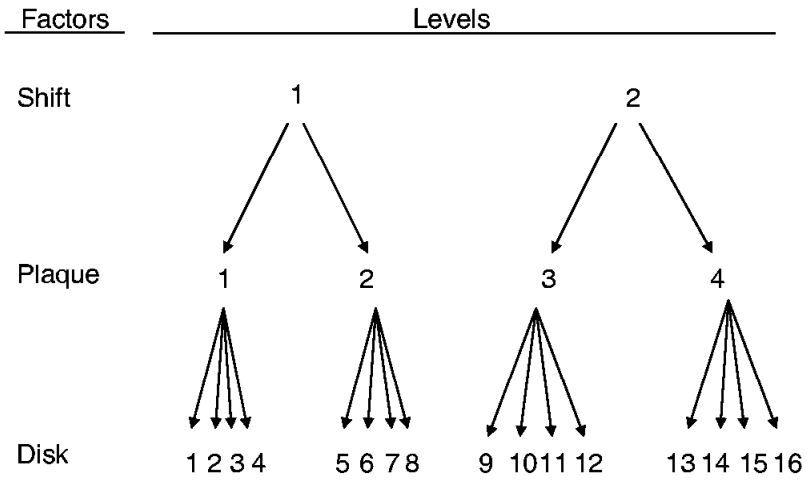


Figure 11.2 Nested factors.

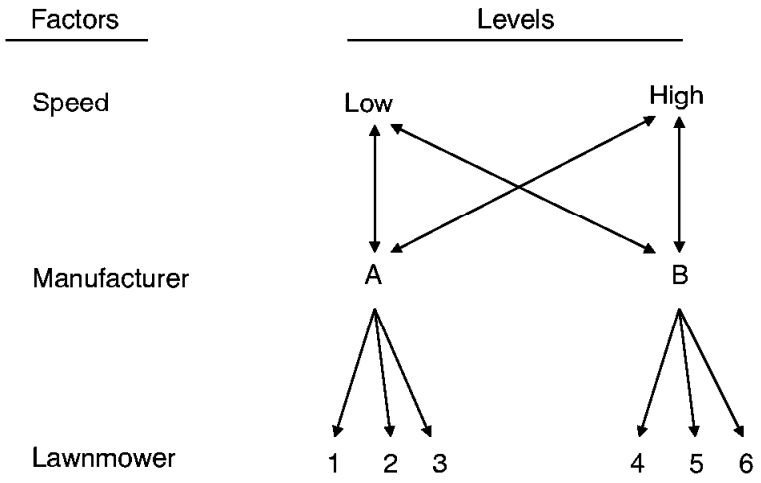


Figure 11.3 Crossed and nested factors.

Both nested and crossed factors can be included in an experimental design. The experiment conducted to investigate automatic cutoff times of lawnmowers (Table 9.1) is an example of a design with both types of factors: manufacturer (A, B) and speed (high, low) are crossed factors, and lawnmower is a nested factor. Figure 11.3 illustrates the crossing of the speed and manufacturer factors and the nesting of lawnmowers within manufacturers.

11.2 HIERARCHICALLY NESTED DESIGNS

Hierarchically nested designs are appropriate when each of the factors in an experiment is progressively nested within the preceding factor. The factors in the polyethylene density study in Section 11.1 are hierarchically nested:

disks within plaques, plaques within shifts. The layout for a hierarchically nested experiment (see Exhibit 11.3) is similar to that of Figure 11.2. The design is usually balanced. Balance is achieved by including an equal number of levels of one factor within each of the levels of the preceding factor. Wherever possible, randomization of testing the factor levels is included in the specification of the design.

EXHIBIT 11.3 CONSTRUCTION OF HIERARCHICALLY NESTED DESIGNS

1. List the factors to be included in the experiment.
2. Determine the hierarchy of the factors.
3. Select, randomly if possible, an equal number of levels for each factor within the levels of the preceding factor.
4. Randomize the run order or the assignment of factor–level combinations to experimental units.

The allocation of experimental resources in hierarchically nested designs yields more information on factors that are lower in the hierarchy than on those that are higher. For example, only two shifts are studied in the design of Figure 11.2, whereas sixteen disks are included in the design. In some circumstances it may not be desirable to have many factor levels at lower stages of the hierarchy. In such circumstances, unbalanced nested designs can be used to reduce the number of factor levels. The most popular of this type of design is the "staggered" nested design.

Figure 11.4 shows a typical layout for a staggered nested design for four factors. In this figure, factor *A* occurs at two levels. Two levels of factor *B*

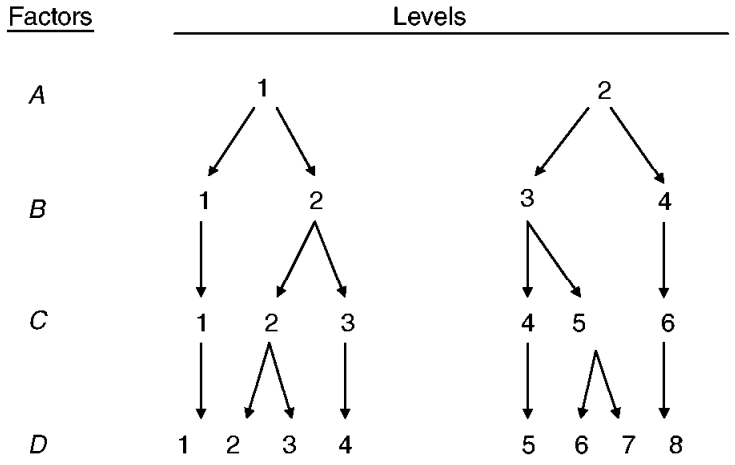


Figure 11.4 Staggered nested design.

are nested within each level of factor A. Rather than selecting two levels of factor C to be nested within each level of factor B, two levels of factor C occur within only one of the levels of factor B. The other level of factor B has only one level of factor C. This process continues at each remaining stage of the construction of staggered nested designs.

Note the pattern in the figure. Whenever a factor has only a single level nested within a level of the previous factor, all subsequent factors also have only a single level nested within them. The unbalanced pattern in Figure 11.4 can easily be extended to more factors or more levels within each factor.

An illustration of the use of this type of design is given in Figure 11.5. This experiment is designed to investigate sources of variability in a continuous polymerization process that produces polyethylene pellets. In this process, lots (100 thousand-pound boxes of pellets) are produced. From each lot, two boxes are selected. Two sample preparations are made from pellets from one of the boxes; one sample preparation is made from pellets from the other box. Two strength tests are made on one of the sample preparations from the first box. One strength test is performed on each of the other preparations. The completed design consists of strength tests from 30 lots.

The advantage of the design just described over the traditional hierarchical design is that many fewer boxes, preparations, and tests need to be included in the design. If one chose to test two boxes, two preparations from each box, and two strength tests on each preparation, the number of factor levels for each type of nested design would be as shown in Table 11.1. Half the number of strength tests are required with the staggered nested design.

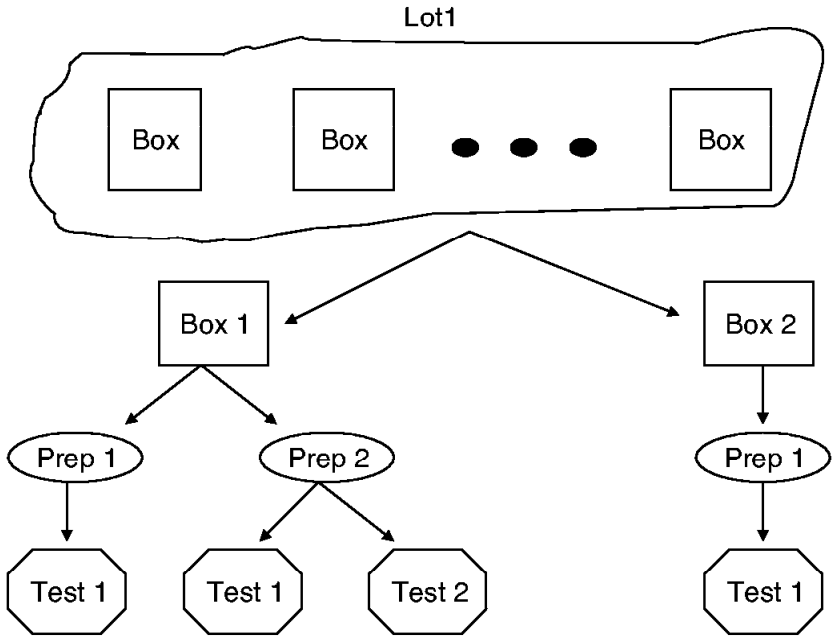


Figure 11.5 Staggered nested design for the polymerization-process study.

TABLE 11.1 Number of Levels for Hierarchical and Staggered Nested Designs: Polymerization Study

Factor	Number of Levels	
	Hierarchical Design	Staggered Nested Design
Lots	30	30
Boxes	60	60
Preparations	120	90
Strength tests	240	120

11.3 SPLIT-PLOT DESIGNS

Split-plot designs are among the most commonly used in modern industrial and agricultural experimentation. While common, they are often analyzed incorrectly because the nested error structure of the design is not recognized. At least two factors contribute to this problem of recognition. First, the classical nomenclature of *split plots* suggests that one can only have a split-plot design structure with nested experimental units, the "whole plots" and the "split plots" that are discussed more fully in Section 11.3.2. Second, the nomenclature itself — split-plot design — suggests that this type of design only is found in agricultural experiments. Both of these misconceptions are addressed in this section and in Section 11.4.

In Section 11.3.1, the basic character of split-plot designs is illustrated with an example that does not fit the classical split-plot design structure of agricultural experiments. In this illustration, there are "whole plots" but no "split plots." In Section 11.3.2, the classical split-plot design construction is detailed, with examples from agriculture and industry. In Section 11.4, split-plot design structure is shown to occur in experiments in which there are no experimental units but there are equivalent restrictions on how factor levels are randomized in the experiment.

11.3.1 An Illustrative Example

The data in Table 11.2 consist of oxides of nitrogen (NO_x) from automobile exhaust emissions of two vehicles from each of three models. The analysis of variance table shown in Table 11.2 shows that the average NO_x emissions for the three models are significantly different ($p = 0.001$.), in spite of the vehicle-to-vehicle variability within each model type. Note that the error variability estimate in this example is from vehicle-to-vehicle variability.

Table 11.3 shows data from an experiment on oxides of nitrogen in which three fuels were tested in each of six vehicles. Both vehicles and fuels have

TABLE 11.2 Analysis of NO_x (gal/mile) Emissions for Two Vehicles of Three Model Types

Model	Vehicle	NO_x
A	A1	0.0707
A	A2	0.0618
B	B1	0.0265
B	B2	0.0295
C	C1	0.1018
C	C2	0.1015

Source	df	S.S.	M.S.	F	p-Value
Models	2	0.00543	0.00272	184.3	0.001
Error	3	0.00004	0.00002		
Total	5	0.00547			

significant effects on the average NO_x emissions because the interaction is statistically significant ($p < 0.001$.). Note that in this example the appropriate error variability estimate is obtained from the repeat test variability.

Modern experiments in the automobile industry are not conducted as described in the previous two experiments. It is very expensive to conduct emissions tests like those illustrated in the previous two examples. Today's current practice is to combine these smaller experiments into one larger, more cost-effective one. This is straightforward from a statistical perspective if one keeps in mind that there are two key sources of error variation in the combined experiment, just as there were two sources of variation in the individual experiments.

The data in Tables 11.2 and 11.3 were collected from one experiment. The data and the corresponding analysis of variance table are shown in Table 11.4. The combined experiment was conducted as a split-plot design. Note that the portion of the experiment labeled "whole-plot analysis" is similar to the analysis shown in Table 11.2. The model effect on the average NO_x emissions is assessed using an F statistic that has the vehicle mean square as the denominator. This is one of the key differences between an analysis of crossed and nested factors. Often this difference is not recognized and the model effect incorrectly is tested against error.

Another key difference is caused by the vehicle effect being a random effect. Because of this, the vehicle effect, the fuel effect, and the fuel × model interaction effect are tested against the fuel × vehicle(model) interaction mean square, not the error mean square. Only the latter interaction is tested against the error mean square. The reader need not be concerned with a detailed understanding of why the analysis is conducted in the manner. The justification for the analysis will be presented in Chapter 13. What is important to realize

TABLE 11.3 Analysis of Fuel and Vehicle Effects on NO_x (gal/mile) Emissions

Vehicle	Fuel	Repeat #1	Repeat #2
1	F1	0.0707	0.0727
	F2	0.0936	0.0987
	F3	0.1337	0.1283
2	F1	0.0618	0.0666
	F2	0.0828	0.0846
	F3	0.1225	0.1183
3	F1	0.0265	0.0337
	F2	0.0413	0.0478
	F3	0.1147	0.1158
4	F1	0.0295	0.0255
	F2	0.0263	0.0316
	F3	0.1032	0.1183
5	F1	0.1018	0.0944
	F2	0.1486	0.1361
	F3	0.3002	0.2956
6	F1	0.1015	0.1013
	F2	0.1580	0.1510
	F3	0.2000	0.2299

Source	df	S.S.	M.S.	F	p-Value
Vehicles	5	0.07586	0.01517	331.9	0.000
Fuel	2	0.06390	0.03195	698.9	0.000
Fuel × Vehicles	10	0.01726	0.00173	37.8	0.000
Error	18	0.00082	0.00005		
Total	35	0.15785			

is that the correct F statistics for testing the various effects in Table 11.4 do not all have the error mean square in the denominators.

One must realize that it was the conduct of the experiment that determined the proper analysis of the data in Table 11.4. Because vehicles were nested within models and all fuels were randomly tested on each vehicle, a split-plot design structure ensued. In the classical split-plot design structure discussed in the next section, vehicles are the whole plots. There are no split plots in this experiment. The repeat tests can, however, be viewed as the equivalent of split plots.

11.3.2 Classical Split-Plot Design Construction

Many experiments require that all combinations of levels of one or more factors occur within levels of one or more other factors. While one could

TABLE 11.4 Split-Plot Analysis of NO_x Emissions Experiment

Model	Vehicle	Fuel	Repeat #1	Repeat #2
A	A1	F1	0.0707	0.0727
		F2	0.0936	0.0987
		F3	0.1337	0.1283
A	A2	F1	0.0618	0.0666
		F2	0.0828	0.0846
		F3	0.1225	0.1183
B	B1	F1	0.0265	0.0337
		F2	0.0413	0.0478
		F3	0.1147	0.1158
B	B2	F1	0.0295	0.0255
		F2	0.0263	0.0316
		F3	0.1032	0.1183
C	C1	F1	0.1018	0.0944
		F2	0.1486	0.1361
		F3	0.3002	0.2956
C	C2	F1	0.1015	0.1013
		F2	0.1580	0.1510
		F3	0.2000	0.2299

Source	df	S.S.	M.S.	F	p-Value
Whole-Plot Analysis					
Model	2	0.07386	0.03693	55.40	0.000
Vehicle(Model)	3	0.00200	0.00067	0.71	0.558
Split-Plot Analysis					
Fuel	2	0.06390	0.03195	34.05	0.000
Model × Fuel	4	0.01163	0.00291	3.10	0.042
Fuel × Vehicle (Model)	6	0.00563	0.00094	20.53	0.000
Error	18	0.00082	0.00005		
Total	35	0.15784			

think of this happening with complete factorials in completely randomized designs, an additional feature of the experiment prevents randomization across all the combinations of factor levels. Two of the most frequently encountered situations where this occurs are (1) when an experiment consists of two types of experimental units and (2) when some factor levels are easy or inexpensive to change while others are very difficult or very costly to change.

When two types of experimental units occur in an experiment, one type is often nested within another (e.g., small agricultural fields within each of

several large farms) and the levels of different factors are applied to each type of experimental unit. The larger experimental units are referred to as *whole plots* (see below) and the smaller ones as *split plots*. When levels of some factors are difficult to change, experiments are often conducted by varying the levels of the easy-to-change factors within fixed levels of the difficult-to-change factors. This type of *restricted randomization* leads to the same type of analysis as for experiments with two types of experimental units. Restricted randomization is discussed further in Section 11.4. In each of these situations, a feature of the experiment prevents the random assignment of all factor–level combinations to the experimental units or to the test sequence, resulting in a nesting of levels. A statistical design that is often used in these types of applications is the split-plot design (Exhibits 11.4 and 11.5).

EXHIBIT 11.4

Split-Plot Design. A split-plot design occurs when all combinations of levels of one or more factors occur within each combination of levels of one or more other factors. The nesting can be due to the assignment of factors to two or more types of experimental units, one of which is nested with the other(s), or to the more frequent randomization of some (combinations of) factor levels than others.

EXHIBIT 11.5 CONSTRUCTION OF SPLIT-PLOT DESIGNS

1. Identify whole plots and the factors to be assigned to them. List all combinations of the whole-plot factors that are to be tested.
2. Identify split plots and the factors to be assigned to them. List all combinations of the split-plot factors that are to be tested.
3. Randomly select a whole plot. Randomly assign a whole-plot factor–level combination to the whole plot.
4. Randomly select split plots from the whole plot chosen in step 3. Randomly assign combinations of the split-plot factors to the split plots.
5. Repeat steps 3 and 4 for all the whole plots.
6. Steps 1–5 can be extended to include additional nesting of experimental units (e.g., split-split-plots) as needed.

The term *split-plot* derives from agricultural experimentation, where it is common to experiment on plots of ground. Levels of one or more factors are assigned to large geographical areas referred to as *whole plots* (e.g., individual farms), while levels of other factors are applied to smaller geographical areas referred to as *split plots* (e.g., fields within each farm). For example, one might need to fertilize crops or control insects by spraying entire farms from airplanes. Thus, different fertilizers or insecticides would be applied to entire

farms. On the other hand, different varieties of crops or different crop spacings could readily be applied to individual fields within a farm. Such an experimental layout, with appropriate randomization, would constitute a split-plot design.

Figure 11.6 illustrates a design for an experiment similar to the one just described. Two insecticides, A and B, are each applied to two farms. Two varieties of seed, seed 1 and seed 2, are planted in each of three fields on each of the farms. In this example, insecticide is the whole-plot factor and the farms are the whole plots (whole-plot experimental units). Seed is the split-plot factor and the fields are the split plots (split-plot experimental units).

The layout of the design shown in Figure 11.6 might, at first glance, appear to be that of a randomized complete block design with farms as blocks. However, a randomized complete block design requires that factor–level combinations be randomly assigned to the experimental units in each block. In the design shown in Figure 11.6, the insecticide-variety combinations cannot be randomly assigned to the fields because each insecticide must be assigned to all the fields in a farm. It is far more cost-effective to spray an entire farm with one insecticide than to apply different insecticides to each field. In contrast, the varieties of seed are easily applied to individual fields.

In the analysis of data from split-plot designs (see Chapter 13), the whole-plot experimental units act like levels of a blocking factor which is included, along with the whole- and split-plot factors themselves, in the analysis of variance model. The replicate farms in this example are blocks. Because of this, the advantages associated with blocking (see Section 9.1) are obtainable with this and with many other split-plot designs. For example, the effect of the seeds is calculated from differences of averages over the fields within each farm (block). These differences eliminate the farm effect, and therefore the

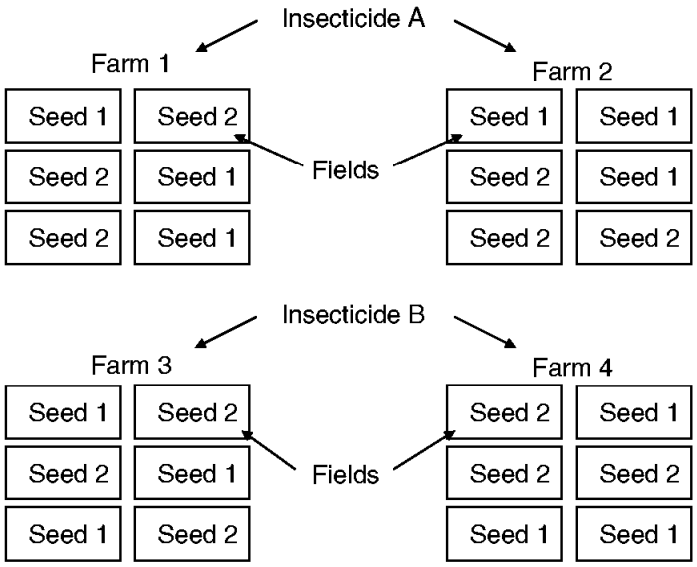


Figure 11.6 A split-plot design layout.

seed effect is measured with greater precision than the effect of the insecticides. The latter effect is calculated from the difference between the averages over all the observations in the replicate farms. These averages are subject to variability due to both fields and farms.

This example illustrates one of the main applications of split-plot designs: the estimation of components of variability due to two or more sources. Frequently responses are subject to variability from a variety of sources, some of which can be controlled through the use of appropriate blocking or nested designs. Hierarchical designs and split-plot designs, as well as the other blocking designs discussed in Chapter 9, can be used to allow estimation of the components of variability associated with the sources.

Although split-plot designs are extremely useful in agricultural experimentation, they are also beneficial in other scientific and in engineering research. As an illustration, consider an investigation of one of the components of an electronic connector. The pins on the connector are made from small-diameter bronze alloy wire. The factors of interest in this investigation are the vendors who supply the wire, the operational environment in which the pins are used, and the heat treatment applied to the pins. The factors and their levels are shown in Table 11.5.

This experiment is to be conducted using a split-plot design because there are two types of experimental units to which the levels of the factors are applied. Each vendor (whole-plot factor) supplies the wire on reels. The whole plots are reels of wire. Six-inch samples of wire are cut from the reels. The environment and heat-treatment factors (split-plot factors) are applied to the wire samples. The split plots are the wire samples taken from each reel.

A second example of the use of split-plot designs in industrial experimentation is the lawnmower automatic-cutoff-time example presented in Table 9.1. The whole-plot factor in this experiment is the manufacturer. The whole plot experimental units are the lawnmowers. The split-plot factor is the operating speed. In this example there are replicates and repeat tests. There are no split-plot experimental units; the individual test runs take the place of the split plots.

TABLE 11.5 Experimental Factors for Connector-Pin Study

Factor	Levels
Vendor	Vendor 1, vendor 2
Environment	Ambient, high humidity
Heat treatment	Yes, no

When designing split-plot experiments one should remember that experimental error is contributed by both the whole-plot and the split-plot experimental units. To estimate both types of experimental error it is necessary to replicate the whole plots and conduct repeat tests on the split plots. Thus, in the split-plot design illustrated in Figure 11.6, the farm whole plots were replicated to provide an estimate of the replicate error variability against which to compare the effect of the insecticides. Similarly, the three fields having the same seed application in each farm provide repeat measurements that can be used to estimate the repeat test error variability against which the seed effect can be compared. Replication and/or repeat tests may not be necessary if one can assume that certain interactions involving the whole-plot and split-plot factors are zero or at least negligible. This consideration will be discussed in the next section and in Chapter 13.

11.4 RESTRICTED RANDOMIZATION

Completely randomized designs require both randomization and operational conditions that sometimes cannot be met. Stressed in Chapter 5 were the randomization requirements on assignment of factor–level combinations to experimental units and to test run sequences. Equally important is the requirement that operational conditions be reset with each test run. For example, two test runs in which some factor levels are unchanged might not represent complete randomization if the variability of the response is affected by resetting levels. Failure to reset levels often causes variance estimates to be attenuated because only measurement error variability is being reflected in the test runs, not measurement error and operational variability.

There are occasions when the randomization of factor–level combinations to experimental units or to the run-order sequence must be restricted. The restrictions may be necessary because some of the factor levels are difficult or expensive to change, such as when components of a device must be disassembled and reassembled between test runs. This restriction of the randomization of factor–level combinations is similar to the restrictions placed on the assignment of factor–level combinations to whole plots and to split plots in split-plot designs.

In split-plot designs the split-plot factors can be randomized over any of the experimental units (plots) in the design. The whole-plot factors must be assigned only to the whole plots and in that respect are similar to factors that are difficult or costly to change.

As an illustration of this type of experiment, consider an investigation of wear on the cylinder walls of automobile engines. The factors of interest in such an investigation might include those shown in Figure 11.7. One might select a half fraction of a complete factorial experiment using the defining

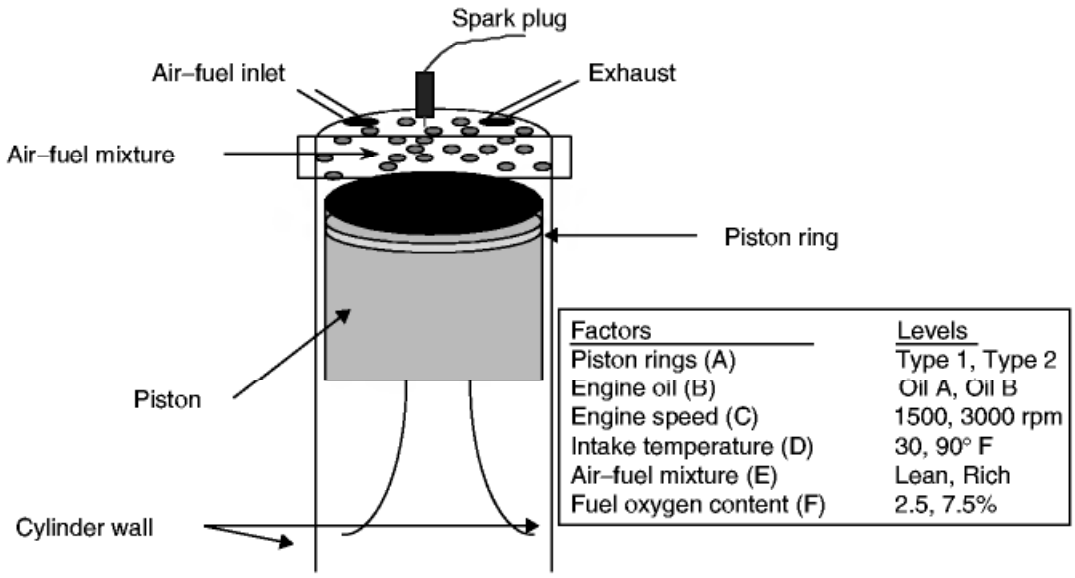


Figure 11.7 Factors for a cylinder wear study.

equation $I = ABCDEF$ for this investigation. A randomized test sequence for the 32 factor–level combinations for this factorial experiment is shown in Table 11.6.

Of concern to the research engineer who is conducting this investigation is the number of times the piston rings must be changed. There are 32 changes of the piston rings in the design shown in Table 11.6 if, as required for a completely randomized design, consecutive test runs with the same assigned ring type each have new rings of that type inserted on the piston. Each change of the piston ring requires that the engine be partially disassembled and reassembled, and that various pieces of test equipment be disconnected and reconnected. The design was considered very burdensome because of these requirements and because it extended the duration of the test program.

One solution to this problem is to change the piston rings only once, as indicated in Table 11.7. Within each set of 16 test runs for a given piston ring, the sequence of factor levels is completely randomized. Note the similarity with split-plot designs: piston-ring levels act as whole plots, and the run order within a piston ring constitutes the split-plot portion of the design.

The design in Table 11.7 consists of two groups of 16 test runs, one group corresponding to each piston ring. If a group effect results from some change in experimental conditions, it will be confounded with the piston-ring effect. This group effect is similar to the whole-plot error effect in a split-plot design. If each half of the design in Table 11.7 can be replicated, estimates of the variation due to the whole-plot (group) testing can be obtained. Likewise, if some of the test runs can be repeated, the split-plot error variation can be estimated. This is similar to the inclusion of replicate farms and several fields on which seed treatments were repeated in Figure 11.6.

TABLE 11.6 Cylinder-Wear Fractional Factorial Experiment: Completely Randomized Design

Run	Piston-Ring Type	Engine Oil	Engine Speed (rpm)	Intake Temperature (°F)	Air–Fuel Mixture	Oxygen Content (%)
1	2	A	3000	90	Lean	7.5
2	1	B	1500	90	Lean	2.5
3	2	A	1500	90	Rich	7.5
4	1	A	3000	90	Lean	2.5
5	1	A	1500	90	Lean	7.5
6	1	B	3000	90	Lean	7.5
7	2	A	3000	30	Lean	2.5
8	2	A	1500	30	Rich	2.5
9	1	B	3000	30	Rich	7.5
10	2	A	3000	90	Rich	2.5
11	1	A	3000	30	Rich	2.5
12	2	B	3000	90	Rich	7.5
13	1	B	1500	30	Lean	7.5
14	1	B	1500	30	Rich	2.5
15	2	A	1500	30	Lean	7.5
16	1	B	3000	90	Rich	2.5
17	1	A	3000	30	Lean	7.5
18	2	B	1500	90	Lean	7.5
19	2	A	1500	90	Lean	2.5
20	2	B	1500	30	Rich	7.5
21	2	B	3000	30	Rich	2.5
22	1	A	1500	90	Rich	2.5
23	1	A	3000	90	Rich	7.5
24	2	A	3000	30	Rich	7.5
25	2	B	1500	90	Rich	2.5
26	1	A	1500	30	Lean	2.5
27	2	B	3000	90	Lean	2.5
28	2	B	1500	30	Lean	2.5
29	1	A	1500	30	Rich	7.5
30	1	B	3000	30	Lean	2.5
31	2	B	3000	30	Lean	7.5
32	1	B	1500	90	Rich	7.5

The use of restricted randomization without replication and/or repeat tests can only be advocated when experiments are so well controlled that the whole-plot error can be ignored; that is, it is nonexistent or negligible. As will be discussed in Chapter 13, if certain interactions involving split-plot

TABLE 11.7 Cylinder-Wear Fractional Factorial Experiment: Split-Plot Design

Run	Piston-Ring Type	Engine Oil	Engine Speed (rpm)	Intake Temperature (°F)	Air–Fuel Mixture	Oxygen Content (%)
1	1	A	3000	30	Rich	2.5
2	1	B	1500	90	Lean	2.5
3	1	A	1500	90	Rich	2.5
4	1	A	1500	30	Rich	7.5
5	1	B	1500	30	Lean	7.5
6	1	B	1500	90	Rich	7.5
7	1	B	1500	30	Rich	2.5
8	1	A	3000	90	Lean	2.5
9	1	B	3000	30	Rich	7.5
10	1	A	1500	90	Lean	7.5
11	1	B	3000	90	Rich	2.5
12	1	A	3000	30	Lean	7.5
13	1	A	3000	90	Rich	7.5
14	1	B	3000	30	Lean	2.5
15	1	A	1500	30	Lean	2.5
16	1	B	3000	90	Lean	7.5
17	2	B	1500	90	Rich	2.5
18	2	A	3000	90	Rich	2.5
19	2	B	1500	90	Lean	7.5
20	2	B	3000	30	Rich	2.5
21	2	A	1500	90	Lean	2.5
22	2	A	3000	30	Lean	2.5
23	2	A	3000	90	Lean	7.5
24	2	B	1500	30	Rich	7.5
25	2	A	1500	30	Lean	7.5
26	2	B	3000	90	Rich	7.5
27	2	B	3000	30	Lean	7.5
28	2	A	1500	30	Rich	2.5
29	2	B	1500	30	Lean	2.5
30	2	A	3000	30	Rich	7.5
31	2	A	1500	90	Rich	7.5
32	2	B	3000	90	Lean	2.5

factors are zero, the split-plot error variation can be estimated from the quantities that would ordinarily be used to measure these interaction effects. So too, if whole-plot error is not negligible but interactions involving only the whole-plot factors are negligible, the quantities used to measure the interaction effects can be used to estimate the whole-plot error variation. If interactions

involving both whole-plot factors and split-plot factors can be assumed to be negligible, replication and/or repeat tests need not be conducted. This is an important consideration in many types of scientific and engineering studies for which fractional factorial experiments are conducted.

REFERENCES

Text References

Nested designs, in particular split-plot designs, are not covered in many elementary engineering statistics texts. The following books include adequate coverage of nested designs, in addition to those included in Chapters 4 and 6.

Hinkleman, K. and Kempthorne, O. (1994). *Design and Analysis of Experiments*, New York: John Wiley & Sons, Inc.

Littell, R. C., Millikin, G. A., Stroup, W. W., and Wolfinger, R. D. (1996). *SAS System for Mixed Models*, Cary, NC: SAS Institute, Inc.

Steel, R. G. D. and Torrie, J. H. (1980). *Principles and Procedures of Statistics: A Biometrical Approach*, New York: McGraw-Hill Book Company.

Staggered nested designs are discussed in the following articles:

Bainbridge, T. R. (1965). "Staggered, Nested Designs for Estimating Variance Components," *Industrial Quality Control*, **22**, 12–20.

Smith, J. R. and Beverly, J. M. (1981). "The Use and Analysis of Staggered Nested Factorial Designs," *Journal of Quality Technology*, **13**, 166–173.
The latter article not only discusses the type of staggered nested designs introduced in Section 11.2, *it also includes designs in which some factors have a factorial relationship and others are nested.*

The use of split-plot designs with factorial or fractional factorial experiments without including replicates and/or repeat tests is discussed in the following article:

Addleman, S. (1964). "Some Two-Level Factorial Plans with Split-Plot Confounding," *Technometrics*, **6**, 253–258.
In addition to discussing the use of interaction effects to estimate whole-plot and split-plot error variation, the article provides a list of defining contrasts for constructing fractional factorial experiments in split-plot designs with high-order confounding of interaction effects.

Restricted randomization and its use with split-plot designs is discussed in the following article:

Lu, J. M. and Ju, H. (1992). "Split-Plotting and Randomization in Industrial Experimentation," *American Society for Quality Annual Quality Congress Transactions*, 374–382.

EXERCISES

1 The plume dispersion of liquid chemicals from a ruptured hole in a chemical tanker is to be studied in a simulated river channel. Two factors of interest are

the speed of the river (10, 100, 300 ft/min) and the spill rate of the chemical (50, 150, 400 gal/min). The experimenter is interested in comparing the plume dispersion of sodium silicate and ethyl alcohol. Design an experiment for this study. Which factors, if any, are crossed? Which, if any, are nested? Sketch a layout of the experiment similar to the figures in this chapter.

2 A ballistics expert wishes to investigate the penetration depth of projectiles fired from varying distances into an armor plate. Three different gun types that fire the same-size projectiles are to be studied. Two guns of each type are to fire four projectiles at ranges of 5, 50, and 100 feet. Design an experiment for this study. Which factors, if any, are crossed? Which, if any, are nested? Sketch a layout of the experimental design.

3 Refer to the heat-transfer study of Exercise 5, Chapter 5. For a specific glass type and cooling temperature, suppose an experiment is to be conducted to assess the variability of the mold temperatures that is inherent in the production process. In this experiment, three disks used to make the drinking glasses at each plant are to be studied. Each disk contains molds to produce three drinking glasses. Five measurements of mold temperature are to be taken on each mold from each disk at each of the two plants. Design an experiment for this study. Which factors, if any, are crossed? Which, if any, are nested? Sketch a layout of the experimental design.

4 Suppose the number of measurements to be taken in the experiment described in Exercise 3 is considered excessive. Design two alternative hierarchically nested designs, one a staggered nested design, which would reduce the number of measurements yet retain the ability to estimate measurement error from repeat measurements on the mold temperature.

5 Sketch a layout of the experimental design for the wing design study in Exercise 8 of Chapter 9. What type of design is it? Identify whether the factors are crossed or nested and whether there is replication or repeat tests.

6 An interlaboratory study of the stress (ksi) of titanium is to be designed, involving three laboratories in the United States and three in Germany. The laboratories in each of the two countries can be considered representative of the many laboratories making stress measurements in their countries. Two temperatures (100, 200°F), and four strain rates (1, 10, 100, 1000 sec^{-1}) are to be investigated. Two titanium specimens are available for each lab–temperature–strain combination. Design an experiment for this study. Identify crossed and nested factors, if any, and sketch a layout of the design.

7 The quality of dye in a synthetic fiber being manufactured by a textile factory is to be investigated. The response of interest is a quantitative measurement of the dye in a specific fiber. Four factors are to be studied: shifts (three shifts of eight hours each on a given work day), operators (two on each shift), packages of bobbins, and individual bobbins. Each

operator produces three packages of bobbins on each shift; each package contains two bobbins. Design an experiment to study these four factors. The manufacturer restricts the experiment to one work day and no more than 20 bobbins. Sketch a layout of the design.

8 A chemical laboratory is experimenting with factors that may contribute to the success of plating gold on a base metal. The factors being studied in the current investigation are plating-bath type (two levels), anode type (two levels), and current density (three levels). It is noted that current density is easy to change but that the other two factors are much more difficult. Design an experiment for this study. Include replicates and/or repeat tests where desirable.

9 A study of alloying processes leads to an experiment to compare the effects of heat treatment on the strength of three alloys. Two ingots are to be made of each of the alloys. These ingots are to be cut into four pieces, each of which will receive one of three heat treatments. Design an experiment for this investigation. Sketch a layout of the design.

10 The technical training office of a large international manufacturer is experimenting with three methods of teaching the concepts of statistical process control to its work force. Five large manufacturing plants are chosen to be part of the study. In each plant, employees on a particular shift are to be given one of the three methods of instruction. The shift to receive each type of instruction is to be randomly selected at each plant. A random sample of the employees in each shift are to be further divided into two groups. One of the groups on each shift is to be trained in a workshop to implement the process-control techniques on microcomputers, with appropriate software for calculations and graphics. The other group is to utilize calculators and graph paper. Design an experiment for this study involving no more than 10 workers from each shift at each plant.

11 A company seeks to evaluate three manufacturers of diesel engines. Each manufacturer supplies three engines. Each engine is equipped with one of three different after-treatment devices, used to reduce engine emissions. Each device is tested over two different engine operating cycles, one at high speed and high load and the other at low speed and low load. Three measurements are taken on the resulting NO_x emissions from each engine. Design an experiment for this study. Identify crossed and nested factors, if any, and sketch a layout of the design.

12 In Exercise 11, suppose that one of the manufacturers is only able to supply two engines, and another manufacturer is only able to supply one engine. Design an experiment for this study. Identify crossed and nested factors, if any, and sketch a layout of the design.

13 Four different paint formulations are to be applied to an aircraft at three different temperatures and then compared in terms of their durability. The

paints are tested by preparing a formulation and applying it to a small sample of metal used on the exterior of the aircraft. The tests are conducted within a single enclosed structure where the temperature can be controlled. It is desired to run four replications of each formulation. Suppose the two designs available for running this experiment are a completely randomized design and a split-plot design. Give a layout for each design, and explain the pros and cons of running each design.

14 Three different types of fuel injectors are to be tested on four different makes of automobiles to determine if there is a reduction in engine deposits, with the constraint that no injector is used for more than one test. Each vehicle is driven 3,000 miles before a new injector is installed. Also, there are two possible driving routes to take for each test. Design an experiment for this study. Identify crossed and nested factors, if any, and sketch a layout of the design.

15 Three chemistry laboratories in different parts of the United States are asked to run analyses on four samples of gasoline, each of which contain a different amount of oxygenate. In each lab three different technicians are randomly chosen to conduct the experiment and each technician examines all four gasoline samples in a random sequence. The technician records the oxygen content of each fuel and takes three repeat measurements on each sample. Design an experiment for this study. Identify crossed and nested factors, if any, and sketch a layout of the design.

16 A study was conducted to measure the combustion deposits on the surface of a certain brand of pistons used in a four-cylinder automobile. Two vehicles were independently selected and the four cylinders were separately examined on each vehicle. For each cylinder, the associated piston surface was measured by identifying 30 locations uniformly distributed across the top of the piston surface. Each location was sampled two times, and three replicates were sampled for each set of 30 locations. Design an experiment for this study. Identify crossed and nested factors, if any, and sketch a layout of the design.

17 An innerspring mattress plant takes wire from carriers and forms it into innerspring mattress units. A carrier is 3000 lbs. of wire. The manufacturer is evaluating four different steel-rod suppliers relative to coiling and assembly efficiency. Three heats are to be chosen from each supplier, where a heat is a batch of steel that is formed into rod. From each heat, it is desired to draw wire onto five carriers. The response is the efficiency of the spring plant. Design an experiment for this study. Identify crossed and nested factors, if any, and sketch a layout of the design.

18 A semiconductor manufacturer desires to evaluate the thickness of the silicon wafers being produced at a production site. Three different furnace runs are to be selected. From each furnace run, five wafers are to be selected, and on each wafer, eight sites are to be measured. Design an

experiment for this study. Identify crossed and nested factors, if any, and sketch a layout of the design.

19 Suppose an experiment was to be run involving five machines. Three technicians of varying levels of training operate each machine and each technician takes two measurements. Describe under what circumstances that this experiment might be viewed as a nested design, and when it might be viewed as a crossed design. Sketch the layout for each of your designs.

20 List at least two advantages and two disadvantages of using restricted randomization in an experimental design. Explain the difference between using complete randomization and restricted randomization in a split-plot experimental design.

CHAPTER 12

Special Designs for Process Improvement

In this chapter we present designs that have been found to be highly successful in process improvement applications. These designs are useful for obtaining information about processes so that critical product and process characteristics can be identified, monitored, and kept on target. In addition, if needed for process improvement or competitive advantage, the inherent variation in these key characteristics can be reduced. The following topics are introduced in this chapter:

- *the use of gage repeatability and reproducibility designs in studies used to assess the adequacy of a measurement system,*
- *process capability designs for evaluating the performance of key product characteristics relative to customer specifications or expectations,*
- *an approach popularized by Genichi Taguchi for designing robust products, and*
- *an integrated approach that allows great flexibility for studying process and product robustness.*

In previous chapters, randomization and blocking techniques were used to minimize the effects of unwanted and uncontrollable factors on the responses of primary interest in an experiment. In the context of block designs in Chapter 9, it was stressed that uncontrollable factors often introduce excess variation, compromising one's ability to identify and to control the primary factor effects that are being studied in the experiment. In this chapter, emphasis will be on designs that can be used to identify these uncontrollable factors through special experiments. One key purpose of conducting such specialized

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

experiments is to determine combinations of controllable factor levels that minimize the variation of the uncontrollable factor levels. This is the concept behind *robust product* or *process design.* A product design is *robust* if uncontrollable factors that vary during production, evaluation, or consumption have minimal effect on product quality and functionality.

Note that *design* is used in two contexts in this chapter. Statistical experimental design is used to aid in the identification of factors that contribute to the variation of key process or product characteristics. This information is then used in the manufacturing design of products and processes to make these products and processes more robust to uncontrollable environmental or manufacturing influences on their performance in meeting customer needs.

The two sections in this chapter deal with statistical methods for assessing quality performance and statistical designs for product or process improvement. Much of this chapter builds on applications of complete and fractional factorials discussed in Chapters 5 and 7.

12.1 ASSESSING QUALITY PERFORMANCE

Product and process improvement goals are to improve customer satisfaction with as efficient use of resources as possible. The term *customers* is used very broadly in this context, meaning users or beneficiaries of the products and processes. Efficient use of resources relates both to the improvement study itself and to the changes that must be made to implement product or process improvement. This section addresses two important process improvement issues:

- Are the measurements adequate to accomplish process improvement goals?
- Does the current process produce results that conform to customer needs?

12.1.1 Gage Repeatability and Reproducibility

Gage repeatability and reproducibility studies, commonly referred to as Gage R&R studies, are used to help ensure that data of sufficiently high reliability are obtained from a measurement process. The measurement process is the collection of operations, procedures, gages, equipment, software, and personnel used to assign a numerical value to the characteristic being measured. These numerical values or measurements are used for a variety of purposes, including determining if the process is adequately controlled, implementing procedures for actually controlling the process, determining if the process is capable of meeting customer requirements, increasing process understanding, modeling the process, quantifying improvement, characterizing manufactured product, conducting sampling inspection, etc.

Numerical values obtained from the measurement process need to be of sufficient reliability to be useful for the purpose intended. The reliability of the measurement is determined by whether the measurement process has sufficiently small bias and sufficiently high precision (small measurement variability). Bias is the difference between the location of the center of the measurement data and a standard or reference value. Sufficiently small measurement bias is generally ensured through calibration procedures. A more common reason for unreliable measurement data than bias is low precision — excessive variability — in the measurement process. Excessive variability in the measurement process may mask the true variation in the manufacturing process (or whatever process is being studied), impairing a decision-maker's ability to make correct decisions about ways to improve the product or process.

Figure 12.1 illustrates these points. Figure 12.1(a) shows the true behavior of a characteristic of interest relevant to a process. Figure 12.1(b) shows 60 measured values of the characteristic of interest taken with modest measurement variability. In Figure 12.1(b), about 30% of the variation is due to the measurement process and about 70% to the true values of the process. The modest measurement variability of the values plotted in Figure 12.1(b) is adequate for drawing inferences about the behavior of the true process. This is not the case for the high measurement variability of the values plotted in Figure 12.1(c). In Figure 12.1(c), the measurement process has high variability. About 70% of the variation is due to the measurement process and about 30% to the true values of the process.

Gage R&R studies are useful for understanding the sources of variation that can influence numerical values produced by a measurement process and for assessing the adequacy of the measurement process in describing the behavior of the true process. Before describing the design requirements of a Gage R&R study, some terms frequently used to describe key features of Gage R&R studies are defined in Exhibit 12.1. Although these definitions and the discussion that follows are specifically germane to parts manufacturing processes, they are easily generalized to other processes, including service-oriented processes.

EXHIBIT 12.1 GAGE R&R TERMINOLOGY

Repeatability. The variation in measurement obtained with one measurement instrument when used several times by the same appraiser while measuring one characteristic on the same part of a product.

Reproducibility. The variation in the average of the measurements made by different appraisers using the same measurement instrument when measuring one characteristic on the same part of a product.

Discrimination. The ability of the measurement process to detect and consistently indicate changes in the measured characteristic.

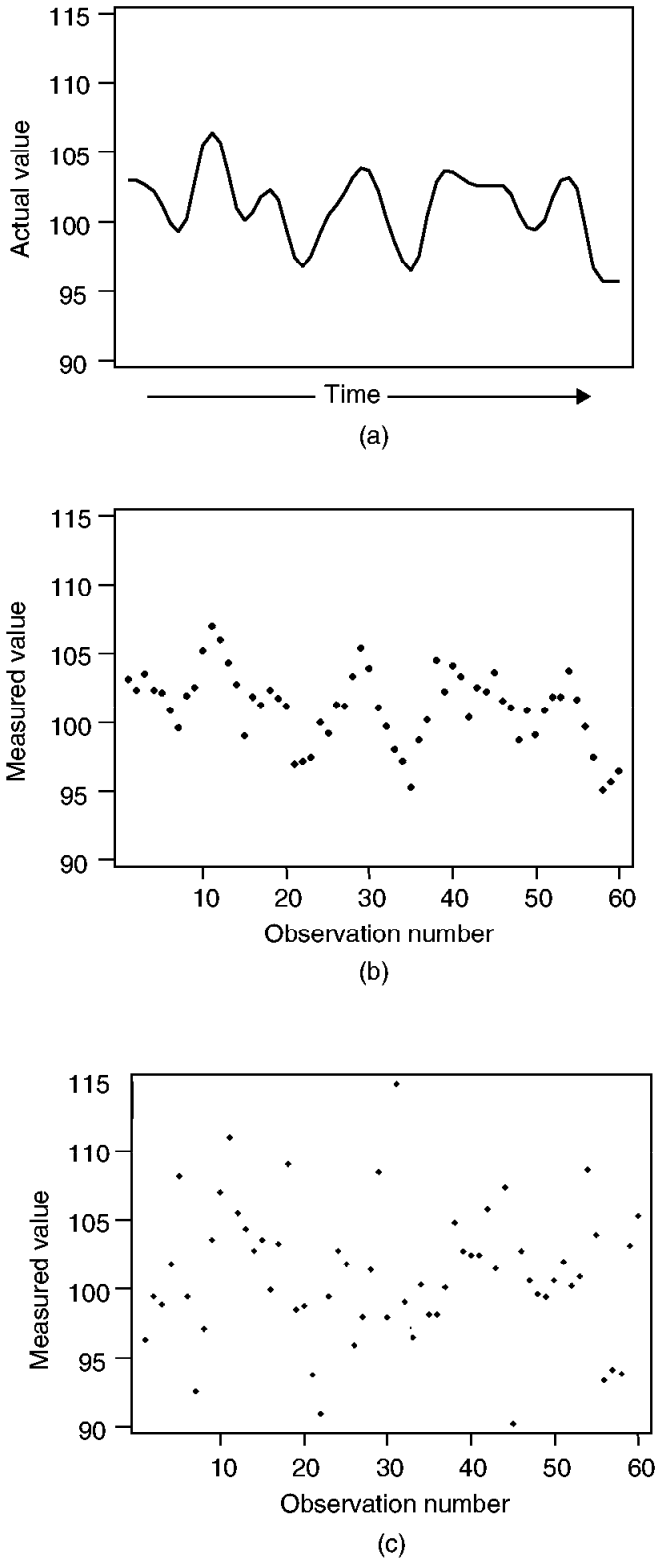


Figure 12.1 Process variability. (a) Actual process values. (b) Measured values with small variation. (c) Measured values with high variation.

Factorial experiments in completely randomized designs are typically used for Gage R&R studies. There are generally only two factors of interest: appraisers who take the measurements and parts. Appraisers should be chosen from among those who typically make the measurements of the characteristic of interest. Three appraisers are common in a Gage R&R study. More may be included, and doing so would provide better estimates of appraiser variability. The sampled parts are selected from the process and should represent the entire operating range of the part. Five parts are considered a minimum, and often ten or more are used. For each part and appraiser combination, three repeat readings are typically taken.

Figure 12.2 shows a data collection form and the actual data for a Gage R&R study on the length measurement (in inches) of a mattress innerspring unit. The measurement process consists of a production operator (appraiser) using a tape measure to measure the innerspring length. The tape measure is calibrated prior to the study. The five sample units, A through E, are chosen at random from several days of production and represent the manufacturing range of the product. Each innerspring unit is presented to each of the three operators (the operators are blind as to which unit they are measuring) in random order on three consecutive days. This design allows for an assessment of reproducibility, repeatability, and discrimination for this measurement process. The analysis of these data will be detailed in Chapter 13.

12.1.2 Process Capability

A capability study is a data collection strategy — a sampling plan — that provides information allowing one to estimate the inherent, common cause (inherently random) variation of a process and to compare that variation to customer requirements. Information from a capability study allows the prioritization of efforts to make products and processes robust.

Assessing process capability is appropriate only when two fundamental issues have been addressed: process standardization and stability. A process is standardized when it operates according to agreed upon procedures and conditions. Simply put, "you say what you do (documentation) and do what you say (compliance)." A stable process has no special (nonrandom) causes of variation present; they have been eliminated through removal or compensation. Process control charts, discussed in the references at the end of this chapter, are valuable tools for assessing process stability. After a process is determined to be standardized and stable, one can determine whether the process is capable of meeting customer requirements through a capability study.

There are two common types of capability studies, machine studies and process studies. A machine capability study is concerned with variation caused

Operator 1

	Sample A	Sample B	Sample C	Sample D	Sample E
Day 1	72.750	73.000	73.125	73.500	73.875
Day 2	72.750	73.031	73.250	73.438	73.750
Day 3	72.813	72.938	73.000	73.531	73.750

Operator 2

	Sample A	Sample B	Sample C	Sample D	Sample E
Day 1	73.000	73.000	73.156	73.438	73.813
Day 2	72.844	73.000	73.250	73.438	73.750
Day 3	72.750	73.125	73.125	73.500	73.500

Operator 3

	Sample A	Sample B	Sample C	Sample D	Sample E
Day 1	72.813	73.063	73.125	73.500	73.750
Day 2	72.813	73.125	73.063	73.500	73.813
Day 3	72.750	72.938	73.125	73.500	73.813

Figure 12.2 Gage R&R study for innerspring length (inches).

by one machine or operation. It is a short-term determination of capability, and data are usually collected on the same day. This type of capability study does not take into account all sources of process variation, and the results of the study can be viewed as a measure of process potential.

A process capability study measures variation caused by many sources: machines, materials, methods, operators, etc. It is a long-term determination of capability, and data are typically collected within a month-long time frame. The results from this type of study begin to measure process performance. Which type of capability study is appropriate depends on the intended needs of the study.

This brief description of process capability studies is intended to stress their importance in analytical studies for process improvement. Depending on the specific features of the study, any of the crossed and nested designs that have been introduced in previous chapters can be used for capability studies. The references at the end of this chapter can be consulted for details on all aspects of process capability studies. The remainder of this chapter discusses statistical experimental designs that are often used in conjunction with process improvement studies.

12.2 STATISTICAL DESIGNS FOR PROCESS IMPROVEMENT

All of the statistical designs discussed in previous chapters can be used in process improvement studies. Often these studies must be conducted with minimal resources, especially with short time frames and as few test runs as possible. One can appreciate the need for such restrictions because process improvement studies often involve current processes that must be temporarily stopped or altered to obtain relevant data. For this reason, screening experiments and highly fractionated factorials conducted in completely randomized designs are common. These designs have been thoroughly discussed in previous chapters so they will not be emphasized in this chapter.

One important strategy for improving a process is to design or redesign it so that product properties essential to the customer or consumer are robust to variation in uncontrollable factors that affect the process. The next section introduces and describes an approach popularized by a Japanese engineer named Genichi Taguchi. Section 12.2.1 describes Taguchi's approach to robust product design using design of experiment techniques. Taguchi's approach is then contrasted in Section 12.2.2 to an integrated, model-based approach that provides additional tools for studying and characterizing process and product robustness.

12.2.1 Taguchi's Robust Product Design Approach

The statistical designs promulgated by Genichi Taguchi are fractional factorials (Chapter 7) and have been widely available for many years. Taguchi has been able to make these designs more useful to engineers and scientists through (1) framing the use of these designs in an engineering context aimed at product design, (2) emphasizing a few basic statistical designs and providing them in convenient tables, and (3) providing a simple approach to the analysis of the data for practitioners. Taguchi's impact on applications of sounder experimental techniques for product and process design is easily seen in the volumes of testimonials obtained from manufacturing leaders. The Taguchi approach has become standard operating procedure in many businesses, particularly in the automotive and electronics industries. The successes created by using Tuguchi's approach have led to increased interest and application of the broader array of design of experiment technologies discussed throughout this book.

Two ideas introduced by Taguchi have been fundamental to the success of his approach. They are the idea of a quality loss function and the idea that products can be designed to minimize quality loss and, hence, improve robustness. Figure 12.3 shows a schematic of quality loss as a function of a product property. Quality loss is assessed on a continuous scale of the product property

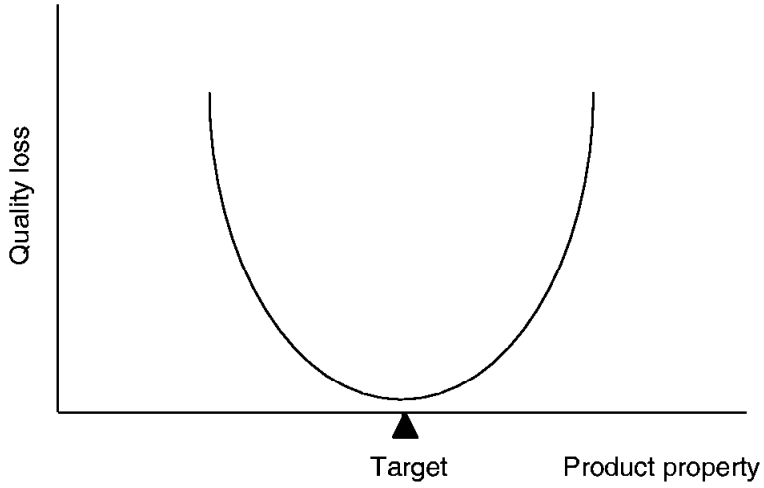


Figure 12.3 Taguchi's quality loss concept.

from the target value and not simply on whether the property is either inside or outside a set of specification limits. *Any deviation from the target value results in quality loss.* The quadratic loss function shown in Figure 12.3 is typically used because this leads to a simple analysis for optimizing the product.

A goal of the study of quality loss is to minimize loss. If a product can be made to be robust to uncontrolled process factors, deviation of the product property from the target value should be minimal. Hence, quality loss should be minimal. The robustness of a product to uncontrolled factors depends on both the value of the product property target and the amount of variation of the product property around the target. Unsatisfactory variability due to bias or uncontrolled variation leads to the need to study the process further. To this end, Taguchi promoted statistical experimental design using a natural approach that was far less technical than that traditionally offered.

Taguchi provided tables of *orthogonal arrays* to be used as statistical designs. These orthogonal array designs are well suited for estimating main effects. They have the same purpose as the screening designs introduced in Chapter 7. In practice, many of these orthogonal array designs are *saturated.* Saturated screening designs have numbers of test runs equal to one less than the number of factors. They are Resolution-III designs (Section 7.2) with all main effects aliased with interaction effects. No estimate of experimental error variation can be calculated with saturated designs.

Some of the more popular Taguchi orthogonal array designs are shown in the appendix to this chapter. The shorthand notation of L_x for these tables derives from the fact that many of these designs are derivable from latin-square designs (Section 9.4.1). The label L_x indicates that the design has x test runs in the experiment. Consistent with the convention introduced in Chapter 5, -1 represents the low level of a factor variable and a $+1$ represents the high level.

TABLE 12.1 Numbers of Factors for Select Orthogonal-Array Designs

Design	Two-Level Factors	Three-Level Factors
L_4	2 to 3	0
L_8	2 to 7	0
L_9	0	2 to 4
L_{16}	2 to 15	0
L_{18}	0	2 to 13
L_{27}	1	1 to 7

A 0 denotes the middle factor level. Alternatively, these tables can be reexpressed with 1, 2, and 3 designating the low, middle, and high factor levels.

Table 12.1 indicates the number of factors that can be investigated in experiments using each of the designs that are tabulated in the appendix. Note that some of the designs are limited to two-level factors, some to three-level factors, and one can have a mixture of two- and three-level factors Additional designs are available and include some that can be used when factors have four or five levels.

In his robust design approach, Taguchi uses these statistical designs in an inner- and outer-array composite design that includes both controllable process factors and uncontrollable process factors. The goal of experiments using inner-outer array designs is to discover controllable factor levels that render key process properties relatively unaffected by changes in the uncontrollable factor levels. Taguchi thus recommends selecting a set of controllable process factors and a set of factors that are uncontrollable during normal operation of the process (e.g., temperature and humidity) but that can be controlled during a process improvement experiment (e.g., in a temperature- and humidity-controlled laboratory). The two sets of factors are called control and environmental factors. The names control factors and environmental factors have several alternatives, a few of which are shown in Exhibit 12.2.

EXHIBIT 12.2 ALTERNATIVE LABELS FOR CONTROL AND ENVIRONMENTAL FACTORS

Environmental Factors	Control Factors
Uncontrolled variables	Design variables
Noise variables	Process variables
Ambient variables	Ingredients
Raw material	Formula variables
Customer use	

Control factors are controlled during production and use. Environmental factors are those that affect product or process functionality but are not controlled during production and use. Both types of factors, environmental and control, must be systematically varied during experimentation. Rather than use randomization techniques to minimize the possible effects of environmental factors when trying to understand the relationships of the controls factors to the response, the environmental factors are incorporated explicitly into the design.

Once the factors to be studied are categorized, an inner array and outer array are chosen. The inner array is an orthogonal array chosen for the control factors. An outer array is an orthogonal array chosen for the environmental factors. The complete design layout is obtained by crossing the inner and outer arrays. That is, the entire design for the environmental factors (outer array) is repeated at each design point of the control factors (inner array).

Lorenzen and Villalobos (1990) provide an interesting example of the use of inner and outer arrays. In this experiment, engineers studied the amount of torque required to separate the receiving yoke from the tube of an intermediate shaft steering column that connects the steering wheel to the power steering motor. Six factors were of interest: four control factors and two environmental factors. Each of the factors were studied at two levels. Table 12.2 lists the factors and their levels.

Using this example to illustrate the Taguchi approach to robust design, an L_8 inner array can be used with an L_4 outer array (the actual design used a different inner array, a half fraction of a 2^4 factorial). The crossed array requires a total of 32 test runs. Figure 12.4 illustrates the design points for this example. The analysis of this type of robust design will be detailed in Chapter 13.

TABLE 12.2 Factor Levels for the Steering Column Example

	Factor Levels	
Control Factors	−1	1
Pocket depth	Shallow	Deep
Yoke concentricity	Off center	On center
Tube length	Short	Long
Power setting	80	87
Environmental Factors		
Clearance	Maximum	Minimum
Line voltage	Low	Best

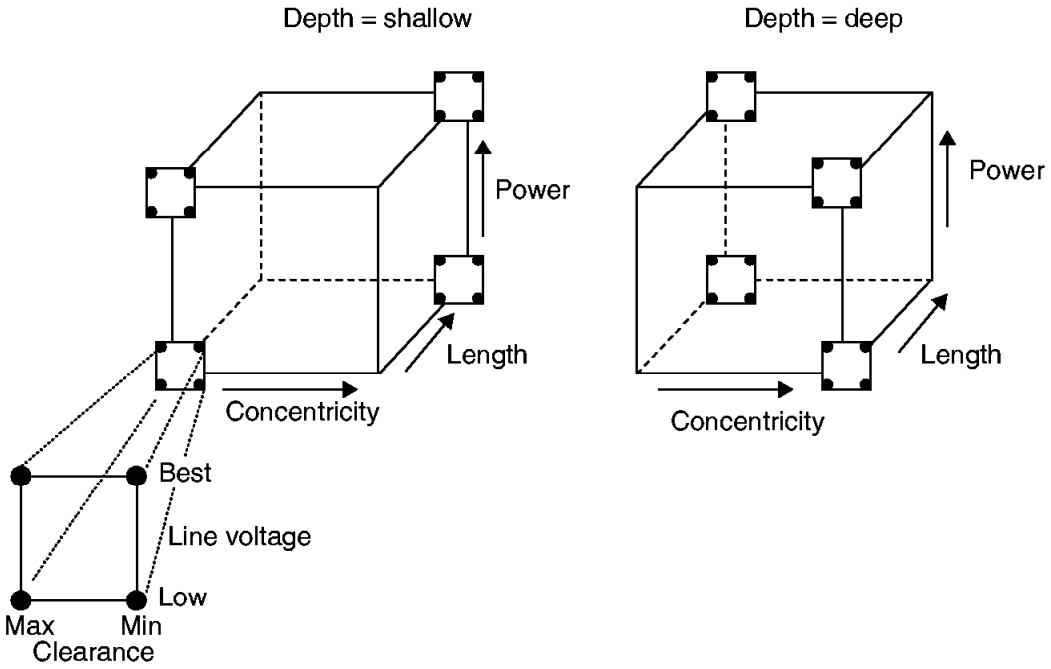


Figure 12.4 Taguchi L_8 inner- and L_4 outer-array design geometry for the steering column example.

12.2.2 An Integrated Approach

Statisticians, engineers, and scientists have widely endorsed Taguchi's concepts of a quality loss function and robust product design. He has also been applauded for bringing a structured design of experiment discipline to the engineering design process. However, the design of experiment methods described in the previous section that are used to put these concepts into practice have been strongly criticized. A number of alternative approaches for executing the experimental design strategy of robust product design are documented in the references at the end of this chapter. We outline an alternative that we believe will generally deliver less expensive and more reliable results than the Taguchi approach.

We advocate an integrated approach that combines the control and environmental factors. Complete factorial and fractional factorial experiments (Chapters 5 and 7), or classes of designs called *response surface designs* that are introduced in this chapter and covered in detail in Chapter 17 are used to study the combined set of control and environmental factors. This integrated approach allows for obtaining fundamental process understanding and permits greater flexibility for studying process robustness. A primary goal of the integrated approach is to design the experiment so that an empirical equation can be estimated that relates the process properties and the control and environmental factors throughout the experimental region. This empirical equation,

the estimated response function or response surface, allows one to estimate the effects that the control and environmental factors have on the process properties and allows a set of control factor levels to be selected that minimize the variation due to the environmental factor levels while achieving the desired targeted process property. Exhibit 12.3 summarizes this approach.

EXHIBIT 12.3 INTEGRATED APPROACH

1. Select the control (C) and environmental (E) factors of interest. All factors must be able to be systematically changed during the experiment.
2. If the total number of factors is large (greater than 6 or 7), use screening experiments to identify important control and environmental factors.
3. Use complete or fractional factorial experiments or response surface designs to develop interaction and curvature information; express this information as $y = f(C, E)$, where y represents the response (process or product property of interest) and f is a function relating the response and the factor levels.
4. Understand and exploit interactions among control factors and environmental factors ($C \times C$ and $C \times E$) to achieve product functionality with minimum variation. Control-factor ($C \times C$) interaction levels can be chosen to locate the response at the desired target. Control ($C \times C$) and control-environmental ($C \times E$) interactions can be used to reduce transmitted variation. Trade-offs are usually necessary.

The integrated approach quantifies two types of variation: experimental error (introduced in Chapter 4) and transmitted variation due to either control factor tolerances or environmental factor variability. Figures 12.5 and 12.6 illustrate these two types of variation. Figure 12.5(a) shows the true relationship between a response, y, and a factor, F_1. The desired target is 100 and that value occurs when F_1 equals x_1. Figure 12.5(b) shows the variation in y due to experimental error. The effect of factor level x_1 is contaminated by experimental error so that the observed value of y will be in a range of values centered at 100.

Figure 12.6(a) demonstrates the variation that is transmitted due to tolerances ($\pm T$) of an environmental factor, E. In the absence of experimental error, the observed value of y is again in a range, but this time it is due to the inability to control exactly the environmental factor E. Figure 12.6(b) shows how the interaction relationship between two control factors, C_1 and C_2, can be exploited to reduce transmitted variation. Factor C_1 is changed from the value c_1 to c_1^*; yielding the lower curve shown in Figure 12.6(b) and a narrower range on y than the curve for c_1. Factor C_2 is then increased to c_2^*, which further narrows the range of transmitted variation in y and gives the desired target of 100.

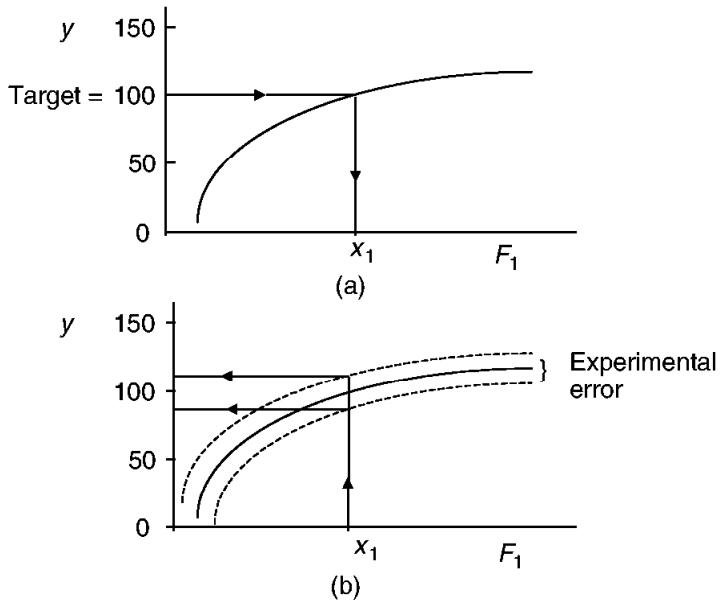


Figure 12.5 Illustration of experimental error variation.

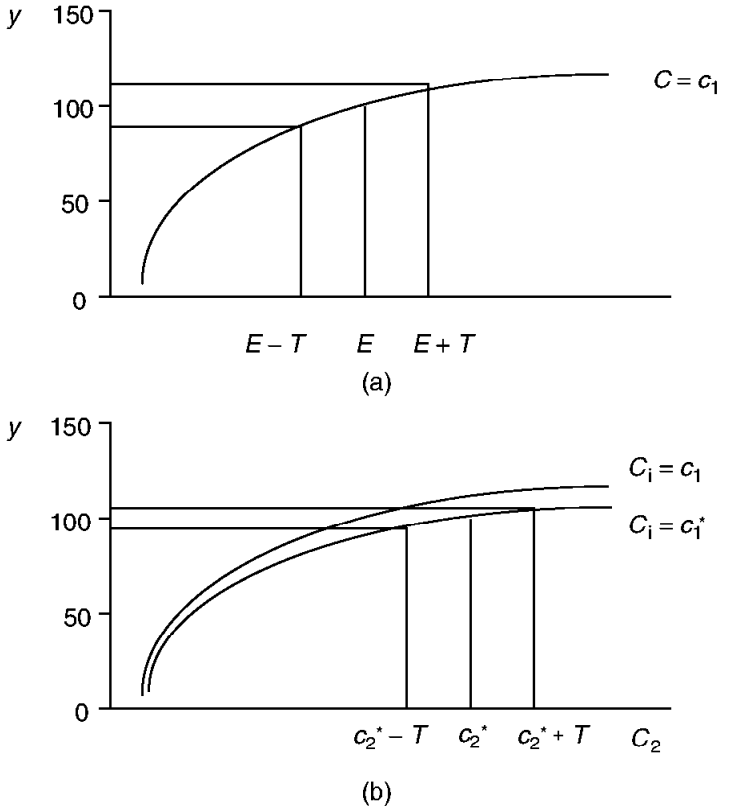


Figure 12.6 Illustration of transmitted variation.

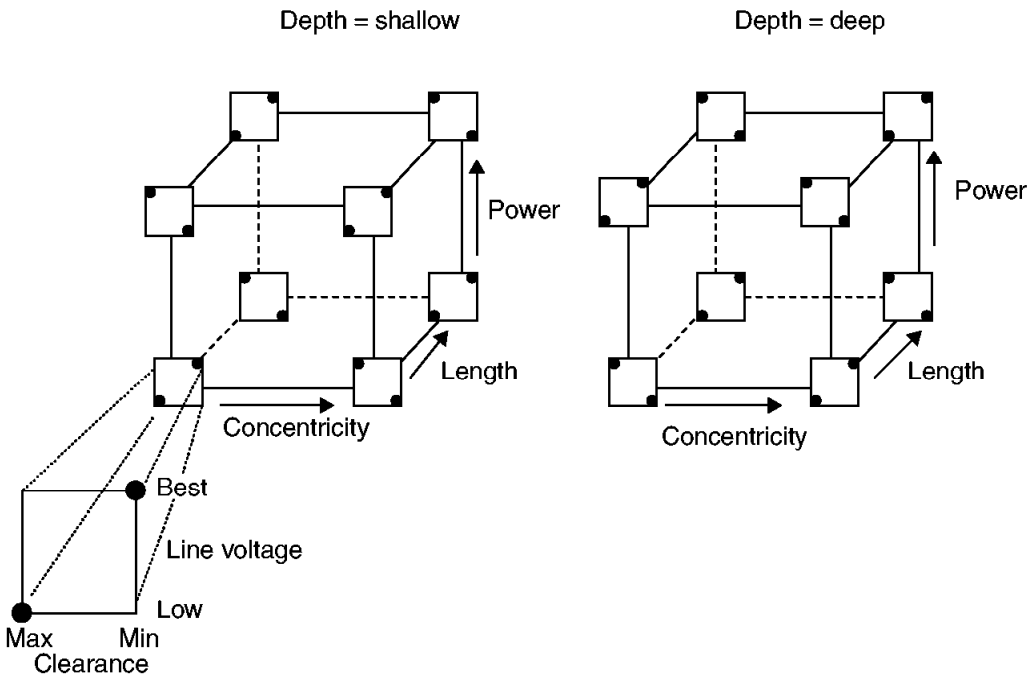


Figure 12.7 Integrated design geometry for steering column example.

Complete and fractional factorial experiments were introduced and discussed in detail in Chapters 5 and 7, respectively. These experiments are available in abundance for factors with two levels and, to a more limited degree, for three-level factors. They allow main effects, interactions, and experimental error variation to be estimated. Response surface designs allow for factors with more than two levels. They are so named because they provide the data to fit a model that allows the response to be graphed as a curve in one dimension (one factor) or a surface in two or more dimensions (two or more factors). Response surface designs allow estimation of curved response surfaces. Some efficiency in the required number of factor–level combinations is also achieved relative to factorial experiments. Two classes of response surface designs, central composite, and Box–Behnken, are detailed in Chapter 17.

This integrated approach is now demonstrated with the steering column example introduced in Section 12.2.1. All six factors were studied at two levels. A half fraction of a 2^6 factorial experiment entails the same number of test runs as the Taguchi approach, 32 test runs. Figure 12.7 shows the design points for the integrated design. The half-fraction run in a completely randomized design is of Resolution VI: all main effects and two-factor interactions can be estimated. No two-factor interactions among the control factors can be estimated with the design in Figure 12.4. Experimental error variation can also be estimated with the integrated design. The analysis of this type of robust design will be described in Chapters 13 and 17.

APPENDIX: SELECTED ORTHOGONAL ARRAYS

L_4

	Factor		
Run	A	B	C
1	−1	−1	−1
2	−1	1	1
3	1	−1	1
4	1	1	−1

L_8

	Factor						
Run	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

L_9

	Factor			
Run	A	B	C	D
1	−1	−1	−1	−1
2	−1	0	0	0
3	−1	1	1	1
4	0	−1	0	1
5	0	0	1	−1
6	0	1	−1	0
7	1	−1	1	0
8	1	0	−1	1
9	1	1	0	−1

L_{16}

	Factor														
Run	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
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	1	1	1	1	1	1	1
3	−1	−1	−1	1	1	1	1	−1	−1	−1	−1	1	1	1	1
4	−1	−1	−1	1	1	1	1	1	1	1	1	−1	−1	−1	−1
5	−1	1	1	−1	−1	1	1	−1	−1	1	1	−1	−1	1	1
6	−1	1	1	−1	−1	1	1	1	1	−1	−1	1	1	−1	−1
7	−1	1	1	1	1	−1	−1	−1	−1	1	1	1	1	−1	−1
8	−1	1	1	1	1	−1	−1	1	1	−1	−1	−1	−1	1	1
9	1	−1	1	−1	1	−1	1	−1	1	−1	1	−1	1	−1	1
10	1	−1	1	−1	1	−1	1	1	−1	1	−1	1	−1	1	−1
11	1	−1	1	1	−1	1	−1	−1	1	−1	1	1	−1	1	−1
12	1	−1	1	1	−1	1	−1	1	−1	1	−1	−1	1	−1	1
13	1	1	−1	−1	1	1	−1	−1	1	1	−1	−1	1	1	−1
14	1	1	−1	−1	1	1	−1	1	−1	−1	1	1	−1	−1	1
15	1	1	−1	1	−1	−1	1	−1	1	1	−1	1	−1	−1	1
16	1	1	−1	1	−1	−1	1	1	−1	−1	1	−1	1	1	−1

L_{18}

	Factor							
Run	A	B	C	D	E	F	G	H
1	−1	−1	−1	−1	−1	−1	−1	−1
2	−1	−1	0	0	0	0	0	0
3	−1	−1	1	1	1	1	1	1
4	−1	0	−1	−1	0	0	1	1
5	−1	0	0	0	1	1	−1	−1
6	−1	0	1	1	−1	−1	0	0
7	−1	1	−1	0	−1	1	0	1
8	−1	1	0	1	0	−1	1	−1
9	−1	1	1	−1	1	0	−1	0
10	1	−1	−1	1	1	0	0	−1
11	1	−1	0	−1	−1	1	1	0
12	1	−1	1	0	0	−1	−1	1
13	1	0	−1	0	1	−1	1	0
14	1	0	0	1	−1	0	−1	1
15	1	0	1	−1	0	1	0	−1
16	1	1	−1	1	0	1	−1	0
17	1	1	0	−1	1	−1	0	1
18	1	1	1	0	−1	0	1	−1

L_{27}

	Factor												
Run	A	B	C	D	E	F	G	H	I	J	K	L	M
1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
2	−1	−1	−1	−1	0	0	0	0	0	0	0	0	0
3	−1	−1	−1	−1	1	1	1	1	1	1	1	1	1
4	−1	0	0	0	−1	−1	−1	0	0	0	1	1	1
5	−1	0	0	0	0	0	0	1	1	1	−1	−1	−1
6	−1	0	0	0	1	1	1	−1	−1	−1	0	0	0
7	−1	1	1	1	−1	−1	−1	1	1	1	0	0	0
8	−1	1	1	1	0	0	0	−1	−1	−1	1	1	1
9	−1	1	1	1	1	1	1	0	0	0	−1	−1	−1
10	0	−1	0	1	−1	0	1	−1	0	1	−1	0	1
11	0	−1	0	1	0	1	−1	0	1	−1	0	1	−1
12	0	−1	0	1	1	−1	0	1	−1	0	1	−1	0
13	0	0	1	−1	−1	0	1	0	1	−1	1	−1	0
14	0	0	1	−1	0	1	−1	1	−1	0	−1	0	1
15	0	0	1	−1	1	−1	0	−1	0	1	0	1	−1
16	0	1	−1	0	−1	0	1	1	−1	0	0	1	−1
17	0	1	−1	0	0	1	−1	−1	0	1	1	−1	0
18	0	1	−1	0	1	−1	0	0	1	−1	−1	0	1
19	1	−1	1	0	−1	1	0	−1	1	0	−1	1	0
20	1	−1	1	0	0	−1	1	0	−1	1	0	−1	1
21	1	−1	1	0	1	0	−1	1	0	−1	1	0	−1
22	1	0	−1	1	−1	1	0	0	−1	1	1	0	−1
23	1	0	−1	1	0	−1	1	1	0	−1	−1	1	0
24	1	0	−1	1	1	0	−1	−1	1	0	0	−1	1
25	1	1	0	−1	−1	1	0	1	0	−1	0	−1	1
26	1	1	0	−1	0	−1	1	−1	1	0	1	0	−1
27	1	1	0	−1	1	0	−1	0	−1	1	−1	1	0

REFERENCES

Text References

Discussions of gage repeatability and reproducibility, and process capability are covered in the following references:

Leitnaker, M. G., Sanders, R. D., and Hild, C. (1996). *The Power of Statistical Thinking: Improving Industrial Processes*. New York: Addison-Wesley Publishing Company.

Measurement Systems Analysis Reference Manual (1995). Detroit, MI: Automotive Industry Action Group (AIAG).

Pitt, H. (1999). *SPC for the Rest of Us*. New York: K.W. Tunnell Company, Inc.

The following texts present Taguchi's approach to robust design:

DeVor, R. E., Chang, T., and Sutherland, J. W. (1992). *Statistical Quality Design and Control: Contemporary Concepts and Methods*, New York: Macmillan Publishing Company.

Ealey, L. A. (1988). *Quality by Design: Taguchi Methods and U.S. Industry*, Dearborn, MI: ASI Press, American Supplier Institute, Inc.

Lochner, R. H. and Matar, J. E. (1990). *Designing for Quality*, New York: Chapman and Hall.

Schmidt, S. R. and Launsby, R. G. (1989). *Understanding Industrial Designed Experiments*, 2nd Ed., Longmont, CO: CQG Ltd Printing.

Taguchi, G. (1986). *Introduction to Quality Engineering*, Tokyo: Asian Productivity Organization.

Taguchi, G. and Konishi, S. (1987). *Taguchi Methods: Orthogonal Arrays and Linear Graphs*, Dearborn, MI: American Supplier Institute.

Wu, C. E. J. and Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. New York: John Wiley & Sons, Inc.

The following texts and references present Taguchi's approach along with alternative methods recommended by the authors:

Lorenzen, T. J. and Villalobos, M. A. (1990). "Understanding Robust Design, Loss Functions, and Signal to Noise Ratios," Research Publication, General Motors Research Laboratories, GMR-7118, Warren, MI.

Lucas, J. M. (1994). "How to Achieve a Robust Process Using Response Surface Methodology," *Journal of Quality Technology*, **26**, 248–260.

Montgomery, D. C. (1997). *Design and Analysis of Experiments*, 4th Ed., New York, NY: John Wiley & Sons, Inc.

Myers, R. H., Khuri, A. I., and Vining, G. (1992). "Response Surface Alternatives to the Taguchi Robust Parameter Design Approach," *The American Statistician*, **46**, 131–139.

Shoemaker, A. C., Tsui, K., and Wu, C. F. J. (1991). "Economical Experimentation Methods for Robust Design," *Technometrics*, **33**, 415–427.

Steinberg, D. M. and Bursztyn, D. (1994). "Dispersion Effects in Robust-Design Experiments with Noise Factors," *Journal of Quality Technology*, **26**, 12–20.

Vining, G. and Myers, R. H. (1990). "Combining Taguchi and Response Surface Philosophies: A Dual Response Approach," *Journal of Quality Technology*, **22**, 38–45.

Wheeler, D. J. (1990). *Understanding Industrial Experimentation*, Knoxville, TN: SPC Press, Inc.

Figures 12.5 and 12.6 are adapted from the following reference:

Bailey, S. P., Chatto, K. A., Fellner, W. H., and Pfeifer, C. G. (1990). "Giving Your Response Surface a Robust Workout," Proceedings of the 34th Annual Fall Technical Conference, Richmond, VA.

EXERCISES

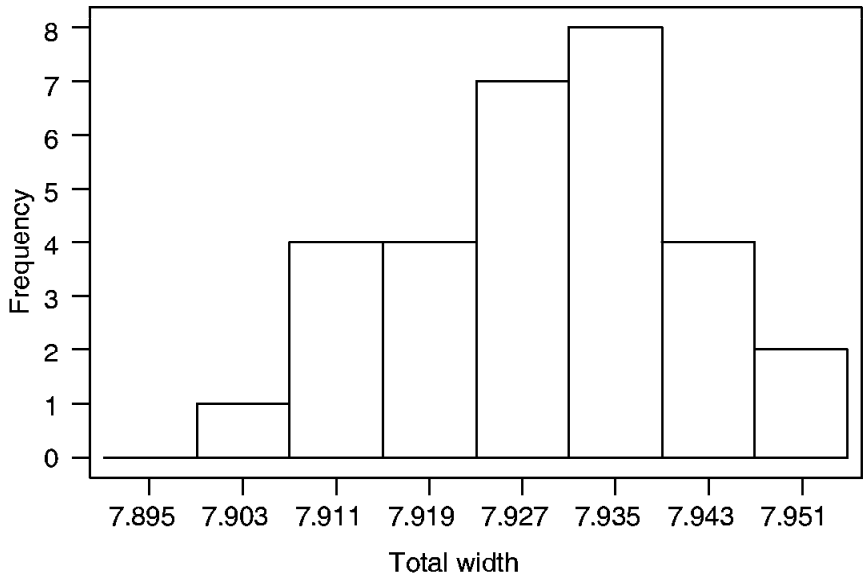
1 The lean of an innerspring mattress unit is measured with a special gage that determines departure from vertical to the nearest one-eighth inch. The lean can be measured for the side, end, and corner of the innerspring unit. Three operators typically make these measurements on the production floor: Ken, Mick, and Tammy. Design a Gage R&R study for this measurement method, and provide instructions for collecting the data.

2 A machine-bored diameter on a die-cast aluminum part is measured with a gage that reads to the nearest thousandth of a millimeter. One technician from each of three manufacturing shifts typically measures bore diameter. This particular diameter is a critical quality characteristic (the specification is 31.2 ± 0.08 mm) and the engineer wants to be sure the gage is adequate. Design a Gage R&R study for this measurement method and provide instructions for collecting the data.

3 Comment on the following statement: A Gage R&R study allows you to quantify all the sources of variation inherent in a measurement method.

4 A machine capability study is desired on a machine that makes formed-wire modules used in the construction of box springs. A consistent total width of the module is essential for efficient assembly of box spring units. The specifications for this product are a minimum total width of 7.75 inches and a maximum of 8.00 inches. Describe how you would collect data to assess the capability of this machine. Why would this be of interest to the customer who assembles box spring units?

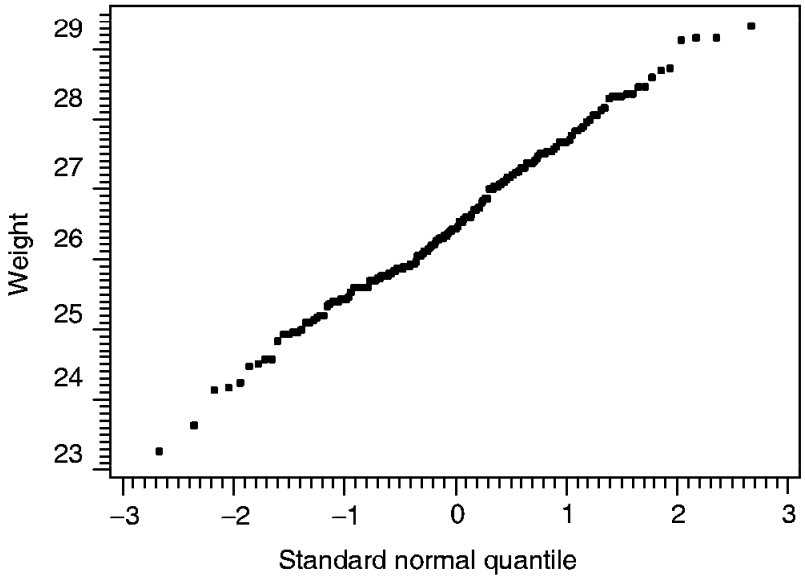
5 For the scenario described in Exercise 4, a quality manager collected thirty consecutive samples from the module machine and the total widths were recorded. The following histogram was constructed from the data.



Compare the histogram to the specifications and comment on the capability of this machine.

6 The manager of a nonwoven fabric plant wanted to know if the plant's process (run on two shifts) was capable relative to the weight of the product. Weight is an important characteristic because it is related to the consumer's perception of a quality product. The lower specification limit on product weight is 24 grams per square foot. Design a process capability study to see if the plant is able to satisfy the weight specification. Will the data collected represent the product the customer is receiving? Why or why not?

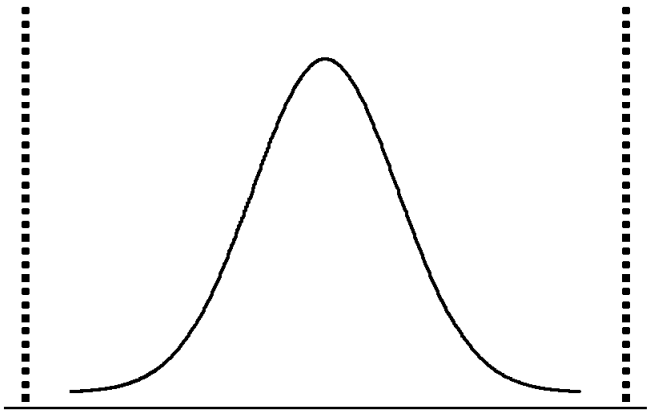
7 For the situation described in Exercise 6, a product engineer collected data every two hours (roughly) over a four-week period. The data are displayed in the following normal quantile-quantile plot. Compare the information in the plot to the specifications and comment on the capability of this machine. Does this data represent the product the customer is receiving? Why or why not?



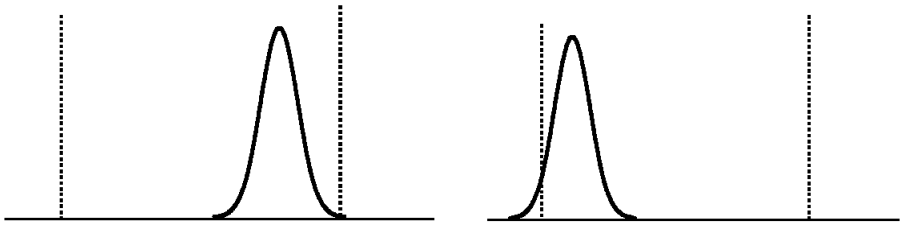
8 The sampling plan used in a capability study includes decisions about the sampling frequency, the number of observations collected (both total and at each sample time), and the elapsed time over which the samples for the study are obtained. Discuss the effect that the sampling plan has on the interpretation of the results of a capability study. Frame your discussion in the context of sources of variation and variance components that were introduced in Chapter 10.

9 The following distribution represents individual observations from a manufacturing process. The vertical dashed lines are the specifications on this

product characteristic. Is the process capable? Why or why not? Suppose this had been a distribution of averages of 10 observations from the process. Would this change your conclusion? How or how not?



10 Two distributions follow for a product characteristic from a manufacturing process. The distributions represent individual observations from the process. The vertical dashed lines are the specifications on this product characteristic. Comment on the process capability for the two situations.



11 A team of engineers and operators has been assembled to address the shrinkage of an extruded plastic product — they want to minimize it. They have identified the following variables that they think may affect shrinkage: air pressure, screw speed, die temperature, cooling tank water temperature, amount of regrind material, and ambient temperature at the machine. Each of these factors can be tested at two levels, and it is not possible to replicate the test conditions. Identify which of these factors are control factors and which are environmental factors. Using coded factor levels construct (a) a Taguchi design and (b) a design using the integrated approach.

12 Discuss the use of Taguchi designs and integrated designs for the following three cases:

(a) All the factors are control factors with no replicates,

(b) All the factors are control factors and replicates are available at each factor level combination in the experiment, and

(c) Some factors are control factors and some are environmental factors.

13 The accuracy of an online sensor for measuring solids is thought to be a function of the following variables:

Variable	Levels
Sight glass	Type A, type B
Orifice type	22 degrees, 25 degrees
Installation torque	120 lbs., 150 lbs.
Line location	Vertical, 10° off vertical

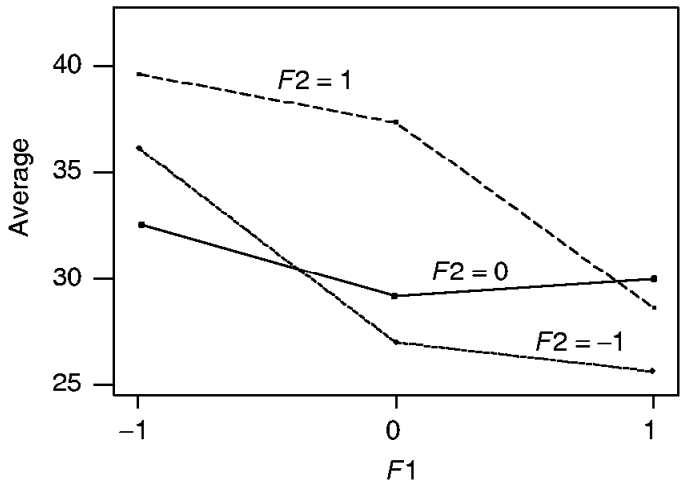
The first two variables are related to the manufacture of the sensor and the remaining two are determined at installation. Identify which of these factors are control factors and which are environmental factors. Using the given factor levels construct (a) a Taguchi design and (b) a design using the integrated approach.

14 Over spray during the final lacquer coat application at a wood furniture plant appears to be related to filter age, line speed, fluid pressure, lacquer temperature, spray tip age, and spray tip size. It is desired to conduct an experiment to test this conventional wisdom and gain insight into how to control over spray. Identify which of these factors are control factors and which are environmental factors. Using coded factor levels construct a two-level design using the integrated approach.

15 A study is being planned to evaluate the top-end performance of racing motorcycles. The following factors are to be included in the study: drive chain, fuel injection, exhaust, air intake, timing plate, cylinder heads, assembly mechanic, and test track. The first six factors are tested with both the standard equipment and with the enhanced equipment. Identify which of these factors are control factors and which are environmental factors. Using coded factor levels construct a design using the integrated approach.

16 In Exercise 11 suppose that three levels of each factor are of interest. How would this change your answers?

17 The interaction plot shown below shows the relationship between the average response on the vertical axis and two quantitative factors, $F1$ and $F2$.



The desired target value for this process is 30. What combination of factor levels will give you the most robust process? Why?

CHAPTER 13

Analysis of Nested Designs and Designs for Process Improvement

In this chapter the analysis of several special types of designs are discussed. These designs can have fixed or random effects; some have crossed factors, while some have nested factors. Analyses for the following designs are detailed in this chapter:

- *Hierarchically nested designs,*
- *Split-plot designs,*
- *Gage repeatability and reproducibility designs, and*
- *Orthogonal array designs.*

Nested designs and special designs for process improvement were introduced in the last two chapters. Each of these designs can be analyzed using the basic analysis of variance, multiple comparison, and variance component analyses that were discussed in Chapters 6, 8, and 10. The modifications to these basic analyses that are needed for nested and process improvement experimental designs are introduced in this chapter.

13.1 HIERARCHICALLY NESTED DESIGNS

Nested designs (Chapter 11) are designs in which the levels of one or more factors differ with each level of one or more other factors. With hierarchically nested designs, each level of nesting is imbedded within a previous level of nesting. The polyethylene density experiment discussed in Section 11.1 (see Figure 11.2) is an example of a hierarchically nested design.

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

Hierarchically nested designs are most frequently used with random factor levels. The analysis we present in this section requires that all but the first factor be a random factor. There are situations in which one or more of the nested factors correspond to fixed factor levels. The analysis of such designs depends on which factors are fixed and which are random. Computation of expected mean squares using the rules in the appendix to Chapter 10 is necessary to determine appropriate F-statistics for testing hypotheses and for estimating variance components for random effects.

Consider an experiment having three hierarchically nested factors A, B, and C. Suppose that factor C is nested within factor B and that factor B is nested within factor A. Assume that there are an equal number of replications at each level of nesting. The model for such an experiment can be written as follows:

$$y_{ijkl} = \mu + a_i + b_{j(i)} + c_{k(ij)} + e_{ijkl}$$
$$i = 1, 2, \ldots, a, \quad j = 1, 2, \ldots, b, \quad k = 1, 2, \ldots, c, \quad l = 1, 2, \ldots, r, \quad (13.1)$$

where the subscripts in parentheses denote the factor levels within which the leading factor level is nested; for example, the kth level of factor C is nested within the ith level of factor A and the jth level of factor B. Because Latin letters have been used to denote the factor levels, we are assuming that all factors are random and denote the variance components of the nested factors by σ_a^2, σ_b^2, and σ_c^2, respectively.

Sums of squares for nested factors are computed differently from those for crossed factors. Each nested sum of squares is computed as a sum of squared differences between the average response for a factor–level combination and the average of the responses at the previous level of nesting. For example, the sums of squares for the model (13.1) are

$$\mathrm{SS}_A = bcr \sum_i (\bar{y}_{i\cdots} - \bar{y}_{\cdots\cdot})^2, \quad \mathrm{SS}_{B(A)} = cr \sum_i \sum_j (\bar{y}_{ij\cdot\cdot} - \bar{y}_{i\cdots})^2,$$
$$\mathrm{SS}_{C(AB)} = r \sum_i \sum_j \sum_k (\bar{y}_{ijk\cdot} - \bar{y}_{ij\cdot\cdot})^2. \quad (13.2)$$

These nested sums of squares are related to the sums of squares for main effects and interactions, treating all the factors as crossed factors, as follows:

$$\mathrm{SS}_{B(A)} = \mathrm{SS}_B + \mathrm{SS}_{AB}, \qquad \mathrm{SS}_{C(AB)} = \mathrm{SS}_C + \mathrm{SS}_{AC} + \mathrm{SS}_{BC} + \mathrm{SS}_{ABC}.$$

Table 13.1 displays a symbolic ANOVA table, along with the expected mean squares, for this hierarchically nested model. Hierarchically nested models having more than three factors have ANOVA tables with the same general

TABLE 13.1 Symbolic ANOVA Table for a Three-Factor, Hierarchically Nested Random-Effect Model

ANOVA

Source	df	SS	MS	Expected Mean Square
A	$a-1$	SS_A	MS_A	$\sigma^2 + r\sigma_c^2 + cr\sigma_b^2 + bcr\sigma_a^2$
$B(A)$	$a(b-1)$	$\mathrm{SS}_{B(A)}$	$\mathrm{MS}_{B(A)}$	$\sigma^2 + r\sigma_c^2 + cr\sigma_b^2$
$C(AB)$	$ab(c-1)$	$\mathrm{SS}_{C(AB)}$	$\mathrm{MS}_{C(AB)}$	$\sigma^2 + r\sigma_c^2$
Error	$abc(r-1)$	SS_E	MS_E	σ^2
Total	$abcr-1$	TSS		

Computations

$$\mathrm{TSS} = \sum_i \sum_j \sum_k \sum_l y_{ijkl}^2 - \mathrm{SS}_M, \quad \mathrm{SS}_M = n^{-1} y_{\bullet\bullet\bullet\bullet}^2$$

$$\mathrm{SS}_A = \sum_i \frac{y_{i\bullet\bullet\bullet}^2}{bcr} - \mathrm{SS}_M$$

$$\mathrm{SS}_{B(A)} = \sum_i \sum_j \frac{y_{ij\bullet\bullet}^2}{cr} - \sum_i \frac{y_{i\bullet\bullet\bullet}^2}{bcr}$$

$$\mathrm{SS}_{C(AB)} = \sum_i \sum_j \sum_k \frac{y_{ijk\bullet}^2}{r} - \sum_i \sum_j \frac{y_{ij\bullet\bullet}^2}{cr}$$

$$\mathrm{SS}_E = \mathrm{TSS} - \mathrm{SS}_A - \mathrm{SS}_{B(A)} - \mathrm{SS}_{C(AB)}$$

$$y_{\bullet\bullet\bullet\bullet} = \sum_i \sum_j \sum_k \sum_l y_{ijkl}, \quad y_{i\bullet\bullet\bullet} = \sum_j \sum_k \sum_l y_{ijkl}, \quad \text{etc.}$$

pattern. Note that each expected mean square contains all the terms of the expected mean square that follows it in the table. Because of this pattern, to test the statistical significance of any factor, the appropriate F-statistic is the ratio of the mean square for the factor of interest divided by the mean square for the factor that follows it. For example, to test the hypotheses

$$H_0: \sigma_b^2 = 0 \quad \text{vs} \quad H_a: \sigma_b^2 \neq 0,$$

the appropriate F-ratio is $F = \mathrm{MS}_{B(A)}/\mathrm{MS}_{C(AB)}$.

Estimation of the variance components for the various model factors follows the general procedures outlined in Section 10.2. Estimation of these variance components is especially easy for hierarchically nested factors. For three-factor random-effects models one simply uses differences between the mean square for the factor of interest and the one following it in Table 13.1. For example,

$$s_b^2 = \frac{\mathrm{MS}_{B(A)} - \mathrm{MS}_{C(AB)}}{rc}.$$

Staggered nested designs are an alternative to hierarchically nested designs when it is desired to obtain more evenly distributed information on each of the factors. A staggered nested design for a polymerization process study was described in Section 11.2 (see Figure 11.5). The data from this study are shown in Table 13.2. Note that there are two boxes selected from each of

TABLE 13.2 Strength Data for Polymerization Process Study

Strength	Lot	Box	Prep.	Strength	Lot	Box	Prep.
11.91	1	1	1	4.08	29	1	1
9.76	1	1	2	4.88	29	1	2
9.24	1	1	2	4.96	29	1	2
9.02	1	2	1	4.76	29	2	1
10.00	7	1	1	6.73	30	1	1
10.65	7	1	2	9.38	30	1	2
7.77	7	1	2	8.02	30	1	2
13.69	7	2	1	6.99	30	2	1
8.02	8	1	1	6.59	32	1	1
6.50	8	1	2	5.91	32	1	2
6.26	8	1	2	5.79	32	1	2
7.95	8	2	1	6.55	32	2	1
9.15	9	1	1	5.77	47	1	1
8.08	9	1	2	7.19	47	1	2
5.28	9	1	2	7.22	47	1	2
7.46	9	2	1	8.33	47	2	1
7.43	10	1	1	8.12	48	1	1
7.84	10	1	2	7.93	48	1	2
5.91	10	1	2	6.48	48	1	2
6.11	10	2	1	7.43	48	2	1
7.01	11	1	1	3.95	49	1	1
9.00	11	1	2	3.70	49	1	2
8.38	11	1	2	2.86	49	1	2
8.58	11	2	1	5.92	49	2	1
11.13	12	1	1	5.96	51	1	1
12.81	12	1	2	4.64	51	1	2
13.58	12	1	2	5.70	51	1	2
10.00	12	2	1	5.88	51	2	1
14.07	13	1	1	4.18	52	1	1
10.62	13	1	2	5.94	52	1	2
11.71	13	1	2	6.28	52	1	2
14.56	13	2	1	5.24	52	2	1
11.25	14	1	1	4.35	53	1	1
9.50	14	1	2	5.44	53	1	2
8.00	14	1	2	5.38	53	1	2

TABLE 13.2 *(continued)*

Strength	Lot	Box	Prep.	Strength	Lot	Box	Prep.
11.14	14	2	1	7.04	53	2	1
9.51	15	1	1	2.57	54	1	1
10.93	15	1	2	3.50	54	1	2
12.16	15	1	2	3.88	54	1	2
12.71	15	2	1	3.76	54	2	1
16.79	16	1	1	3.48	55	1	1
11.95	16	1	2	4.80	55	1	2
10.58	16	1	2	4.46	55	1	2
13.08	16	2	1	3.18	55	2	1
7.51	19	1	1	4.38	56	1	1
4.34	19	1	2	5.35	56	1	2
5.45	19	1	2	6.39	56	1	2
5.21	19	2	1	5.50	56	2	1
6.51	21	1	1	3.79	57	1	1
7.60	21	1	2	3.09	57	1	2
6.72	21	1	2	3.19	57	1	2
6.35	21	2	1	2.59	57	2	1
6.31	23	1	1	4.39	58	1	1
5.12	23	1	2	5.30	58	1	2
5.85	23	1	2	4.72	58	1	2
8.74	23	2	1	6.13	58	2	1
4.53	24	1	1	5.96	59	1	1
5.28	24	1	2	7.09	59	1	2
5.73	24	1	2	7.82	59	1	2
5.07	24	2	1	7.14	59	2	1

30 lots, but only one preparation is made from the second box of each lot, whereas two preparations are made from the first box. Likewise, only one strength test was made on preparation 1 from each box, but two strength tests were made on preparation 2 from the first box. By reducing the number of preparations and the number of strength tests (i.e., not having two preparations from each box and two strength tests for each preparation), the total number of tests was reduced from 240 to 120 (see Table 11.1). In spite of the reduction in the number of tests, there are still ample observations that can be used to assess the effects of the various experimental factors.

The analysis of staggered nested designs is similar to the analysis of hierarchically nested designs. The model (13.1), with A = lot, B = box, and C = preparation, is an appropriate model for the polymerization process data; however, because staggered nested designs are unbalanced, observations are not taken for all the responses indicated by the subscripts in the model (13.1).

TABLE 13.3 ANOVA Table for Polymerization-Process Data

Source	df	SS	MS	F	p-Value
Lots	29	855.96	29.52	17.67	0.000
Boxes (lots)	30	50.09	1.67	0.73	0.803
Preparations (lots, boxes)	30	68.44	2.28	3.52	0.000
Error	30	19.44	0.65		
Total	119	993.92			

A consequence of the imbalance is that the ANOVA table is constructed using the principle of reduction in error sums of squares.

Three models are fitted: one with lots as the only factor, one with lots and boxes, and one with all three factors. The first fit provides the sum of squares for the lots, the difference in error sums of squares for the first and second models provides the sum of squares for boxes, and the difference in error sums of squares for the last two fits gives the sum of squares for preparations. The error sum of squares for the ANOVA table is obtained from the last fit. The ANOVA table for the polymerization process study is shown in Table 13.3. Note from this table the nearly equal numbers of degrees of freedom for the three experimental factors. An ordinary hierarchical design would have many more degrees of freedom for the preparations than for the lots.

Estimation of variance components with staggered nested designs is more complicated than with hierarchically nested designs. The difficulty arises because of the imbalance in the staggered designs. The interested reader is referred to the references on staggered nested designs at the end of Chapter 11.

13.2 SPLIT-PLOT DESIGNS

Split-plot designs were introduced in Section 11.3 as designs in which crossed factors are nested within levels of other factors. Split-plot designs can be used when experimental units have different sizes, as in agricultural field trials in which some factors are assigned to entire farms and others to fields within the farms. Split-plot designs can also be used when there are different levels of randomization, as in the cylinder-wear example of Section 11.4. Because split-plot designs contain both crossed and nested factors, the analysis of such designs includes features of both factorial experiments and nested experiments.

In Section 11.3 (see also Section 9.1), an investigation of lawnmower automatic-cutoff times was described. The experiment was conducted in a split-plot design. Responses for each of two speeds were taken on each of three

lawnmowers from each of the two manufacturers. Figure 11.3 shows the nested and crossed relationships among the factors.

There are two fixed factors in this study (manufacturer and speed), and one random factor (lawnmower). The two fixed factors are crossed, and an interaction between the factors is possible. An ANOVA model for this experiment is

$$y_{ijkl} = \mu + \alpha_i + b_{j(i)} + \gamma_k + (\alpha\gamma)_{ik} + e_{ijkl}$$
$$i = 1, 2, \quad j = 1, 2, 3, \quad k = 1, 2, \quad l = 1, 2. \tag{13.3}$$

It is important to note that the above model indicates the random effects for lawnmowers are only nested within the levels of the manufacturers. The lawnmowers are not nested within the levels of speed, because all lawnmowers are tested at both speeds. Had some lawnmowers only been tested at the high speed and others only at the low speed, lawnmowers would be nested within the joint levels of manufacturer and speed. In that case the model would be written as

$$y_{ijkl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + c_{k(ij)} + e_{ijkl},$$

where factor B (fixed) would represent speed and factor C (random) would represent lawnmowers.

The mean of a response for model (13.3) is $\mu + \alpha_i + \gamma_k + (\alpha\gamma)_{ik}$, where the α_i and γ_k model terms correspond to the manufacturer and speed factors, respectively. The random lawnmower and error terms are assumed to be independently normally distributed with zero means. The $b_{j(i)}$ terms have standard deviation σ_b. The experimental error terms, e_{ijkl}, have standard deviation σ. If there is reason to suspect that the lawnmowers and the speed factors interact, another set of random terms representing the interaction could be added to the above model.

Note that the lawnmowers constitute the whole plots for this split-plot design. The repeat tests on the speeds for each lawnmower constitute the split plots. The whole-plot error, the variability due to replicate test runs on each lawnmower, consists of variation due to different lawnmowers, and to the uncontrollable experimental error. The split-plot error in this example is simply the uncontrollable experimental error. These components of variability are quantified in the expected mean squares.

Table 13.4 displays a symbolic ANOVA table for the experiment on lawnmower automatic-cutoff times. Note that both fixed and random effects and crossed and nested factors occur in this split-plot analysis. The expected mean squares identify the proper F-ratios for testing the factor effects. Observe that the manufacturers are tested against the whole-plot (lawnmower) error

TABLE 13.4 Symbolic ANOVA Table for Split-Plot Analysis of Automatic-Cutoff-Time Experiment

Source	df	SS	MS	F	Expected MS
Manufacturers (M)	$a-1$	SS_A	MS_A	$MS_A/MS_{B(A)}$	$\sigma^2 + cr\sigma_b^2 + bcrQ_\alpha$
Lawnmowers	$a(b-1)$	$SS_{B(A)}$	$MS_{B(A)}$	$MS_{B(A)}/MS_E$	$\sigma^2 + cr\sigma_b^2$
Speed (S)	$c-1$	SS_C	MS_C	MS_C/MS_E	$\sigma^2 + abrQ_\gamma$
$M \times S$	$(a-1)(c-1)$	SS_{AC}	MS_{AC}	MS_{AC}/MS_E	$\sigma^2 + brQ_{\alpha\beta}$
Error	$a(bcr-b-c+1)$	SS_E	MS_E		σ^2
Total	$abcr-1$	TSS			

TABLE 13.5 Lawnmower Automatic Cutoff Times

		Time (10^{-2} sec)				
Manufacturer	Lawnmower	Low Speed		High Speed		Average
A	1	211	230	278	278	249.25
	2	184	188	249	272	223.25
	3	216	232	275	271	248.50
B	4	205	217	247	251	230.00
	5	169	168	239	252	207.00
	6	200	187	261	242	222.50
Average		197.50	203.67	258.17	261.00	230.08

mean square, and the speed effects against the split-plot (experimental) error mean square.

Table 13.5 lists the cutoff times for the individual test runs. The ANOVA table for these data is shown in Table 13.6. The significance probabilities in the ANOVA table indicate that the interaction between manufacturers and speeds is not statistically significant. The main effect for manufacturers is not statistically significant at the 5% significance level; nevertheless, one may wish to examine the average cutoff times and compute a confidence interval for the difference in the mean cutoff times for the two manufacturers.

In the construction of a confidence interval for the difference of the means for the two manufacturers, the mean square for the lawnmowers is used in the confidence-interval formula, not the error mean square. The mean square used in confidence-interval procedures is always the same mean square that would be used in the denominator of an F-test for the significance of the effect.

TABLE 13.6 ANOVA Table for Split-Plot Analysis of Automatic Cutoff Times

Source	df	SS	MS	F	p-value
Manufacturer (M)	1	2,521.50	2,521.50	3.54	0.133
Lawnmower	4	2,852.83	713.21	5.39	0.006
Speed (S)	1	20,886.00	20,886.00	157.87	0.000
$M \times S$	1	10.67	10.67	0.08	0.780
Error	16	2,116.83	132.30		
Total	23	23,387.83			

Both the main effect for the speeds and the variability due to lawnmowers are statistically significant. Interestingly, the estimates of the two components of variability are approximately equal:

$$s_e = (132.30)^{1/2} = 11.50,$$
$$s_b = \left(\frac{713.21 - 132.30}{4}\right)^{1/2} = 12.05.$$

Thus, the variability incurred by using several lawnmowers in the experiment is of approximately the same magnitude as that from all the sources of uncontrolled variability.

If only one lawnmower had been obtained per manufacturer, no test of the manufacturer main effect would be possible unless the assumption of no whole-plot error could be made. From the above analysis, we see that such an assumption is unreasonable for this experiment, because the lawnmower effect is substantial. In experiments where combinations of levels of two or more factors are assigned to whole plots, tests of factor effects of interest can be obtained without assuming that there is no whole-plot error if the assumption of negligible higher-order interactions among the whole-plot factors is reasonable. Note that in the case of a two-factor factorial experiment in the whole plots, it would be necessary to assume that the two-factor interaction was negligible if no replication of the whole plot was included in the experiment.

The speed factor has only two levels in this experiment. The significant F-ratio for this effect indicates that the two factor levels produce significantly different effects on the cutoff times. For situations in which a factor that is studied at more than two levels is significant, or where a significant interaction is found, it may not be clear which pairs of factor–level averages are significantly different. Procedures such as those discussed in Chapter 6 can be used to determine which pairs of levels are statistically significant. The key to using these procedures is an estimate of the standard error of the difference of two averages.

Obtaining a standard-error estimate is more complicated for a split-plot design than for the other designs discussed, since the proper standard error depends on the averages being compared. There are four situations for distinguishing among averages that need to be considered: factor–level averages for a whole-plot factor, factor–level averages for a split-plot factor, interaction averages from a common level of a whole-plot factor, and interaction averages from different whole-plot factor levels. It was mentioned above that in general, the correct estimate of the error variance is the mean square used to test the significance of a factor effect. In Exhibit 13.1, we make this statement more explicit for the comparison of averages for a split-plot design.

EXHIBIT 13.1 STANDARD ERRORS FOR SPLIT-PLOT AVERAGES

1. For the comparison of two whole-plot averages,

$$\widehat{\mathrm{SE}}(\bar{y}_i - \bar{y}_j) = [\mathrm{MS}_{\mathrm{WP}}(n_i^{-1} + n_j^{-1})]^{1/2},$$

   where $\mathrm{MS}_{\mathrm{WP}}$ is the whole-plot error mean square.

2. For the comparison of two split-plot averages,

$$\widehat{\mathrm{SE}}(\bar{y}_i - \bar{y}_j) = [\mathrm{MS}_{\mathrm{SP}}(n_i^{-1} + n_j^{-1})]^{1/2},$$

   where $\mathrm{MS}_{\mathrm{SP}}$ is the split-plot error mean square, ordinarily MS_E.

3. For the comparison of two interaction averages, two cases must be considered.
   (a) If the two averages have the same level of the whole-plot factor, the standard error is estimated as in instruction 2.
   (b) If the two averages do not have the same level of the whole-plot factor, no exact standard-error formula exists. In place of $\mathrm{MS}_{\mathrm{SP}}$ in instruction 2, insert a Satterthwaite-type estimate of the sum of the subplot and whole-plot error variances.

The estimated standard errors in Exhibit 13.1 can be used in conjunction with any of the multiple-comparison procedures discussed in Section 6.4. The general approach outlined here can be extended to models that have additional levels of nesting, such as in split-split-plot experiments.

13.3 GAGE REPEATABILITY AND REPRODUCIBILITY DESIGNS

The Gage R&R design described in Section 12.1.1 was used to investigate the effects of operators (appraisers) and samples (parts) on the variability of an innerspring measurement process. The measured lengths of the innerspring units are shown in Figure 12.2. The two factors of interest in this study, operators and samples, are both random effects. It was decided to block on the day the measurement was obtained because of potential day-to-day differences. An appropriate model for this investigation is a randomized complete block design with two-factor random effects (and their interaction). The statistical model is similar to that of Equation (10.1) with a random main effect for

TABLE 13.7 ANOVA Table for the Gage R&R Study of Figure 12.2

Source	df	SS	MS	F	p-Value	Expected MS
Days	2	0.02228	0.01114	1.889	0.170	$\sigma^2 + 15\sigma_d^2$
Operator (O)	2	0.00159	0.00079	0.135	0.875	$\sigma^2 + 3\sigma_{OS}^2 + 15\sigma_O^2$
Sample (S)	4	5.12365	1.28091	217.163	0.000	$\sigma^2 + 3\sigma_{OS}^2 + 9\sigma_S^2$
O × S	8	0.05124	0.00640	1.086	0.401	$\sigma^2 + 3\sigma_{OS}^2$
Error	28	0.16516	0.00590			σ^2
Total	44	5.36392				

blocks (days) added. An ANOVA table corresponding to this model is shown in Table 13.7.

Denote the part-to-part (sample) variance component by σ_s^2, the appraiser (operator) variance by σ_o^2, their interaction by σ_{os}^2, the block (day) variance by σ_d^2, and the error variance by σ^2. In the context of Gage R&R studies, the measurement variation is composed of variance components for repeatability, σ^2, and reproducibility, $\sigma_O^2 + \sigma_{OS}^2$. If the appraiser by part interaction is not statistically significant, common practice dictates that the variance component for the interaction is set equal to zero. Then the other variance components are estimated using appropriate expected mean squares (see Chapter 10) with σ_{OS}^2 set equal to zero. The last column of Table 13.7 shows the expected means squares for this analysis.

The ANOVA Table 13.7 for this study shows no significant appraiser × part interaction. Hence, the operator by sample interaction variance component in the last column in Table 13.7 is set to zero. The other variance components are then estimated by equating the mean squares to the expected mean squares and solving. Thus, $\hat{\sigma}_{OS} = 0$,

$$\hat{\sigma} = \sqrt{0.00590} \quad \text{and} \quad \hat{\sigma}_O = \sqrt{(0.00079 - 0.00590)/15} \equiv 0.$$

The estimated operator variance component is set to zero because the error mean square is larger than the operator mean square. These results are often summarized in a table similar to Table 13.8. Note from Table 13.8 that in Gage R&R studies it is the estimated standard deviations (square roots of the estimated variance components) that are of interest rather than the ANOVA table itself.

Figure 13.1 shows two scatter plots that are useful in interpreting Gage R&R study results: plots of the innerspring lengths for each operator and for each sample part, with averages superimposed and connected by line segments. It is immediately clear why reproducibility is not an issue in this study: the small variation in operator measurements for each of the five parts (samples) is

TABLE 13.8 Variance Component Estimates for Innerspring Gage R&R Study

Source	Variance Component Estimate	Standard Deviation	Symbol
Total measurement	$(0.00590 + 0) = 0.00590$	0.0768	$\hat{\sigma}_{\mathrm{m}}$
Repeatability	0.00590	0.0768	$\hat{\sigma}$
Reproducibility	$(0.00079 - 0.00590)/(15) \equiv 0$	0	$\hat{\sigma}_{\mathrm{O}}$
Part (sample)	$(1.28091 - 0.00590)/(9) = 0.14167$	0.3764	$\hat{\sigma}_{\mathrm{s}}$
Total	0.14757	0.3841	$\hat{\sigma}_{\mathrm{T}}$

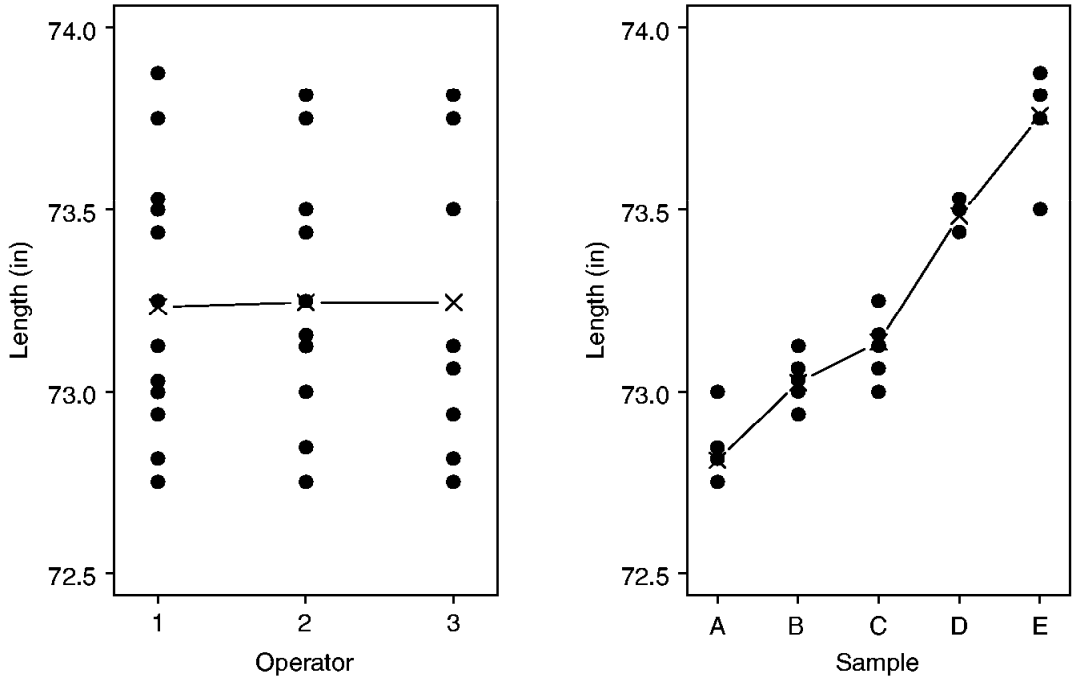


Figure 13.1 Length measurements for innerspring Gage R&R study.

evident from Figure 13.1. Hence, the measurement variation, which ordinarily consists of variability due to repeatability and reproducibility, is primarily due to repeatability in this study. Finally, the parts variation satisfactorily covers the known manufacturing range of ±0.5 inches.

Evaluating various ratios of standard deviations obtained from Gage R&R studies allows investigators to assess the adequacy of the measurement method. For example, the ratio of $\hat{\sigma}_{\mathrm{s}}/\hat{\sigma}_{\mathrm{m}}$ is used to give insight into the ability of the measurement method to discriminate among parts in spite of the process variation. A second example is the calculation of the expected range (99% of the range of the normal distribution) due to the measurement method only. This

expected range is estimated as $5.15\hat{\sigma}_{\mathrm{m}}$. Finally, the percent variation for the measurement method is calculated as $100\hat{\sigma}_{\mathrm{m}}/\hat{\sigma}_{\mathrm{T}}$. Benchmarks for the percent variation in many Gage R&R studies are 10% and 30%. If the value for a Gage R&R study is less than 10%, the measurement method is considered acceptable. For values between 10% and 30%, the measurement is usually considered adequate but a candidate for improvement. For values greater than 30%, the measurement is deemed suspect and needs to be improved. In the innerspring example, the percent study variation is 20.0%, which was considered acceptable for the measurement application.

13.4 SIGNAL-TO-NOISE RATIOS

Genichi Taguchi's approach to robust product design is described in Section 12.2.1. Analysis of all the designs proposed by Taguchi can be conducted using the methods presented in this chapter and in previous chapters. Taguchi popularized the use of signal-to-noise (SN) ratios in the analysis of robust design data. Analysis of SN ratios can add valuable insight into the effects of factors on responses. They should only be used, however, in conjunction with a comprehensive assessment of factor effects using the methods discussed previously (see the discussion at the end of this section).

In Section 12.2.1, it was noted that a quadratic loss function is typically advocated by Taguchi to represent product or process loss because it leads to a simple analysis for optimizing, in a robust sense, the product or process. The average quadratic loss function for a response y can be written as a multiple of the sum of two components:

$$L \propto \sigma_y^2 + (\overline{y} - T)^2$$

where σ_y^2 is the variance of the response, $\overline{y}$ is the response average, and T is the target that minimizes the quality loss (see Figure 12.3). Minimizing L depends on minimizing both the response dispersion σ_y^2 *and* the response location, estimated by the average, relative to T.

Taguchi uses analysis techniques that do not depend on a model that relates the response to the factor effects. Rather, he uses a calculated SN ratio and the response average to identify factors and their corresponding levels that yield minimum quality loss. A SN ratio combines into one measure the trade-offs between the magnitude of the response distribution average and the variation in the response values. Many SN ratios have been proposed, the most common of which are:

$$\mathrm{SN} = \begin{cases} -10\log_{10}(\overline{y}^2/s^2) & 0 < T < \infty \\ -10\log_{10}(\Sigma y_i^2/n) & T = 0 \\ -10\log_{10}(\Sigma(1/y_i)^2/n) & T = \infty \end{cases}$$

The analysis of SN ratios is a two-step process. First, factors (and factor–level combinations) are identified which maximize the appropriate version of SN. Second, additional factors (not substantively influencing SN) are identified that affect the average response $\overline{y}$. Factor level combinations from among this second set of factors are chosen to yield a response average that is close to the target, T. This is a "pick the winner" approach in the sense that experimental conditions are identified among the suite of experimental runs which do the best job of satisfying both objectives: large values of SN and small values of $|\overline{y} - T|$. Insight into possible factor effects can be obtained with main effect plots.

We return to the steering column example introduced in Section 12.2.1. In this example the desired target is a torque of 30. Engineers used a Resolution-IV fractional factorial experiment consisting of eight factor–level combinations as an inner array for the four control factors. An L_4 outer array was used for the two environmental factors. The data for the 32-run experiment are shown in Table 13.9, and the calculated values, $\overline{y}$ and SN, for each point of the inner array are summarized in Table 13.10. Also summarized in Table 13.10 are the averages needed to construct main effect plots. Figures 13.2 and 13.3 show the main effects plots for the average and SN, respectively. Note that Pocket Depth appears to be the only active factor for both SN and the average response. In this example, a factor does not exist that can be used to adjust the location of the response distribution but not change the SN ratio. Fortunately, the average is near the desired target for four of the inner array combinations — the four corresponding to a deep pocket depth. There appear to be at least two winners to pick from: test conditions 5 and 6.

Many authors have pointed out the inefficiencies of this analysis approach. Among the cited difficulties with analyzing SN ratios are the following.

- A SN analysis might not reliably identify active factors.
- SN and $\overline{y}$ are linked; this dependency makes trade-offs difficult to assess.
- Factors that affects $\overline{y}$ and not SN may not exist.
- ANOVA methods for the responses and the logarithms of the standard deviations for outer array factors often provide more detailed and informative assessments of active and environmental factors than an analysis of SN ratios.

The details of these issues are left to the references. An integrated approach to robust product design is recommended as an alternative to the Taguchi approach. This integrated approach is a more efficient and effective methodology, both from design and analysis perspectives. It is described in Section 12.2.2 (design) and Chapter 17 (analysis).

TABLE 13.9 Torque Data for Steering Column Example

Pocket	Yoke	Tube	Power	Clearance:	Maximum		Minimum	
Depth	Conc.	Length	Setting	Line Voltage:	Low	Best	Low	Best
Shallow	Off Center	Short	80		23.4	33.4	6.7	17.8
Shallow	Off Center	Long	87		34.7	21.7	17.9	6.3
Shallow	On Center	Short	87		33.8	22.3	18.1	7.0
Shallow	On Center	Long	80		22.6	34.1	7.1	16.9
Deep	Off Center	Short	87		26.9	17.0	42.6	30.7
Deep	Off Center	Long	80		16.5	28.6	33.5	43.1
Deep	On Center	Short	80		14.9	28.5	33.0	42.6
Deep	On Center	Long	87		28.1	14.6	43.6	32.4

TABLE 13.10 Torque Data Summaries for Steering Column Example

Test Condition	Pocket Depth	Yoke Conc.	Tube Length	Power Setting	Average	s	SN
1	Shallow	Off Center	Short	80	20.3	11.14	5.22
2	Shallow	Off Center	Long	87	20.2	11.70	4.72
3	Shallow	On Center	Short	87	20.3	11.08	5.26
4	Shallow	On Center	Long	80	20.2	11.28	5.05
5	Deep	Off Center	Short	87	29.3	10.58	8.85
6	Deep	Off Center	Long	80	30.4	11.07	8.79
7	Deep	On Center	Short	80	29.8	11.51	8.24
8	Deep	On Center	Long	87	29.7	11.99	7.87

		Averages for Main Effect Plots	
		Average	SN
Pocket Depth			
	Shallow	20.2	5.06
	Deep	29.8	8.44
Yoke Conc.			
	Off Center	25.1	6.89
	On Center	25.0	6.61
Tube Length			
	Short	24.9	6.89
	Long	25.1	6.61
Power Setting			
	80	25.2	6.83
	87	24.9	6.68

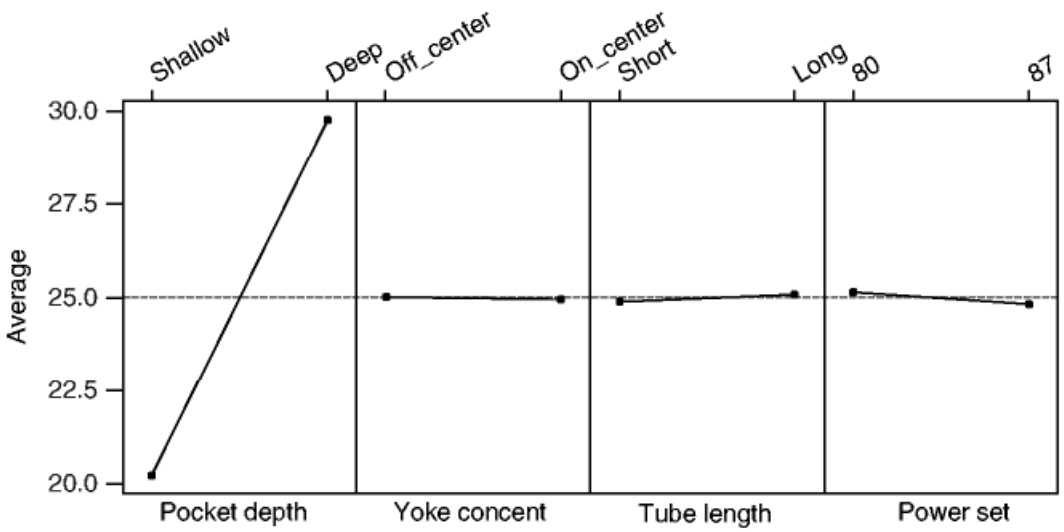


Figure 13.2 Main effects plot for torque data: averages.

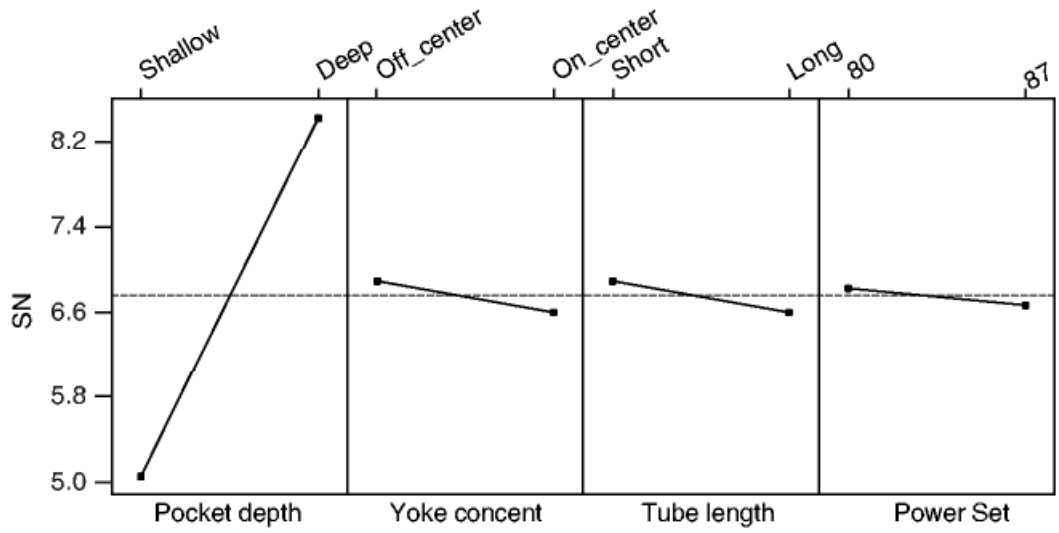


Figure 13.3 Main effects plot for torque data: SN ratios.

REFERENCES

The references listed at the end of Chapter 12 can be consulted for additional information on the topics covered in this chapter.

Data References

The data for Exercise 16 were taken from *The National Institute of Standards and Technology* web site: www.itl.nist.gov/div898/handbook/pri/section5/pri56.htm. July, 2001

EXERCISES

1 The following data are a portion of the responses collected during an interlaboratory study. Each of the several laboratories was sent a number of materials that were carefully chosen to have different measurement values on the characteristic of interest. The laboratories were required to perform three separate analyses of the test material. Devise an appropriate ANOVA model for these data, assuming no laboratory-by-material interaction. Analyze the statistical significance of the design factors by constructing an ANOVA table, including the expected mean squares.

Laboratory	Material	Repeat Measurements		
1	*A*	12.20	12.28	12.16
	B	15.51	15.02	15.29
	C	18.14	18.08	18.21
2	*A*	12.59	12.30	12.67
	B	14.98	15.46	15.22
	C	18.54	18.31	18.60
3	*A*	12.72	12.78	12.66
	B	15.33	15.19	15.24
	C	18.00	18.15	17.93

Obtain estimates of repeatability and reproducibility for this interlaboratory study.

2 Refer to the experiment in Chapter 5, Exercises 5 and 6, where the mold temperature is being measured during the formation of drinking glasses. Five temperatures are sampled from each mold in the experiment. The following data are collected on this hierarchically nested design:

Factory	Disk	Mold	Temperature				
1	1	1	459	467	465	472	464
1	1	2	464	468	461	464	469
1	2	1	465	466	469	463	465
1	2	2	466	464	463	466	466
2	1	1	472	471	472	471	470
2	1	2	472	468	468	473	470
2	2	1	471	469	473	471	470
2	2	2	470	468	472	465	472

Analyze these data, assuming that factory is a fixed factor and disk and mold are random factors.

3 Refer to the experiment designed to study the dye quality in a synthetic fiber in Chapter 11, Exercise 7. The following data were collected in a staggered nested design that involved 18 bobbins. Analyze these data and determine which of the factors significantly affect the dye rating.

Dye Quality Rating	Shift	Operator	Package	Bobbin
87	1	1	2	3
65	3	6	17	34
34	2	4	10	19
54	1	2	5	10
86	2	4	12	23
91	3	6	16	31
3	1	1	1	2
65	3	6	18	36
29	2	3	9	18
83	3	6	16	32
75	2	4	11	22
33	1	1	1	1
26	1	1	3	5
74	3	5	13	26
35	2	4	11	21
75	3	6	18	35
81	1	1	3	6
29	2	4	12	24
44	1	1	2	3
28	3	6	17	34
34	2	4	10	19
64	1	2	5	10
79	2	4	12	23
94	3	6	16	31
10	1	1	1	2
53	3	6	18	36
38	2	3	9	18
82	3	6	16	32
70	2	4	11	22
41	1	1	1	1
28	1	1	3	5
72	3	5	13	26
39	2	4	11	21
59	3	6	18	35
87	1	1	3	6
26	2	4	12	24

4 An experiment was performed to assess the effects of three catalysts on the yield of a chemical process. Four replications of the experiment were performed. In each replication, three supposedly identical process runs were begun, one process run using each catalyst. After three and five days, samples were taken and the chemical yields were determined. Analyze the data below to determine whether the catalysts or the sampling periods (three or five days) significantly affect chemical yields of the process.

	Catalyst *A*		**Catalyst *B***		**Catalyst *C***	
Replicate	**3 Days**	**5 Days**	**3 Days**	**5 Days**	**3 Days**	**5 Days**
1	83	93	76	82	67	81
2	65	77	65	72	82	78
3	67	88	63	75	65	74
4	81	89	68	74	63	75

5 At an elastomer plant the viscosity of the final product is measured according to the following test protocol: one lot of production (composed of several pallets) is sampled per production shift; one sample if taken from bag #40 on a pallet chosen at random; samples are taken from bags #1, #2, #20, and #21 on the next successive pallet; a sample is taken from bag #15 on the next pallet in the production sequence; samples from bags #1, #20, #15, and #40 are analyzed on the shift that they are sampled; and samples from bags #2 and #21 are analyzed the next day. The laboratory results for 20 lots follow. What components of variance can be estimated from these data? Estimate the variance components. What types of process improvement efforts would have the best chance of reducing variation in the product? Why?

Lot	Pallet	Test Time	Pallet Location	Viscosity
1	1	1	1	115.0
1	2	1	2	114.0
1	2	2	2	117.0
1	2	1	3	115.0
1	2	2	3	117.0
1	3	1	4	116.0
2	1	1	1	108.0
2	2	1	2	110.0
2	2	2	2	117.0
2	2	1	3	115.0
2	2	2	3	116.0

2	3	1	4	112.5
3	1	1	1	114.5
3	2	1	2	116.0
3	2	2	2	115.5
3	2	1	3	116.5
3	2	2	3	114.0
3	3	1	4	114.5
4	1	1	1	114.0
4	2	1	2	114.0
4	2	2	2	113.5
4	2	1	3	116.0
4	2	2	3	115.5
4	3	1	4	115.0
5	1	1	1	115.0
5	2	1	2	114.0
5	2	2	2	114.5
5	2	1	3	115.0
5	2	2	3	114.0
5	3	1	4	114.0
6	1	1	1	117.0
6	2	1	2	117.0
6	2	2	2	114.0
6	2	1	3	116.0
6	2	2	3	114.0
6	3	1	4	117.0
7	1	1	1	111.5
7	2	1	2	111.0
7	2	2	2	116.5
7	2	1	3	110.5
7	2	2	3	116.0
7	3	1	4	111.0
8	1	1	1	114.0
8	2	1	2	114.5
8	2	2	2	115.5
8	2	1	3	114.0
8	2	2	3	116.0
8	3	1	4	116.5
9	1	1	1	116.0
9	2	1	2	114.0
9	2	2	2	115.5
9	2	1	3	113.5
9	2	2	3	117.0
9	3	1	4	117.0

Lot	Pallet	Test Time	Pallet Location	Viscosity
10	1	1	1	114.0
10	2	1	2	116.0
10	2	2	2	116.0
10	2	1	3	115.5
10	2	2	3	116.0
10	3	1	4	114.0
11	1	1	1	114.0
11	2	1	2	110.5
11	2	2	2	116.0
11	2	1	3	112.5
11	2	2	3	118.0
11	3	1	4	113.5
12	1	1	1	113.5
12	2	1	2	117.0
12	2	2	2	120.0
12	2	1	3	118.0
12	2	2	3	117.5
12	3	1	4	118.0
13	1	1	1	112.0
13	2	1	2	113.0
13	2	2	2	116.0
13	2	1	3	115.0
13	2	2	3	116.0
13	3	1	4	115.0
14	1	1	1	113.0
14	2	1	2	113.5
14	2	2	2	111.0
14	2	1	3	112.0
14	2	2	3	111.0
14	3	1	4	111.5
15	1	1	1	111.0
15	2	1	2	110.0
15	2	2	2	111.5
15	2	1	3	112.0
15	2	2	3	110.5
15	3	1	4	110.0

16	1	1	1	115.0
16	2	1	2	115.0
16	2	2	2	117.0
16	2	1	3	113.0
16	2	2	3	116.0
16	3	1	4	116.5
17	1	1	1	115.5
17	2	1	2	115.5
17	2	2	2	114.5
17	2	1	3	115.5
17	2	2	3	115.0
17	3	1	4	117.5
18	1	1	1	121.5
18	2	1	2	120.5
18	2	2	2	124.0
18	2	1	3	121.0
18	2	2	3	124.0
18	3	1	4	123.5
19	1	1	1	119.0
19	2	1	2	120.0
19	2	2	2	118.0
19	2	1	3	118.0
19	2	2	3	119.0
19	3	1	4	119.0
20	1	1	1	114.0
20	2	1	2	113.0
20	2	2	2	113.0
20	2	1	3	113.0
20	2	2	3	116.0
20	3	1	4	111.5

6 Questions have been raised about the effects of pollution filters on the noise level performance of automobiles. The data shown below are for three different sizes of automobiles (1 = small, 2 = medium, and 3 = large) with two different filter types (1 = standard filter and 2 = new design). The filter designs are unique depending on the size of the automobile. Noise measurements were taken on both sides of the vehicle (1 = right side and 2 = left side). Which factors are nested and which are crossed? Is there evidence in the data that the new design reduces the noise level?

Noise	Size	Type	Side
810	1	1	1
820	1	1	1
820	1	1	1
840	2	1	1
840	2	1	1
845	2	1	1
785	3	1	1
790	3	1	1
785	3	1	1
835	1	1	2
835	1	1	2
835	1	1	2
845	2	1	2
855	2	1	2
850	2	1	2
760	3	1	2
760	3	1	2
770	3	1	2
820	1	2	1
820	1	2	1
820	1	2	1
820	2	2	1
820	2	2	1
825	2	2	1
775	3	2	1
775	3	2	1
775	3	2	1
825	1	2	2
825	1	2	2
825	1	2	2
815	2	2	2
825	2	2	2
825	2	2	2
770	3	2	2
760	3	2	2
765	3	2	2

7 Fifteen rail cars of a chemical intermediate for nylon were sampled with two samples being taken from distinct locations in the car. Each sample from a location in a rail car was split and an aliquot was analyzed for an important impurity related to the quality of the nylon (measured in parts per million). The lower the amount of the impurity, the better the quality of the nylon. The resulting data are shown in the table below. What components of variance can be estimated from these data? Estimate the variance components. What types of process improvement efforts would have the best chance of reducing variation in the product? Why? Is reduced variation desirable? Why?

Rail Car	Location 1 Test Time 1	Location 1 Test Time 2	Location 2 Test Time 1	Location 2 Test Time 2
1	17	28	22	19
2	23	24	22	28
3	24	22	23	23
4	25	22	23	13
5	19	27	28	27
6	24	27	31	21
7	35	32	10	34
8	33	39	36	46
9	36	36	38	34
10	30	34	27	45
11	35	35	38	39
12	35	40	38	41
13	22	35	38	47
14	29	40	41	47
15	40	44	45	39

8 An engineer wants to understand the variation associated with the manufacture of aluminum die-castings. The molten aluminum is held in bull furnaces at the casting machine for which the desired temperature setting is 1275°F. There are 15 total plants located in North America where castings are made, and four plants are randomly sampled to be included in the study. Furnace temperatures are measured three times during a production shift for each of two operators in each of three different shifts. Estimate the variance components for the data in the table that follows. What effect do the operators have on furnace temperature? What opportunities do the data indicate for improving the consistency of the aluminum temperature in the bull furnaces?

Temperature	Plant	Operator	Shift
1277	1	1	1
1272	1	1	1
1281	1	1	1
1278	1	1	2
1275	1	1	2
1274	1	1	2
1272	1	1	3
1275	1	1	3
1268	1	1	3
1271	1	2	1
1274	1	2	1
1270	1	2	1
1279	1	2	2
1282	1	2	2
1277	1	2	2
1270	1	2	3
1277	1	2	3
1283	1	2	3
1284	2	1	1
1277	2	1	1
1281	2	1	1
1277	2	1	2
1282	2	1	2
1281	2	1	2
1279	2	1	3
1277	2	1	3
1282	2	1	3
1272	2	2	1
1275	2	2	1
1275	2	2	1
1272	2	2	2
1275	2	2	2
1270	2	2	2
1272	2	2	3
1277	2	2	3
1275	2	2	3
1275	3	1	1
1270	3	1	1
1269	3	1	1
1277	3	1	2
1271	3	1	2

1274	3	1	2
1269	3	1	3
1273	3	1	3
1268	3	1	3
1270	3	2	1
1266	3	2	1
1268	3	2	1
1271	3	2	2
1273	3	2	2
1276	3	2	2
1278	3	2	3
1280	3	2	3
1274	3	2	3
1284	4	1	1
1277	4	1	1
1280	4	1	1
1276	4	1	2
1275	4	1	2
1274	4	1	2
1275	4	1	3
1270	4	1	3
1269	4	1	3
1269	4	2	1
1275	4	2	1
1279	4	2	1
1282	4	2	2
1283	4	2	2
1279	4	2	2
1277	4	2	3
1279	4	2	3
1275	4	2	3

9 An experiment is run to examine how three processing factors might be related to the electroplating of aluminum on copper strips. The three factors of interest are current, solution temperature, and the solution concentration of the plating agent. There are 16 copper strips available for the experiment. Four copper strips are placed in the solution at one time. A specified level of current is applied to an individual strip within the solution; two of the four strips are randomly assigned to the low current level and the remaining strips are assigned to the high level. The plating is performed and the plating rate is measured for each strip. This procedure is carried out for all four temperature/concentration factor–level combinations. The data are listed below with coded factor levels. Which factors are whole

plot factors, and which are subplot factors? Analyze the data and determine which factors affect plating rate and by how much.

Current	Temperature	Concentration	Rate
−1	−1	−1	56.2
−1	−1	1	65.7
−1	1	−1	74.2
−1	1	1	73.0
1	−1	−1	61.3
1	−1	1	75.0
1	1	−1	77.0
1	1	1	71.8
−1	−1	−1	57.1
−1	−1	1	70.2
−1	1	−1	82.7
−1	1	1	83.8
1	−1	−1	70.7
1	−1	1	76.0
1	1	−1	73.2
1	1	1	68.7

10 Suppose that the replicated 2^3 factorial experiment in Exercise 9 had been run under the following protocol: There is only one copper strip in the solution at one time; however, two strips (one at each level of the current factor) are processed one right after the other under the same temperature and concentration setting. Once the two strips have been processed, the concentration is changed and the temperature is reset to another combination. The next two strips are then processed one after the other under this temperature and concentration setting. This process is continued until all 16 copper strips have been tested. Which factors are whole plot factors, and which are subplot factors? Analyze the data and determine which factors affect plating rate and by how much. How does your analysis change compared to Exercise 9? Did you conclusions change?

11 A semiconductor chip must be implanted and annealed during manufacture. An experiment was conducted to examine three factors (A, B, and C) in the implant step and three (D, E, and F) in the anneal step. A group of eight chips go through the implant step first. A specific implant treatment combination is applied to all eight chips at once. Once the first implant treatment is finished, another set of eight chips gets implanted with a different treatment combination. This process continues until all $2^3 = 8$ treatment combinations for factors A, B, and C have been run. The anneal

step gets underway with the first anneal treatment combination applied to a set of eight chips where each of the eight chips comes from one of each of the eight implant combinations. After this group of chips have been annealed, the second anneal treatment combination is applied to a second group of eight chips (again, where each of the eight chips comes from one of each of the eight implant combinations). This is continued until all $2^3 = 8$ treatment combinations for factors D, E, and F have been run. Each chip is then assigned a quality rating. The data from this experiment are shown in the following table. Analyze the data and determine which factors affect the quality rating and by how much. Assume that three-factor and higher interactions are negligible.

A	B	C	D	E	F	Rating
−1	−1	−1	−1	−1	−1	84
1	−1	−1	−1	−1	−1	96
−1	1	−1	−1	−1	−1	90
1	1	−1	−1	−1	−1	99
−1	−1	1	−1	−1	−1	85
1	−1	1	−1	−1	−1	83
−1	1	1	−1	−1	−1	93
1	1	1	−1	−1	−1	94
−1	−1	−1	1	−1	−1	88
1	−1	−1	1	−1	−1	92
−1	1	−1	1	−1	−1	88
1	1	−1	1	−1	−1	88
−1	−1	1	1	−1	−1	84
1	−1	1	1	−1	−1	88
−1	1	1	1	−1	−1	91
1	1	1	1	−1	−1	97
−1	−1	−1	−1	1	−1	91
1	−1	−1	−1	1	−1	91
−1	1	−1	−1	1	−1	85
1	1	−1	−1	1	−1	93
−1	−1	1	−1	1	−1	95
1	−1	1	−1	1	−1	89
−1	1	1	−1	1	−1	93
1	1	1	−1	1	−1	91
−1	−1	−1	1	1	−1	91
1	−1	−1	1	1	−1	83
−1	1	−1	1	1	−1	93

A	B	C	D	E	F	Rating
1	1	−1	1	1	−1	85
−1	−1	1	1	1	−1	77
1	−1	1	1	1	−1	95
−1	1	1	1	1	−1	82
1	1	1	1	1	−1	89
−1	−1	−1	−1	−1	1	83
1	−1	−1	−1	−1	1	95
−1	1	−1	−1	−1	1	96
1	1	−1	−1	−1	1	81
−1	−1	1	−1	−1	1	95
1	−1	1	−1	−1	1	99
−1	1	1	−1	−1	1	94
1	1	1	−1	−1	1	92
−1	−1	−1	1	−1	1	88
1	−1	−1	1	−1	1	90
−1	1	−1	1	−1	1	92
1	1	−1	1	−1	1	91
−1	−1	1	1	−1	1	80
1	−1	1	1	−1	1	89
−1	1	1	1	−1	1	99
1	1	1	1	−1	1	86
−1	−1	−1	−1	1	1	88
1	−1	−1	−1	1	1	88
−1	1	−1	−1	1	1	99
1	1	−1	−1	1	1	93
−1	−1	1	−1	1	1	86
1	−1	1	−1	1	1	87
−1	1	1	−1	1	1	93
1	1	1	−1	1	1	89
−1	−1	−1	1	1	1	88
1	−1	−1	1	1	1	91
−1	1	−1	1	1	1	93
1	1	−1	1	1	1	95
−1	−1	1	1	1	1	85
1	−1	1	1	1	1	85
−1	1	1	1	1	1	89
1	1	1	1	1	1	84

12 The following data were obtained for the scenario described in Exercise 2 of Chapter 12. Is the gage adequate for this process? Why or why not?

	Operator		
Part	1	2	3
1	31.200	31.198	31.206
	31.203	31.194	31.206
	31.203	31.198	31.205
2	31.189	31.192	31.192
	31.190	31.190	31.194
	31.189	31.190	31.192
3	31.195	31.194	31.202
	31.197	31.194	31.200
	31.195	31.196	31.201
4	31.196	31.190	31.194
	31.194	31.188	31.195
	31.196	31.190	31.195
5	31.200	31.198	31.198
	31.198	31.196	31.200
	31.200	31.198	31.200
6	31.190	31.204	31.190
	31.192	31.202	31.191
	31.190	31.204	31.190
7	31.197	31.194	31.198
	31.198	31.192	31.198
	31.195	31.194	31.198
8	31.202	31.202	31.208
	31.202	31.200	31.207
	31.203	31.200	31.208
9	31.190	31.198	31.192
	31.194	31.196	31.192
	31.193	31.198	31.190
10	31.200	31.206	31.206
	31.204	31.204	31.205
	31.205	31.204	31.204

13 The following data were obtained for the scenario described in Exercise 1 of Chapter 12. The side-lean response is reported in one-eight-inch multiples; the sign of the response denotes the direction of the lean — negative is to the left and positive to the right. Is the gage adequate for this process? Why or why not?

	Sample				
Person	A	B	C	D	E
Ken	1	−1	1	2	1
	1	−1	1	−2	1
	−1	−1	−1	−1	−1
	−1	−1	−1	−1	1
Mick	−2	−2	2	2	1
	−2	−2	2	2	1
	2	−2	−1	−2	2
	2	2	−2	2	−2
Tammy	−1	−1	−1	−2	1
	−1	−1	−1	−2	−1
	−1	−1	−2	−2	−1
	−1	−1	−2	−2	1

14 For the situation described in Exercise 11 of Chapter 12, the team of engineers and operators decided to investigate the three environmental factors, Cooling Tank Water Temperature, Amount of Regrind Material, and Ambient Temperature at the Machine, at two different process conditions (two particular combinations of Air Pressure, Screw Speed, and Die Temperature — it was thought that both of these process conditions yield an average near the desired target of 31). The following data were obtained:

Process Condition	Cooling Tank Temperature	Regrind Material	Ambient Temperature	Shrinkage
1	−1	−1	−1	29.8
1	1	−1	−1	29.1
1	−1	1	−1	32.8
1	1	1	−1	31.4
1	−1	−1	1	29.4
1	1	−1	1	31.4
1	−1	1	1	30.2
1	1	1	1	32.0

2	−1	−1	−1	27.3
2	1	−1	−1	29.4
2	−1	1	−1	35.2
2	1	1	−1	30.1
2	−1	−1	1	31.2
2	1	−1	1	28.4
2	−1	1	1	32.5
2	1	1	1	34.9

Identify the designs used for the inner array and the outer array. Analyze the data using (a) a SN ratio when a target value is best, (b) a SN ratio when smaller-is-better, and (c) the average and standard deviation at each process condition. Which analysis is appropriate for these data? Is one of the process conditions preferred? Why or why not?

15 For the scenario described in Exercise 13 of Chapter 12 the following data were obtained:

Sight Glass	Orifice Type	Torque*	Location*	Response
Type A	22	120	Vertical	16
Type A	22	120	Off Vertical	17
Type A	22	150	Vertical	12
Type A	22	150	Off Vertical	13
Type A	25	120	Vertical	11
Type A	25	120	Off Vertical	19
Type A	25	150	Vertical	12
Type A	25	150	Off Vertical	15
Type B	22	120	Vertical	11
Type B	22	120	Off Vertical	16
Type B	22	150	Vertical	11
Type B	22	150	Off Vertical	11
Type B	25	120	Vertical	12
Type B	25	120	Off Vertical	11
Type B	25	150	Vertical	13
Type B	25	150	Off Vertical	15

Identify the designs used for the inner array and the outer array (the two factors with the "*" are environmental factors). Analyze the data using the appropriate SN ratio. What is the best process condition to minimize the response and minimize variation due to environmental factors? Describe the results you found to someone who is not familiar with the study, and use graphical displays as necessary.

16 The starter motor of an automobile has to perform reliably in many different ambient temperatures and varying states of battery strength. The design engineer can control the number of armature turns, gage of armature wire, and the ferric content of the magnet alloy. A laboratory study resulted in the following data:

Turns	Gauge	Ferric Content	Ambient Temperature	Battery Voltage	% Maximum Torque
−1	−1	−1	−1	−1	75
−1	−1	−1	1	−1	86
−1	−1	−1	−1	1	67
−1	−1	−1	1	1	98
1	−1	−1	−1	−1	87
1	−1	−1	1	−1	78
1	−1	−1	−1	1	56
1	−1	−1	1	1	91
−1	1	−1	−1	−1	77
−1	1	−1	1	−1	89
−1	1	−1	−1	1	78
−1	1	−1	1	1	8
1	1	−1	−1	−1	95
1	1	−1	1	−1	65
1	1	−1	−1	1	77
1	1	−1	1	1	95
−1	−1	1	−1	−1	78
−1	−1	1	1	−1	78
−1	−1	1	−1	1	59
−1	−1	1	1	1	94
1	−1	1	−1	−1	56
1	−1	1	1	−1	79
1	−1	1	−1	1	67
1	−1	1	1	1	94
−1	1	1	−1	−1	79
−1	1	1	1	−1	80
−1	1	1	−1	1	66
−1	1	1	1	1	85
1	1	1	−1	−1	71
1	1	1	1	−1	80
1	1	1	−1	1	73
1	1	1	1	1	95

A high percent maximum torque is desirable. Identify the designs used for the inner array and the outer array. Analyze the data by (a) using the appropriate SN ratio, and (b) calculating the average and standard deviation at each process condition. What is the best process condition to maximize the percent maximum torque and to minimize variation due to environmental factors? Can you "pick the winner" in this experiment or are trade-offs needed? Describe the results you found to someone who is not familiar with the study, and use graphical displays as necessary.

17 An experiment was conducted to understand how three processing variables (A, B, and C) affect the strength of high-carbon steel wire. The following data were obtained:

A	B	C	Strength
−1	−1	−1	36
−1	−1	−1	34
−1	−1	−1	38
−1	−1	−1	38
1	−1	−1	24
1	−1	−1	26
1	−1	−1	27
1	−1	−1	23
−1	1	−1	61
−1	1	−1	62
−1	1	−1	64
−1	1	−1	61
−1	−1	1	43
−1	−1	1	37
−1	−1	1	40
−1	−1	1	42
1	1	−1	46
1	1	−1	44
1	1	−1	47
1	1	−1	45
1	−1	1	27
1	−1	1	29
1	−1	1	30
1	−1	1	30
−1	1	1	58
−1	1	1	60
−1	1	1	62

A	B	C	Strength
−1	1	1	64
1	1	1	51
1	1	1	46
1	1	1	48
1	1	1	52

Analyze the data using the appropriate SN ratio. What is the best process condition to maximize strength and minimize variation? Describe the results you found to someone who is not familiar with the study, and use graphical displays as necessary.

18 In Exercise 17, suppose that steel wire can be either too strong or too weak. A nominal value of 56 gives the best performance in the field. How does this change your analysis? What process conditions would you recommend?

PART IV

Design and Analysis with Quantitative Predictors and Factors

CHAPTER 14

Linear Regression with One Predictor Variable

Linear regression is used to model one quantitative variable as a function of one or more other variables. In this chapter we introduce regression modeling with the fitting of a response variable as a linear function of one predictor variable. As with the introductory chapters to the previous parts of this book, general principles are introduced, principles which are elaborated in later chapters. The topics covered in this chapter include:

- *uses and misuses of regression modeling,*
- *a strategy for regression modeling,*
- *scatterplot smoothing,*
- *least-squares parameter estimation, and*
- *procedures for drawing inferences.*

Regression modeling is one of the most widely used statistical modeling techniques for fitting a quantitative response variable y as a function of one or more predictor variables $x_1, x_2, \ldots, x_p$. Regression models can be used to fit data obtained from a statistically designed experiment in which the predictor variables either are quantitative or are indicators of factor levels. All the ANOVA models described in earlier chapters of this book are special types of regression models. In ANOVA models the predictor variables are specially coded *indicator variables*, for example, the effects representation in Tables 5.5 to 5.7, in which the upper level of a factor is indicated by a + 1 and the lower level by a − 1. Analysis of covariance (ANACOVA) models (Chapter 16) are also special types of regression models in which some of the variables are indicator variables and others are quantitative.

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

One of the reasons for the widespread popularity of regression models is their use with nonexperimental data, such as observational or historical data. Another reason is that the regression analysis procedures contain diagnostic techniques for (a) identifying incorrect specifications of the model (Chapter 18), (b) assessing the influence of outliers on the fit (Chapter 18), and (c) evaluating whether redundancies (collinearities) among the predictor variables are adversely affecting the fit (Chapter 19). Perhaps most pragmatically, regression models are widely used because they often provide excellent fits to a response variable when the true functional relationship, if any, between the response and the predictors is unknown.

In this chapter we introduce regression modeling for the special case in which the response is to be modeled as a function of a single predictor variable. In the next section we introduce the particular form of regression model that we shall stress in this part of the text: the linear regression model.

14.1 USES AND MISUSES OF REGRESSION

Linear regression models are, apart from the random error component, linear in the unknown parameters (see Exhibit 14.1). The coefficients in the model (14.1) appear as either additive constants (β_0) or as multipliers on the predictor variable (β_1). This requirement extends to multiple linear regression models (see Chapter 15). Note too that the predictor variable can be a function, linear or nonlinear, of other predictor variables, for example, $x = \ln z$ or $x = \sin z$ for some variable z. The predictor variable cannot, however, be a function of unknown parameters, for example, $x = \ln(z + \phi)$ or $x = \sin(z - \phi)$ for some unknown ϕ. Models which are nonlinear functions of unknown model parameters are referred to as nonlinear models.

EXHIBIT 14.1

Linear Regression Models. Linear regression models that relate a response variable (y) to one predictor variable (x) are defined as

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, \ldots, n, \tag{14.1}$$

where e_i represents a random error component of the model.

As with all the models discussed in this text, several assumptions usually accompany the model. These assumptions are listed in Table 14.1. It is important to note that violation of one or more of these assumptions can invalidate the inferences drawn on parameter estimates and can result in other, often not apparent, difficulties with fitting linear regression models using least-squares

TABLE 14.1 Assumptions for Linear Regression Models

1. Over the range of applicability, the true relationship between the response and the predictor variable(s) is well approximated by a linear regression model.
2. The predictor variable(s) is (are) nonstochastic and measured without error.
3. The model errors are statistically independent and satisfactorily represented by a normal distribution with mean zero and constant, usually unknown, standard deviation σ.

TABLE 14.2 Common Uses and Misuses of Regression Modeling

Common Uses

- Prediction (forecasting)
- Interpolation
- Data fitting
- Testing for significant factor or predictor variable effects on a response
- Determination of predictor values that maximize or minimize a response
- Control of the response variable by selection of appropriate predictor values

Common Misuses

- Extrapolation
- Misinterpreting spurious relationships
- Overreliance on automated-regression results
- Exaggerated claims about the validity of empirical models

estimators (Section 14.4). Techniques that can be used to assess the reasonableness of these model assumptions are discussed in Chapter 18.

The regression texts referenced at the end of this chapter contain much discussion on the appropriate uses and abuses of regression modeling. Table 14.2 lists a few of the more common of each. The uses of regression modeling will be amply illustrated throughout the next several chapters of this text. In the remainder of this section, we wish to comment on some of the more prevalent misuses listed in the table.

Figure 14.1 is a sequence plot for rotational velocities of a bearing as a function of time. Initially, the bearing was spun at a rate of 44.5 revs/sec. The data plotted in Figure 14.1 are the velocities recorded at 10-second intervals following the termination of the spinning of the bearing. The experiment was replicated six times; each plotted point is therefore the average of six velocities. Note that the data appear to follow a straight line; hence, a linear regression model of the form (14.1) should provide an excellent fit to the averages.

An appropriate use of a linear fit to the data in Figure 14.1 would be to describe the rotational velocities (the response) as a function of time (the predictor). The slope of the fit would be a useful estimate of the deceleration

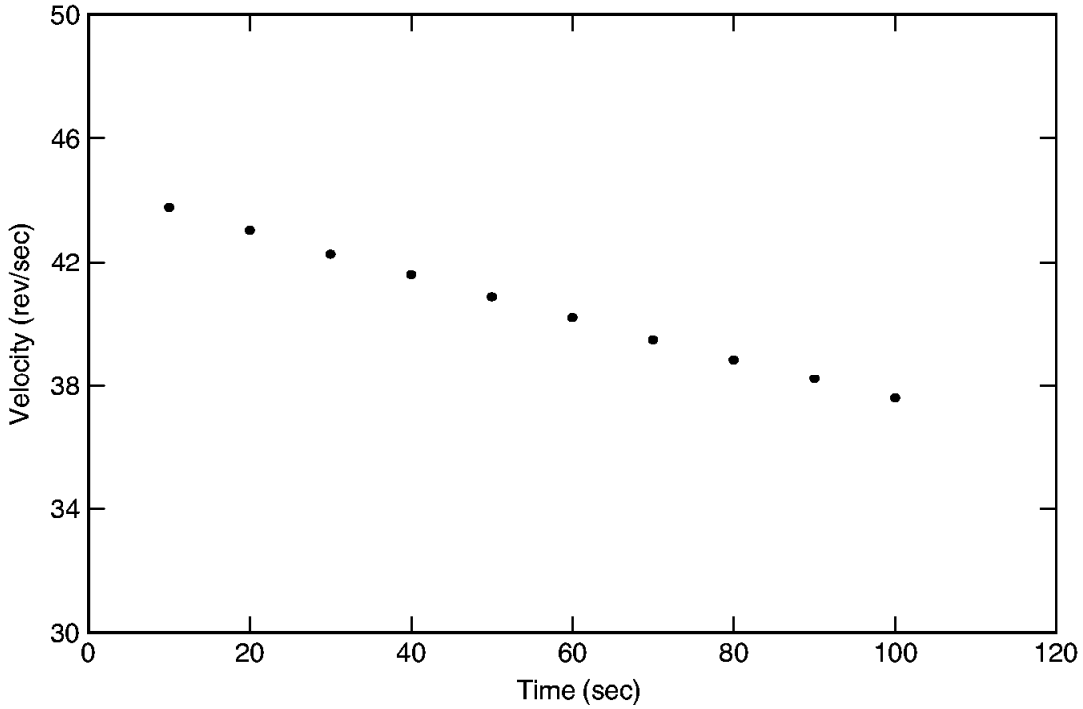


Figure 14.1 Average rotational velocities.

rate of the bearing. One could also use the fit to interpolate rotational velocities for any times between 10 and 100 seconds.

An inappropriate use of the fit would be to extrapolate the rotational velocities at, say, 120 or 150 seconds. Based solely on the data plotted in Figure 14.1, one cannot conclude that the rotational velocities remain linear outside the time frame plotted. One cannot tell from the data in the figure whether the rotational velocities continue to decelerate at a constant rate and then abruptly stop or the deceleration rate itself begins to slow down so that the velocities decrease in a nonlinear fashion. Thus, extrapolation of velocities is highly speculative because of the unknown nature of the deceleration rate beyond 100 seconds. Another extrapolation would be to conclude that the relationship depicted in Figure 14.1 holds for bearing types that are different from the one tested.

Another potential difficulty with the application of regression analysis to the data in Figure 14.1 concerns the possible violation of one of the model assumptions. Because the averages are calculated from individual velocities that are measured 10 seconds apart, it is likely that the measurement errors for the velocities on a single replicate of the experiment are correlated. If so, the averages plotted in Figure 14.1 are also correlated. Before either statistical interval estimates for model parameters or predictions are calculated or tests of hypotheses are conducted, the model assumptions must be investigated. If the correlations are found to be sufficiently large, time-series

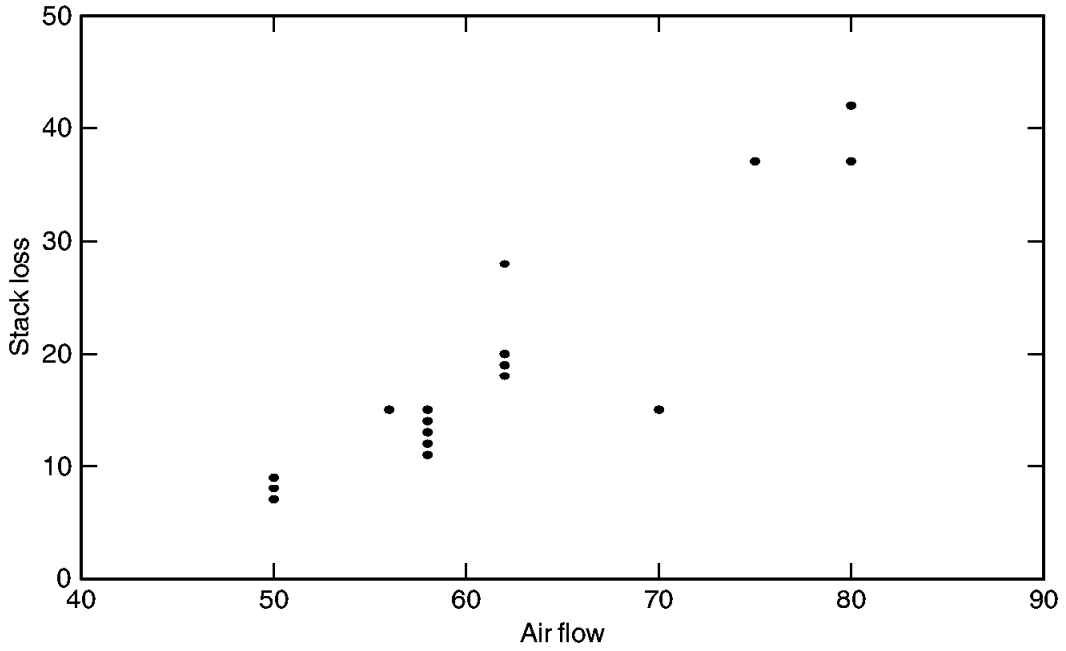


Figure 14.2 Stack loss versus air flow. Data from Brownlee, K.A. (1965). *Statistical Theory and Methodology in Science and Engineering*. New York: John Wiley & Sons, Inc. Copyright 1965 John Wiley and Sons, Inc. Used by permission.

models or other alternative analyses should be conducted, rather than a regression analysis.

Another potential misuse of regression modeling is illustrated with the scatterplot in Figure 14.2. These data represent measurements taken on 21 operating days of a chemical plant. The "stack loss" is ten times the percentage of ammonia that is left unconverted in an oxidation process for converting ammonia to nitric acid. The predictor variable in this plot is the rate of air flow in the process. It is apparent that the stack loss is increasing with the air flow. Having just discussed the dangers of extrapolation, it should be clear that a straight-line fit to these data should not be used to infer the effects of plant operating conditions outside the range of air-flow values plotted in the figure.

Often studies are conducted with nonexperimental data such as these with the intent to determine reasons for the trends. When analyzing experimental or nonexperimental data in which some of the variables are not controlled, extreme care must be taken to ensure that proper inferences are drawn when statistically significant results are obtained. In particular, one must be concerned about confounded predictor variables and spurious relationships.

Confounding was introduced in Section 7.1 in the context of designing fractional factorial experiments. In general, a factor is said to be confounded whenever its effect on the response cannot be uniquely ascribed to the factor. In the present context, a predictor variable is said to be confounded whenever its relationship with the response variable is unclear because of the predictor

variable's relationship with one or more other predictors. Failure to recognize the potential for confounding of predictor variables is one of the ways in which exaggerated claims can be made about a fitted model.

Researchers sometimes conclude from trends such as that in Figure 14.2 or from regression fits to the data that increases in air flow rates are a major contributor to increases in the stack loss. While such a relationship may be true, the conclusion is not warranted solely on the basis of a regression fit to the data in the figure.

Figure 14.3 is a plot of the stack-loss values versus the inlet water temperature. Note that this plot suggests that increases in inlet water temperature rates are associated with increases in the stack loss. Without a controlled experiment that includes both air flow and inlet water temperature, as well as any other predictor variables that are believed to affect stack loss, no definitive conclusion can be drawn about which of the variables truly affects the stack loss.

With controlled experimentation, many of the abuses listed in Table 14.2 are less likely to occur than with observational or historical data. Because variables are controlled and assigned predetermined values in designed experiments, confounding of variables is usually lessened, sometimes completely absent.

Regression analysis is appropriate when the relationship between two or more variables is clearly in one direction — that is, the design factors or the

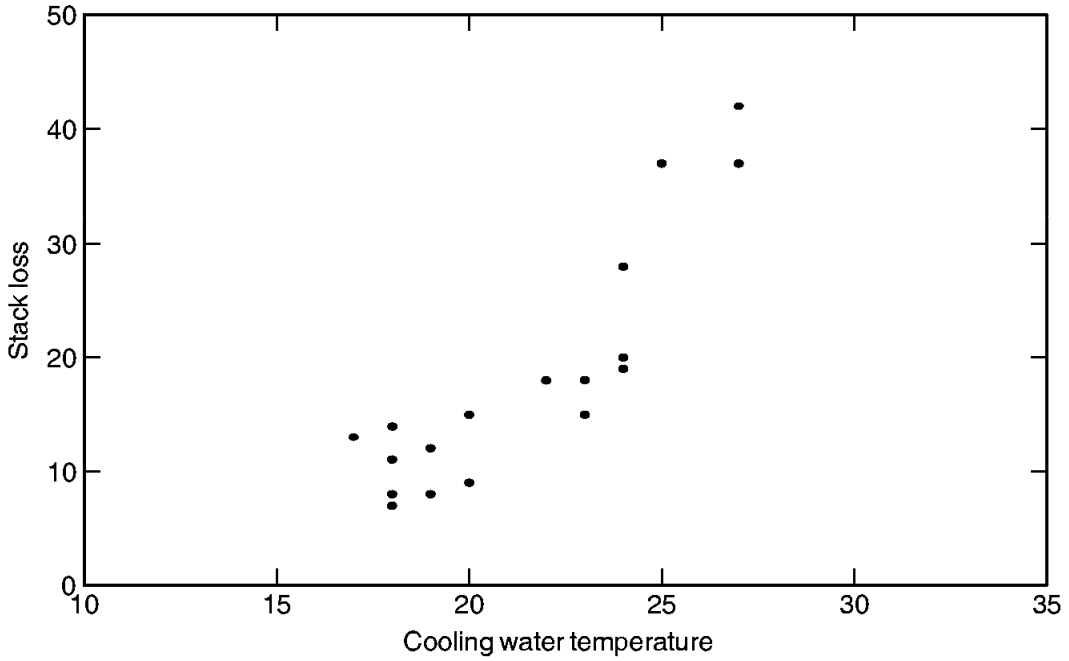


Figure 14.3 Stack loss versus cooling water inlet temperature. Data from Brownlee, K.A. (1965). *Statistical Theory and Methodology in Science and Engineering*. New York: John Wiley & Sons, Inc. Copyright 1965 John Wiley and Sons, Inc. Used by permission.

predictor variables influence but are not influenced by the response variable. Situations in which there are two or more variables that are related, none of which can be said to cause or determine the values of the others, are analyzed using multivariate statistical methods. Multivariate ANOVA models and multivariate regression models are appropriate when the design factors or predictor variables simultaneously influence the values of two or more response variables.

A correlation analysis is used to assess linear relationships between two variables when both are random and neither variable is believed to cause or determine values of the other. An example of the type of data for which a correlation analysis is appropriate is the tire-wear data plotted in Figure 4.5 and listed in Table 14.3. The linear scatter observable in Figure 4.5 is not due to one of the methods of measuring wear influencing the other method. Rather, both measurements are affected by differences in tires and the road conditions under which each tire is tested.

TABLE 14.3 Estimates of Tread Life Using Two Measurement Methods

	Tread Life (10^2mi)	
Tire No.	Weight-Loss Method	Groove-Depth Method
1	459	357
2	419	392
3	375	311
4	334	281
5	310	240
6	305	287
7	309	259
8	319	233
9	304	231
10	273	237
11	204	209
12	245	161
13	209	199
14	189	152
15	137	115
16	114	112

Sample Correlation Coefficient Calculation

x = weight-loss measurement
y = groove-depth measurement
$s_{xx} = 9073.59$ $s_{yy} = 6306.93$
$s_{xy} = 7170.07$ $r = 0.948$

A measure of the strength of a linear association between two random variables is the sample *correlation coefficient* — more precisely, Pearson's product-moment correlation coefficient, or Pearson's r. It is defined in Exhibit 14.2.

EXHIBIT 14.2 CALCULATION OF PEARSON'S r (THE SAMPLE CORRELATION COEFFICIENT)

1. Calculate the two sample means, $\bar{x}$ and $\bar{y}$; e.g.,

$$\bar{x} = n^{-1} \sum x_i .$$

2. Calculate the two sample variances, s_{xx} and s_{yy}; e.g.,

$$s_{xx} = (n-1)^{-1} \sum (x_i - \bar{x})^2$$
$$= (n-1)^{-1} \left(\sum x_i^2 - n\bar{x}^2 \right).$$

3. Calculate the sample covariance, s_{xy}:

$$s_{xy} = (n-1)^{-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$
$$= (n-1)^{-1} \left(\sum x_i y_i - n\bar{x}\,\bar{y} \right).$$

4. Calculate the sample correlation coefficient:

$$r = s_{xy} / (s_{xx} s_{yy})^{1/2}. \tag{14.2}$$

The value of Pearson's r for the tire-wear data can be calculated using Equation (14.2) and the summary statistics given in Table 14.3. Doing so, one obtains $r = 0.948$. This value of Pearson's r can be interpreted with the aid of the properties listed in Exhibit 14.3.

EXHIBIT 14.3 PROPERTIES OF PEARSON'S r

- Pearson's r measures the strength of a linear association between two quantitative variables.
- The closer r is to zero, the weaker is the linear association.
- The closer r is to -1 or $+1$, the stronger is the linear association, with the sign of r indicating a decreasing ($r < 0$) or an increasing ($r > 0$) linear association.
- The sample correlation coefficient can only equal $+1$ (-1) when the observations lie exactly on a straight line having positive (negative) slope.

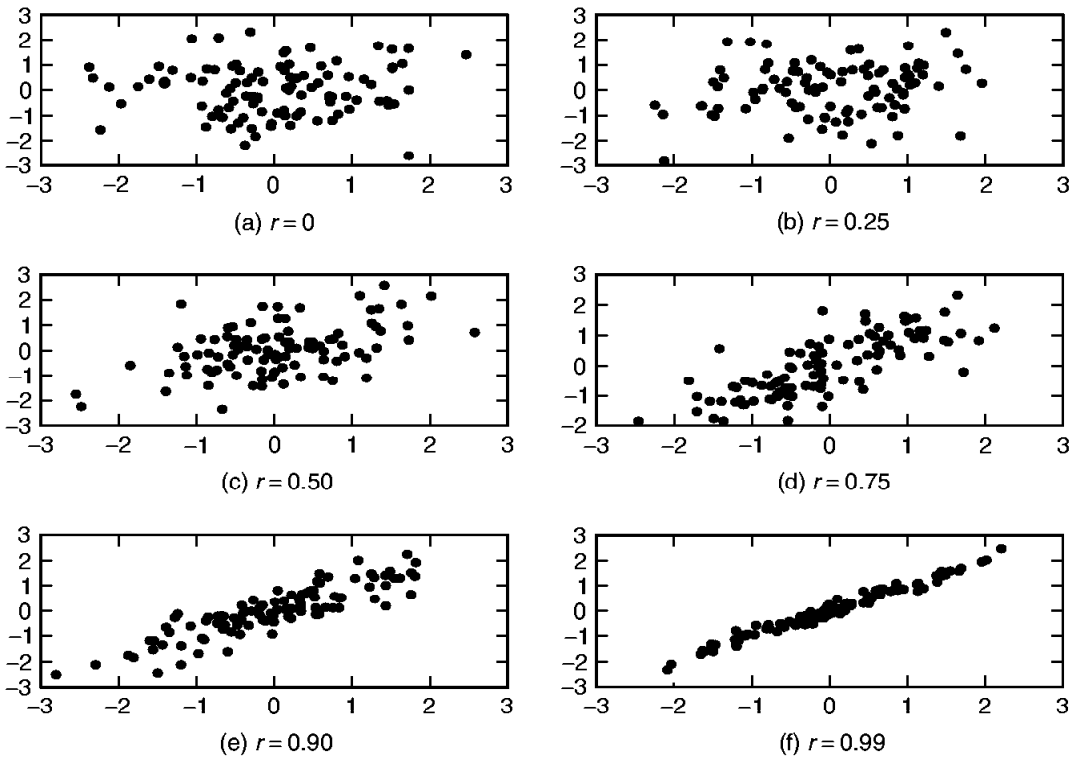


Figure 14.4 Scatterplots of data having various sample correlations.

Figure 14.4 is a series of scatterplots of pairs of variates that have different positive values of the sample correlation coefficient. Variates having negative correlations would have similar scatterplots with downward rather than upward trends. Thus, the size of the sample correlation coefficient for the tire-wear data indicates that the linear relationship between the two measures of tire wear is extremely strong.

The statistical significance of the sample correlation coefficient can be assessed using a t-statistic. Let ρ denote the correlation coefficient for a population of pairs of quantitative observations. Alternatively, ρ could denote the correlation coefficient for a conceptual population of pairs of observations arising from measurements on some process. The following t statistic can be used to test H_0: $\rho = 0$ vs H_a: $\rho \neq 0$,

$$t = \frac{r(n-2)^{1/2}}{(1-r^2)^{1/2}}. \tag{14.3}$$

Values of $|t|$ that exceed a two-tailed t-value from Table A3 in the appendix lead to rejection of the null hypothesis. The degrees of freedom of the t-statistic are $v = n - 2$. One-sided tests can also be made using (14.3) and the appropriate one-sided t-value from this table.

The t-statistic for the tire-wear measurements is $t = 11.14$. The t value from Table A3 corresponding to a two-sided 5% significance level and $v = 14$ degrees of freedom is 2.145. Thus, the correlation coefficient is statistically significant, leading to the conclusion that there exists a linear relationship between the two tire-wear measurements. The magnitude of the t-statistic and the scatterplot in Figure 4.5 indicate that the linear relationship between the variates is a strong one.

To test for nonzero values of a population correlation coefficient, a different test statistic must be used. If one wishes to test H_0: $\rho = c$ vs H_a: $\rho \neq c$, where $|c| < 1$, the following test statistic should be used:

$$z = (n - 3)^{1/2}(\tanh^{-1} r - \tanh^{-1} c), \tag{14.4}$$

where $\tanh^{-1}(r)$ is the inverse hyperbolic-tangent function,

$$\tanh^{-1} r = \tfrac{1}{2} \ln \frac{1 + r}{1 - r}.$$

The statistic z in Equation (14.4) is approximately distributed as a standard normal random variable under the null hypothesis. Using critical values for two-tailed tests from Table A2 of the appendix, one can test the above hypothesis. One-tailed tests are conducted in a similar fashion.

Confidence intervals for correlation coefficients can be obtained using the statistic z in Equation (14.4). One uses the standard normal distribution of z to place confidence limits on $\tanh^{-1}\rho$ and then uses the hyperbolic-tangent function to transform these limits into limits on ρ. The interested reader is referred to the exercises for further details.

As mentioned above, the use of Pearson's r is appropriate only when both of the variates are random variables. Pearson's r can be calculated when one or both of the variables are not random, but none of the inference procedures discussed in this section should be used to draw conclusions about a linear relationship between the variables. The calculated value of r is simply a measure of the degree of linearity in the observed data values.

14.2 A STRATEGY FOR A COMPREHENSIVE REGRESSION ANALYSIS

A comprehensive regression analysis involves several different phases, each of which is important to the successful fitting of regression models. We describe these phases by the acronym PISEAS (plan, investigate, specify, estimate, assess, select). Table 14.4 summarizes some of the components of each of these phases. In this section we briefly describe them. Subsequent chapters explore each more fully.

TABLE 14.4 Components of a Comprehensive Regression Analysis

Plan. Plan the data collection effort.
Investigate. Investigate the data base; calculate summary statistics; plot the variables.
Specify. Specify a functional form for each variable; formulate an initial model, reexpressing variables as needed.
Estimate. Estimate the parameters of the model; calculate statistics which summarize the adequacy of the fit.
Assess. Assess the model assumptions; examine diagnostics for influential observations and collinearities.
Select. Select statistically significant predictor variables.

The first and one of the most important steps in any study for which a regression analysis is anticipated is to carefully plan (P) the data collection effort. Much of this textbook is devoted to such planning. Often, however, factors are not controllable — precisely the setting in which regression analysis derives its prominence among data analytic methods. When factors are not controllable, much of the planning can involve ensuring that the observations that are to be taken will be done in a manner that will provide data that are representative of the system, process, or population of interest. The goal in this setting is to acquire, to the greatest extent possible, data that one would collect if the factors of interest could be controlled.

The second step in a regression analysis is to investigate (I) the data base. A listing of the data base, especially if it is entered into a computer file, should be scanned in order to detect any obvious errors in the data entry — for example, negative values for variates that can only be positive, or decimal points entered in the wrong position. Erroneous or incorrectly coded data values can have a serious effect on the estimation of the model parameters; consequently, the detection of outliers at this stage of the analysis can save much time and effort later on and can lessen the chance of erroneous inferences due to outlier effects.

Along with the scanning of the data base, summary statistics for each of the variables should be computed. Incorrectly coded or erroneous data values often can be identified by examining the maximum and minimum data values for each variable. This is especially important when data bases are so large that scanning the individual data values is difficult.

Plotting response and predictor variables is another important step in the first phase of a regression analysis. A sequence plot of the response variable against its run number or against a time index can often identify nonrandom patterns in the experiment. For example, abrupt shifts could indicate unforeseen or unplanned changes in the operating conditions of an experiment. Cyclic trends could indicate a time dependence among the responses.

The response and predictor variables should also be plotted against one another. Such scatterplots may indicate that one or more of the variates need to be reexpressed prior to the fitting of a regression model. For example, plotting the response variable against each of the predictor variables may indicate the need to reexpress the response variable if several of the plots show the same nonlinear trend. If only one or two of the plots show nonlinear trends, the predictors involved should be reexpressed.

Box plots, histograms, and point plots are valuable aids for identifying possible outliers in a single variable. Scatterplots are useful for identifying combinations of data values that might not be apparent from an examination of each variable separately. In a scatterplot an observation might appear in a corner without having an extremely large or small value on either of the variates.

The third stage of a regression analysis is to initially specify (S) the form of the regression model. Not only must functional forms for each of the variables be specified, but one must consider whether interaction terms, polynomial terms, or nonlinear functions of the predictor variables should be included in the model.

The specification of a regression model relies heavily on one's knowledge of the experiment or process being studied. Not only does such knowledge aid one in selecting the variables to be included in the experiment and, thus, in the model, but it is also important when questions concerning the functional form of the model are discussed. Plots of the variables, as suggested above, can aid in the initial specification of the model, especially when an experiment is being conducted or a process is being studied for which little is known about the effects of variables on the response.

Ordinarily, if no indications to the contrary are shown by plots or by one's knowledge of the problem being studied, all variables are entered into the model in a linear fashion with no transformations. The model (14.1) is an example of the specification for a single predictor variable.

The next stage in a regression model is to estimate (E) the model parameters, ordinarily using appropriate computer software. In addition to just the estimates of the model parameters, regression programs usually provide statistics that assist one in assessing the fit. Thus, the estimation stage leads directly to the assessment stage of the analysis.

When assessing (A) the adequacy of the model fit one must consider several questions. Are the model assumptions reasonable (Chapter 18), including the functional form of the response and predictor variables? If so, does the model adequately fit the data? Can the errors be considered to be normally distributed with zero means and constant standard deviations? Are any observations unduly influencing the fit?

Following the assessment of the specified model, perhaps following a respecification of the model or some other remedial action, one proceeds to a

selection (S) of the individual predictors (Chapter 19). This is done because often one is unsure, prior to the experiment, which of the predictor variables affect the response.

As indicated throughout this discussion, each of these topics is explored more fully in later sections of this chapter and in subsequent chapters. Table 14.4 is intended to provide a general scenario for the analysis of data using regression models. One's decision at one stage of the analysis can cause the reevaluation of conclusions reached at a previous stage. For example, if an observation is deleted because it has an undue influence on the coefficient estimate of a predictor variable and then that variable is deleted from the model at a later stage of the analysis, one should reevaluate the need to delete the observation. Thus, the process is often an iterative one.

14.3 SCATTERPLOT SMOOTHING

Of the steps listed in Table 14.4 for comprehensively performing a regression analysis, correctly specifying (*S*) an initial regression model is often overlooked. Most initial least squares fits simply use the response and predictor variables as they are given in the database. Yet, there is nothing inherent in most measurements that guarantees that the variables are the most appropriate for least squares fitting of regressions models. This is why data plots are so indispensable when assessing whether the variables should be used in regression model fits as they are given in the database. A great aid in this endeavor is the use of diagnostic plots, such as box plots of individual variables and scatterplots of the response variable versus each predictor variable. A helpful graphical aid for assessing the nature of a relationship in a scatterplot is a method for smoothing variability in the plot. A *scatterplot smoother* filters variability in a scatterplot and more clearly shows trends that might be masked by the variability of the variables.

A scatterplot smoother produces a smooth curve through the data. Many different smoothing techniques are commonly available in statistical software. These include *kernel smoothers, smoothing splines*, and *locally weighted smoothers*. All of these help an analyst visualize a trend in the data and assist the data analyst in determining whether a variable transformation is needed.

One popular locally weighted smoother is commonly referred as a *loess* smoother. It is based on using a low-order polynomial regression model to predict each individual value, y, of the response variable. The term *local* refers to the fact that only data (x_i, y_i) in a neighborhood of y is used to fit the low-order polynomial at the point. The local regression model is given by

$$y = g(x) + e,$$

where y is the response variable, x is the predictor variable, and $g(x)$ is a smooth function of x. Commonly $g(x)$ is chosen to be a linear function,

$g(x) = \beta_0 + \beta_1 x$, or a quadratic function, $g(x) = \beta_0 + \beta_1 x + \beta_2 x^2$. The linear or quadratic fit at the point (x,y) uses only a neighborhood of points x_i around the given x value, and each (x_i, y_i) value is weighted by the distance of x_i from x. The smoothed curve for all the data is produced using the loess fitted values for each of the x values in the data set.

The number of data values, q, in the neighborhood around the point x will have an effect on the smoothness of the fit. A *span f*, or fraction of the data, is selected. Then $q = [f \cdot n]$, where $[u]$ is the integer part of u. The q data values (x_i, y_i) whose x_i are closest to x are used to fit y. As the span f increases towards 1, the fit becomes smoother but may only capture gross, overall trends in the data. On the other hand, as the span f approaches zero, the fit to each point becomes better but smoothness and any inherent trends are lost because the fit can rapidly change from one point to the next. In the limit as f approaches zero, each point is fit exactly.

The distance from x_i to x is often defined as

$$d(x, x_i) = \frac{\sqrt{(x - x_i)^2}}{s_x},$$

where s_x is the sample standard deviation of the predictor variable. A weight function that is often used with these distances is the *tricube* weight function, given by

$$W(u) = \begin{cases} (1 - u^3)^3 & 0 \leqslant u \leqslant 1 \\ 0 & \text{otherwise} \end{cases}$$

Using this function, the weights assigned to the points in a given neighborhood of x are defined as

$$w_i(x) = W\left\{\frac{d(x, x_i)}{d_q(x)}\right\},$$

where $d_q(x)$ is defined to be the qth smallest value of $d(x, x_i)$; that is, the largest distance in the neighborhood around x. For a locally linear loess predictor of y, the weighted least squares intercept and slope estimators are (with $w_i = w_i(x)$)

$$b_1 = \frac{\sum_{i=1}^{q} w_i(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{q} w_i(x_i - \overline{x})^2}$$

and

$$b_0 = \overline{y} - b_1\overline{x}.$$

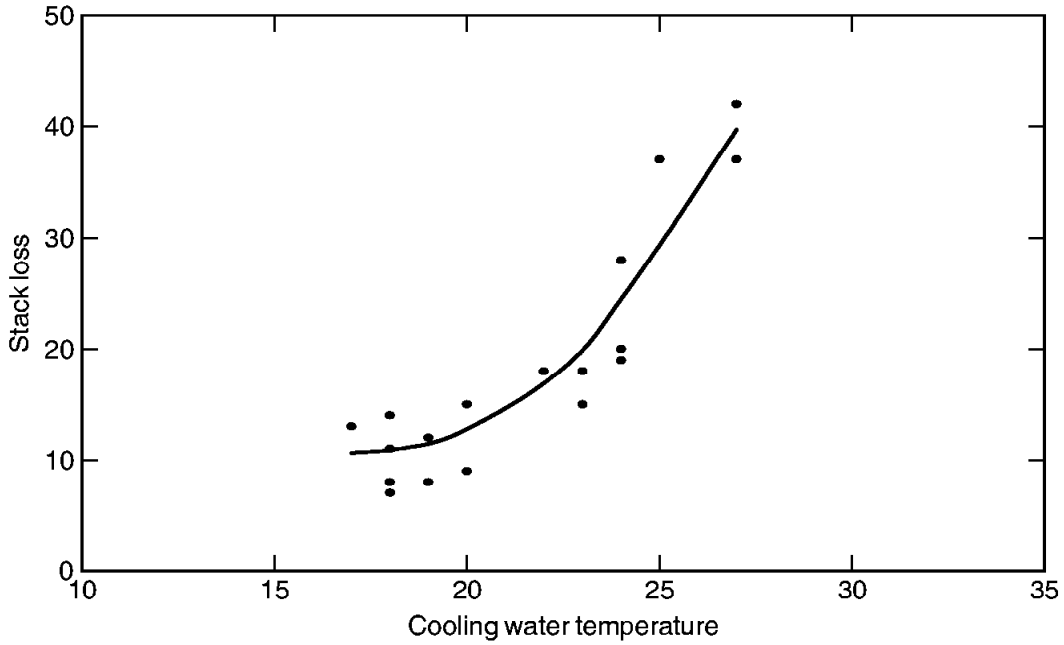


Figure 14.5 Loess-smoothed stack loss data values. Data from Brownlee, K.A. (1965). *Statistical Theory and Methodology in Science and Engineering*. New York: John Wiley & Sons, Inc. Copyright 1965 John Wiley and Sons, Inc. Used by permission.

The smoothed value of y is given as $\hat{y}(x) = b_0 + b_1 x$. In the next section, least squares estimators for intercepts and slopes will be discussed in detail. If all the weights w_i are set equal to 1 in the above equations, the weighted least squares estimators are identical to the least squares estimators shown in Exhibit 14.4. For locally smoothed predictions, the smoothed values must be calculated using separate fits for each (x, y).

As an example of the usefulness of loess smoothers, reconsider the stack loss versus cooling water temperature plot given in Figure 14.3. If there is interest in fitting these data, the scatterplot in Figure 14.3 suggests that there might be some curvature in the data. To confirm whether this is true, a loess-smoothed curve is plotted on the same data in Figure 14.5 using a span of $f = 0.75$ and a local linear function. Note the smoothness of the curve as well as the clear indication that in specifying the model it would be preferable to reexpress (Chapter 18) either the stack loss variable or the temperature variable.

14.4 LEAST-SQUARES ESTIMATION

The linear regression model used to relate a response variable to a single predictor variable was defined in Equation (14.1). The assumptions that ordinarily accompany this model are given in Table 14.1. We now wish to consider the estimation of the model parameters. Because we focus on several different

aspects of fitting regression models in this and the next section, we break each into several subsections.

14.4.1 Intercept and Slope Estimates

The model (14.1) can be written in a form similar to the ANOVA models that were discussed in Chapters 6 and 8:

$$y_i = \mu_i + e_i, \qquad i = 1, 2, \ldots, n. \tag{14.5}$$

This expression of the model coupled with the assumptions in Table 14.1 imply that the responses are independent normal random variables with means μ_i and common standard deviation σ. On comparing the model (14.5) with the model (14.1) it is apparent that the means of the responses are

$$\mu_i = \beta_0 + \beta_1 x_i. \tag{14.6}$$

Figure 14.6 depicts the linear relationship between the response μ and the predictor variable x, as well as the variability in the observed response y associated with the normally distributed errors for each value of x.

Let b_0 and b_1 denote some estimators of the intercept and slope parameters. Insertion of these estimators into (14.6) provides estimators of the means for each of the n responses. We term the equation

$$\hat{y} = b_0 + b_1 x \tag{14.7}$$

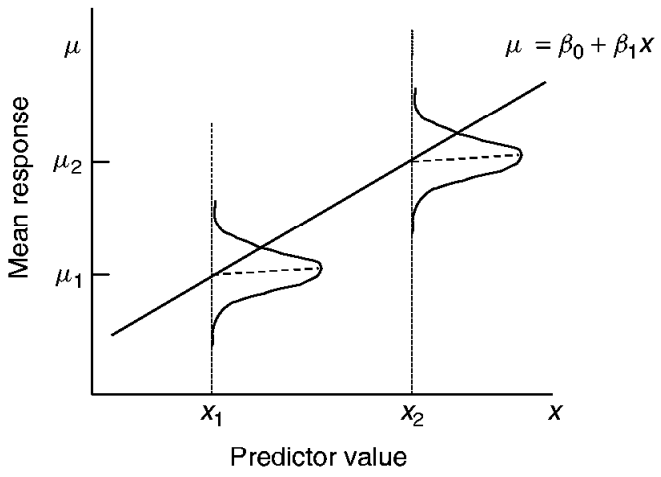


Figure 14.6 Regression model mean responses versus predictor values, with two error distributions overlaid.

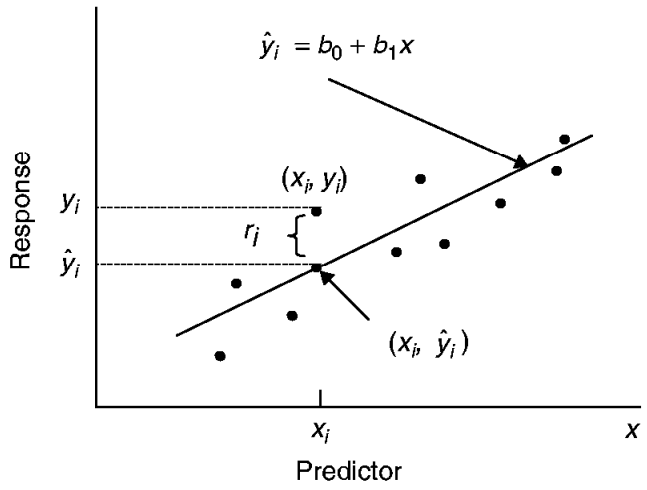


Figure 14.7 Relationships between observed responses (y_i), predicted responses ($\hat{y}_i$), and residuals (r_i).

the *fitted* regression model. Note that the fitted model is an estimator of the mean of the regression model, that is, the deterministic portion of the model (14.1). When the n predictor-variable values x_i are inserted into the prediction equation (14.7), the resulting quantities are estimates not only of the respective means μ_i but also of the actual responses y_i. The latter use of the fitted model results in the estimates being referred to as *predicted* responses. The differences between actual and fitted responses are termed *residuals*, r_i:

$$r_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i. \tag{14.8}$$

Figure 14.7 shows the relationship among the actual responses, fitted responses, and residuals.

The residuals (14.8) provide a measure of the closeness of agreement of the actual and the fitted responses; hence, they provide a measure of the adequacy of the fitted model. The sum of the squared residuals, SS_E, is an overall measure of the adequacy of the fitted model:

$$SS_E = \sum r_i^2 = \sum (y_i - b_0 - b_1 x_i)^2. \tag{14.9}$$

Note that SS_E is a function of the intercept and slope parameter estimators. A reasonable criteria for selecting intercept and slope estimators is to choose those that minimize the residual sum of squares. This is the principle of *least squares*, and the resulting estimators are called *least-squares estimators* (see Exhibit 14.4).

EXHIBIT 14.4 LEAST-SQUARES ESTIMATORS

1. Assume a linear regression model of the form (14.1).
2. Let $\hat{y}_i = b_0 + b_1 x_i$ denote predicted responses using some estimators b_0 and b_1 of the intercept and slope. Denote the residuals by $r_i = y_i - \hat{y}_i$.
3. Least squares estimators b_0 and b_1 minimize the sum of the squared residuals, $SS_E = \sum r_i^2$. The least-squares intercept and slope estimators are

$$b_0 = \bar{y} - b_1 \bar{x} \quad \text{and} \quad b_1 = s_{xy}/s_{xx}, \tag{14.10}$$

where $\bar{y}$ and $\bar{x}$ are the sample means of the response and the predictor variables, respectively, and

$$s_{xy} = (n-1)^{-1} \sum (y_i - \bar{y})(x_i - \bar{x}),$$

$$s_{xx} = (n-1)^{-1} \sum (x_i - \bar{x})^2.$$

Table 14.5 lists two measurements of the acid content of a chemical. The measurements of the response variable are obtained by a fairly expensive extraction and weighing procedure. The measurements of the predictor variable are obtained from an inexpensive titration method. The measurement error in the titration method is believed to be negligible. A scatterplot of the two variables indicates that a straight line should provide an excellent fit to the data.

Least-squares intercept and slope parameter estimates are also shown in Table 14.5. Because the linear fit approximates the observed data points so well, the residuals are all very small. Small residuals are one important indicator of the adequacy of a regression fit.

14.4.2 Interpreting Least-Squares Estimates

The least-squares intercept estimate is usually interpreted as the predicted response associated with the zero level of the predictor variable. Geometrically, it is the point where the fitted line crosses the vertical axis, when $x = 0$. Interpreting the intercept in this manner is an extrapolation when the predictor-variable values in the data set do not include the origin.

The slope is usually interpreted as the change in the fitted response associated with a unit (1) change in the predictor variable:

$$[b_0 + b_1(x+1)] - [b_0 + b_1 x] = b_1.$$

This interpretation is usually adequate but sometimes requires modification. We point out one occasional misinterpretation in this section and another one in Chapter 19.

TABLE 14.5 Acid-Content Data and Least-Squares Fit*

Acid Number x	Acid Content y	Predicted Acid Content	Residual
123	76	75.01	0.98
109	70	70.51	−0.51
62	55	55.40	−0.40
104	71	68.91	2.09
57	55	53.79	1.21
37	48	47.36	0.64
44	50	49.61	0.39
100	66	67.62	−1.62
16	41	40.60	0.40
28	43	44.46	−1.46
138	82	79.84	2.16
105	68	69.23	−1.23
159	88	86.59	1.41
75	58	59.58	−1.58
88	64	63.76	0.24
164	88	88.20	−0.20
169	89	89.81	−0.81
167	88	89.17	−1.17
149	84	83.38	0.62
167	88	89.17	−1.17

*$b_0 = 35.458$, $b_1 = 0.322$.

Data from Daniel, C. and Wood, F. S. (1971), *Fitting Equations to Data*, New York: John Wiley & Sons, Inc., Copyright 1971 John Wiley & Sons, Inc. Used by permission.

The sample correlation coefficient between the response variables y_i and the least-squares predicted responses $\hat{y}_i = b_0 + b_1 x_i$ is related to the least-squares slope estimator (14.10). If r is calculated as in (14.2) using $\hat{y}_i$ instead of x_i, the relationship is as follows:

$$b_1 = r(s_{yy}/s_{xx})^{1/2}. \tag{14.11}$$

Furthermore, because $\hat{y}_i$ and x_i are linearly related, the calculation of r using $\hat{y}_i$ gives the same result as the calculation using x_i in (14.2). Note that the slope estimate and the correlation coefficient are equal when the standard deviations of the two variables are the same. Whenever the standard deviations are the same the least-squares slope estimate will be less than 1 (in absolute value), because $b_1 = r$.

Occasionally researchers draw incorrect conclusions because they do not understand or know about the relationship (14.11). For example, in pre- and

posttesting of the same subjects prior to and following the administration of medical or psychological treatment, researchers often fit regression lines to assess the effectiveness of the treatment. An observed slope that is less than one is sometimes interpreted to mean that the treatment has lowered the posttest measurements from the pretest measurements.

Such an effect is to be expected, regardless of whether the treatment is effective, because the variability of the test scores is likely to be similar for the two sets of scores. Thus, the slope is less than one only because it is approximately equal to the sample correlation coefficient. The supposed effect occurs because of the mathematical relationship (14.11) and not because of a real treatment effect. This phenomenon is referred to as the "regression effect" or "regression to the mean effect," and its (erroneous) interpretation as the "regression fallacy."

14.4.3 No-Intercept Models

Least-squares estimators can be calculated for models that do not have intercepts (see Exhibit 14.5). Only minor modification of Equation (14.10) is required. These models are appropriate when (a) the response is believed to be zero when the predictor variable is zero and (b) the true model is believed to be linear from the origin to the range covered by the values of the predictor variable contained in the data set. The second condition is sometimes difficult to ensure when theoretical models are not known in advance of the data collection. In such cases it is wise to include an intercept term in the model and then test for its statistical significance.

EXHIBIT 14.5 LEAST SQUARES ESTIMATORS: NO-INTERCEPT MODELS

1. Assume a linear regression model of the form (14.1) but without the constant term β_0.
2. Let $\hat{y}_i = b_1 x_i$ denote predicted responses using some estimator b_1 of the slope. Denote the residuals by $r_i = y_i - \hat{y}_i$.
3. The least-squares estimator b_1 minimizes the sum of the squared residuals, $SS_E = \Sigma r_i^2$. The least-squares slope estimator is

$$b_1 = s_{xy}/s_{xx},$$

where s_{xy} and s_{xx} are not adjusted for the averages $\bar{y}$ and $\bar{x}$:

$$s_{xy} = n^{-1}\sum y_i x_i \quad \text{and} \quad s_{xx} = n^{-1}\sum x_i^2.$$

14.4.4 Model Assumptions

The calculation of least-squares estimates does not require the use of any of the assumptions listed in Table 14.1. Least-squares estimates can be obtained for any data set, regardless of whether the model is correctly specified. Statistical properties of the intercept and slope estimators, however, do depend on the assumptions that accompany the model. In particular, if one wishes to draw inferences about the model parameters or to make predictions using the fitted model, the model assumptions may be of critical importance.

The last abuse of regression modeling listed in Table 14.2 is of great concern because of the ability to calculate least-squares estimates for any regression data. In the next section, a number of inferential techniques, from examination of simple descriptive statistics to testing hypotheses about model parameters, will be discussed. All too frequently only the descriptive statistics are examined when the adequacy of the fit is assessed. The references at the end of this chapter contain numerous examples of how such a superficial examination can lead to misleading conclusions. We again stress the need to perform a comprehensive regression analysis using all the steps (PISEAS) listed in Table 14.4 to ensure that proper conclusions about the relationship between the response and the predictor variables are drawn.

14.5 INFERENCE

Inference on regression models can take the form of simply interpreting the numerical values of the intercept and slope estimates or the construction of tests of hypotheses or confidence intervals. In this section we present several inferential techniques that can be used with regression models.

14.5.1 Analysis-of-Variance Table

The same procedures used to derive the ANOVA table in Section 6.1. can be used to derive similar tables for regression models. We again use as a measure of the total variation in the response variable the total sum of squares:

$$\text{TSS} = \sum (y_i - \bar{y})^2.$$

The error sum of squares is again equal to

$$\text{SS}_E = \sum (y_i - \hat{y}_i)^2.$$

The model sum of squares is also the same, but it is now referred to as the regression sum of squares:

$$\text{SS}_R = \sum (\hat{y}_i - \bar{y})^2 = b_1^2 \sum (x_i - \bar{x})^2. \tag{14.12}$$

TABLE 14.6 Symbolic ANOVA Table for Regression Models Having One Predictor Variable

Source of Variation	df	Sum of Squares	Mean Squares	F-Value	p-Value
Regression	1	SS_R	$MS_R = SS_R$	$F = MS_R/MS_E$	p
Error	$n-2$	SS_E	$MS_E = SS_E/(n-2)$		
Total	$n-1$	TSS			

A symbolic ANOVA table for regression models having a single predictor variable is shown in Table 14.6. The source of variation labeled "Regression" reflects the variability in the response variable that can be attributed to the predictor variable. There is only one degree of freedom for regression, because only one coefficient, namely β_1, must be estimated to obtain the regression sum of squares (14.12).

The estimate of the standard deviation, $s_e = (MS_E)^{1/2}$, is a measure of the adequacy of the fitted regression line. It is important to realize that the standard deviation of the error terms in the model (14.1) measures the variability of the observations about the regression line. Thus, an estimated standard deviation that is small relative to the average response indicates that the observed responses are tightly clustered around the fitted line, just as a small model standard deviation σ indicates that the response values are clustered tightly around the mean regression line (14.6).

The estimate s_e is model-dependent. The value of s_e measures not only the uncontrolled variation of the responses but also contributions due to any model misspecification that may occur. This *lack of fit* of the data to the assumed model can be a substantial portion of the magnitude of SS_E and, hence, of s_e. In Section 15.2 we describe a technique for assessing possible lack of fit of a regression model when repeat observations are available for each of several values of the predictor variable. Residual-based techniques for assessing model misspecification when repeat observations are not available are detailed in Section 18.2.

The sample correlation r between the actual and fitted responses is another measure of the adequacy of the fit. The square of this sample correlation is readily computed from the statistics in the analysis of variance table and is termed the *coefficient of determination* (R^2):

$$R^2 = [\text{corr}(y, \hat{y})]^2 = 1 - \frac{SS_E}{TSS}. \tag{14.13}$$

The coefficient of determination could be calculated directly by calculating Pearson's r between the observed and predicted responses, but the above formula is simpler and less subject to roundoff error. Coefficients of determination are usually expressed as a percentage by multiplying Equation (14.13) by 100. Then R^2 values lie between 0 and 100%. The closer they are to the upper bound, the better is the fit.

Although coefficients of determination are very popular measures of the adequacy of the fitted model, there is a tendency to overrely on them because of their straightforward interpretation. Coefficients of determination can be made arbitrarily large even when the fit to the data is poor. For example, a single observation, if sufficiently removed from the bulk of the data, can produce very large R^2 values even though there is no linear association for the majority of the observations in the data base. For this reason, we again stress that a comprehensive regression analysis. (PISEAS) is required before satisfactory conclusions can be drawn about the adequacy of the fit.

The F-statistic in the ANOVA table allows one to test the statistical significance of the slope parameter. Unlike tests for the significance of factor effects in the analysis of designed experiments, the test for significance of the slope parameter is ordinarily conducted using a large significance level, say $\alpha = 0.25$. The change in the significance level is due to the desire to protect against Type II errors (see Section 2.6); that is, erroneously concluding that the predictor variable is not useful in modeling the response. Even though the Type II error probability is unknown (because the true value of β_1 is unknown), we seek to reduce it by using a large significance level.

The ANOVA table for the acid content data in Table 14.5 is shown in Table 14.7. Note that the F-statistic is highly significant. The estimated model standard deviation is $s_e = 1.23$. The coefficient of determination, $R^2 = 99.47\%$, also indicates that the prediction equation fits the observed responses well.

TABLE 14.7 ANOVA Table for the Acid-Content Data

Source of Variation	df	Sum of Squares	Mean Squares	F-Value	p-Value
Regression	1	5071.57	5071.57	3352.33	0.000
Error	18	27.23	1.51		
Total	19	5098.80			

14.5.2 Tests and Confidence Intervals

A t-statistic can be constructed for testing H_0: $\beta_1 = c$ versus H_a: $\beta_1 \neq c$, where c is a specified constant, using the general procedures outlined in Section 2.6. From the assumptions listed in Table 14.1 it follows that the least-squares slope estimator b_1 has a normal distribution with mean β_1 and standard error $\sigma/\left[\sum(x_i - \bar{x})^2\right]^{1/2}$. The following statistic has a Student t-distribution with $n - 2$ degrees of freedom:

$$t = \frac{b_1 - \beta_1}{s_e / s_{xx}^{1/2}}, \qquad s_{xx} = \sum (x_i - \bar{x})^2. \tag{14.14}$$

Inserting c for β_1 in Equation (14.14) provides the t-statistic that is suitable for testing the above hypotheses.

Inserting $c = 0$ in Equation (14.14) and squaring the resulting t-statistic yields the F-statistic from the ANOVA table, Table 14.6. This t-variate can also be used to form confidence intervals on the slope parameter. Using the same type of procedures discussed in Section 2.4, the following limits for a $100(1 - \alpha)\%$ confidence interval for β_1 can be derived:

$$b_1 - t_{\alpha/2} \frac{s_e}{s_{xx}^{1/2}} \leqslant \beta_1 \leqslant b_1 + t_{\alpha/2} \frac{s_e}{s_{xx}^{1/2}}, \tag{14.15}$$

where $t_{\alpha/2}$ is a $100(\alpha/2)\%$ upper-tail t critical value having $n - 2$ degrees of freedom.

For the acid-content data, a 95% confidence interval for β_1 using (14.15) is

$$-0.322 \pm 2.101 \frac{1.230}{(54, 168.2)^{1/2}},$$

or

$$-0.333 \leqslant \beta_1 \leqslant -0.311.$$

The tightness of this interval is in part due to the small estimated model standard deviation, $s_e = 1.230$.

Tests of hypotheses and confidence intervals for the intercept parameter can be made using the following t-variate:

$$t = \frac{(b_0 - \beta_0)}{s_e (n^{-1} + \bar{x}^2 / s_{xx})^{1/2}}. \tag{14.16}$$

14.5.3 No-Intercept Models

When the regression model does not contain an intercept term, modifications of several of the above statistics are required. The total sum of squares is now unadjusted, and the regression sum of squares is not adjusted for the average of the predictor variables:

$$\text{TSS} = \sum y_i^2 \quad \text{and} \quad \text{SS}_R = b_1^2 \sum x_i^2. \tag{14.17}$$

Likewise the number of degrees of freedom for error and the total number of degrees of freedom are now $n - 1$ and n, respectively. The t-variate that is used to test for the significance of the slope parameter and to form confidence intervals is adjusted by replacing s_{xx} in (14.14) with

$$s_{xx} = \sum x_i^2 \tag{14.18}$$

and using $n - 1$ degrees of freedom rather than $n - 2$.

14.5.4 Intervals for Responses

It is often of interest to form confidence intervals around $\mu = \beta_0 + \beta_1 x$ for a fixed value of x. Under the assumptions listed in Table 14.1, the predicted response $\hat{y}$ has a normal probability distribution with mean $\mu = \beta_0 + \beta_1 x$ and standard deviation $\sigma[a_1 + (x - a_2)^2/s_{xx}]^{1/2}$, where

$$a_1 = n^{-1}, \quad a_2 = \bar{x}, \quad \text{and} \quad s_{xx} = \sum (x_i - \bar{x})^2 \qquad \text{for intercept models}$$

and

$$a_1 = 0, \quad a_2 = 0, \quad \text{and} \quad s_{xx} = \sum x_i^2 \qquad \text{for no-intercept models.}$$

Thus the following t-variate can be used to form confidence intervals for an expected response μ:

$$t = \frac{\hat{y} - \mu}{s_e[a_1 + (x - a_2)^2/s_{xx}]^{1/2}}. \tag{14.19}$$

If one wishes to form prediction intervals for an actual future response, not the expected value of a response, equation (14.19) can again be used if one replaces μ by y_f and a_1 by $1 + a_1$. In this formulation y_f represents the future response and $\hat{y}$ its predicted value. The change from a_1 to $1 + a_1$ in the formula for the standard error occurs because the future response has a standard deviation σ, an added component of variability to that of the predicted response $\hat{y}$. Thus the standard deviation of $\hat{y} - y_f$ is $\sigma\{1 + a_1 + (x - a_2)^2/s_{xx}\}^{1/2}$.

REFERENCES

Text References

There are many excellent texts on regression analysis. Among the more current, comprehensive texts are the following:

Chatterjee, S., Hadi, A., and Price, B. (2000). *Regression Analysis by Example, Third Edition*. New York: John Wiley & Sons, Inc.

Cook, R. D. and Weisberg, S. (1999). *Applied Regression Analysis Including Computing and Graphics*. New York: John Wiley & Sons, Inc.

Daniel, C. and Wood, F. S. (1971). *Fitting Equations to Data*, New York: John Wiley and Sons, Inc.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis, Third Edition*, New York: John Wiley & Sons, Inc.

Gunst, R. F. and Mason, R. L. (1980). *Regression Analysis and Its Application: A Data-Oriented Approach*, New York: Marcel Dekker, Inc.

Montgomery, D. C., Peck, E. A and Vining, C. G. (2001). *Introduction to Linear Regression Analysis, Third Edition*, New York: John Wiley & Sons, Inc.

Myers, R. H. (1986). *Classical and Modern Regression with Applications*, Boston, MA: PWS and Kent Publishing Co.

Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996). *Applied Linear Regression Models, Third Edition*, New York: Richard D. Irwin, Inc.

Ryan, T. P. (1997). *Modern Regression Methods*, New York: John Wiley & Sons, Inc.

Theoretical information on smoothing and different smoothing procedures is given in the following two references:

Chambers, J. M. and Hastie, T. J. (1993). *Statistical Methods in S*. New York: Chapman & Hall, Inc.

Cleveland, W. S. and Devlin, S. J. (1988). "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, **83**, 596–610.

Data References

The tire-wear data are taken from Natrella (1963); see the references at the end of Chapter 4. The stack-loss data are taken from:

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering, Second Edition*, New York: John Wiley & Sons, Inc.

The acid content data are taken from Daniel and Wood's text, referenced above.

EXERCISES

1 An electronics firm is interested in documenting a relationship between sales price (y) of one of its videocassette recorders (VCR) and its

factory-authorized discount (x) used in promotions of the VCR. The firm authorizes different discounts over a 15-week period and determines the average sales price of the recorders sold during each week. The data are shown below:

Week	Price ($)	Discount($)	Week	Price($)	Discount($)
1	320	40	9	340	15
2	320	46	10	335	25
3	315	48	11	305	50
4	310	49	12	350	13
5	340	14	13	325	40
6	315	45	14	350	15
7	335	20	15	330	25
8	335	30			

Investigate the data set. Pay special attention to the presence of extreme data values and the possible need for transformations of the variables. Superimpose a loess smooth on a scatterplot to assist in evaluating the need for a transformation.

2 Using the data in Exercise 1, estimate the least-squares intercept and slope coefficients. Plot the fit on a scatterplot of the data values. Is the fit visually satisfactory? Why (not)?

3 An engineering research firm is conducting an experiment to compare the strength of various metals under dynamic load conditions. A cylindrical weight is allowed to drop freely down a rod and hit a spring at the base of the rod. Thin rectangular strips of different metals are attached to the weight so that they extend radially outward from the weight perpendicular to the direction of fall. These cantilever beams each experience permanent deformation, depending on the drop height. The deformation values tabulated below are from beams of 5052-0 aluminum having a free length of 10 inches, a width of 0.5 inches, and a thickness of 0.04 inches:

Drop Height (in.)	Deformation
20	4.00
20	4.85
25	5.45
25	5.60
30	6.25
30	6.10
35	7.00
35	7.40
40	8.00
40	7.30

Investigate the suitability of performing a linear regression analysis by making a scatterplot (include a loess smooth) of the observations. Should a no-intercept model be fit to the data? Why (not)? Place a confidence interval on the intercept parameter. Does this interval confirm your conclusion about the desirability of using a no-intercept model?

4 The following measurements are of beta-particle emissions of a several-hundred-year-old geological artifact. The counts are taken from two channels of the same decay counter, the counts differing because one counter counts emissions over a wider energy range than the other. Calculate Pearson's r for these data. State two reasons why a correlation analysis of these data is more appropriate than a regression analysis.

Channel A	Channel B	Channel A	Channel B
3287	2418	3184	2325
3130	2328	3129	2318
3183	2366	3147	2359
3168	2405	3181	2350
3294	2445	3246	2439
3300	2454	3216	2389
3213	2352	3229	2394
3263	2416	3294	2420
3195	2401	3212	2416
3252	2412	3252	2418
3288	2404	3254	2389

5 Construct an ANOVA table for the beam deformation data in Exercise 3. Does the R^2 value suggest that the fit is adequate? Construct a confidence interval for the mean response for a drop height of 30 in. Construct a second confidence interval for the mean response, this time for a drop height of 50 in. Based on the widths of these two confidence intervals, what is the effect on the confidence interval of moving from the center to the extremes of the predictor variable values?

6 In a tire-treadwear study, a factor believed to influence the amount of wear is the ambient temperature. The data below represent wear rates of one brand of tires that were measured in an experiment. Each of the test runs was made at a different temperature, and all were made using the same vehicle. Perform a comprehensive regression analysis on these data. Are the intercept and slope coefficients statistically significant? Do you consider the fit to the data satisfactory? Why (not)?

Temperature (°F)	Tire Wear (mm)
66.0	1.17
70.3	1.00
53.2	1.15
53.3	1.07
88.1	0.97
89.6	0.94
78.4	1.33
76.9	1.36

7 A study was conducted to investigate factors that might contribute to tooth discoloration in humans. Of particular interest was the possible relationship between fluoride levels and tooth discoloration. Dental examinations were given to 20 participants at the beginning and the end of the year-long study. The participants were selected from communities that differed in the fluoride levels in the public water supply. The response of interest was the percentage increase in discoloration over the course of study. Perform a regression analysis of the data given below and assess the adequacy of the fit. In particular, evaluate whether fluoride level alone is satisfactory in predicting the response variable.

Fluoride Level (ppm)	Discoloration (%)	Fluoride Level (ppm)	Discoloration (%)
0.7	12	0.5	11
1.3	10	1.8	15
1.5	14	2.1	22
2.4	20	0.4	21
2.6	18	3.4	27
2.9	18	2.9	25
3.0	25	2.1	18
3.6	21	3.4	21
3.8	36	3.0	18
4.0	44	1.7	20

8 An investigation was performed to determine whether two measures of the strength of a specific type of metal could be considered to be equivalent. One measure is a tensile strength measurement and the other is a Brinell hardness number. Two questions are of importance in this investigation:

(a) Are the measures linearly related?

(b) If so, how strong is the relationship?

Use the data below to answer these questions.

Sample:	1	2	3	4	5	6	7
Brinell Hardness:	105	107	106	107	101	106	100
Tensile Strength:	38.9	40.4	39.9	40.8	33.7	39.5	33.0

9 A study was performed to determine the relationship between NO_x emissions and the age, using mileage as a surrogate, for a specific automobile model. A total of thirty eight different vehicles with differing mileages were used and an emissions test was run on each engine to determine the NO_x emissions (in grams/horsepower-hour). The data are given below. Construct a scatterplot of the data and include a loess smooth. Discuss the possible need for a transformation of the data.

Mileage	**NO_x (g/mile)**	**Mileage**	**NO_x (g/mile)**	**Mileage**	**NO_x (g/mile)**
11200	0.271	25100	0.124	20300	0.254
14500	0.166	39100	0.207	26300	0.288
92100	0.548	58100	0.203	26800	0.269
46000	0.422	43200	0.192	26100	0.277
54100	0.671	19800	0.294	48900	0.345
93500	0.538	28800	0.197	91200	0.341
71500	1.153	35000	0.255	58800	0.368
52000	0.491	38300	0.178	83200	0.501
30800	0.111	28200	0.248	68200	0.582
60800	0.381	50500	0.347	65300	0.428
48800	0.315	45500	0.416	67800	0.577
63500	0.541	43300	0.238	77600	0.316
25000	0.205	62800	0.459		

10 Using the data in Exercise 9, estimate the least-squares intercept and slope coefficients after applying any needed transformations. Plot the fitted equation on a scatterplot of the data. Is the fit satisfactory? Why (not)?

11 Use the data given in Exercise 4 to answer the following questions.

(a) Compute the pairwise correlation coefficient, r, between the Channel A and Channel B measurements.

(b) Use the results of (a) to test whether the population correlation coefficient, ρ, between the Channel A and B measurements is zero versus it is nonzero. Use a 0.10 significance level.

(c) Compute a 90% confidence interval for the population correlation coefficient in (a). Hint: The inverse hyperbolic-tangent transformation $y = \tanh^{-1}(r) = \ln\{(1 + r)/(1 - r)\}/2$ is approximately normally distributed with mean $\tanh^{-1}(\rho)$ and variance $1/(n - 3)$.

12 Plot the data in Exercise 6 on a scatter plot. Superimpose on it confidence intervals about the line and prediction intervals about the line. Give the prediction interval for tire wear when the ambient temperature is 85, and explain how this differs from the confidence interval at this temperature value.

13 Birds near airports present a hazard to aircraft with jet engines, as the large intakes of the engines can draw the birds into the engine. This ultimately can lead to an engine failure. A study was conducted to determine how the heart rates of captive wild birds respond to approaching aircraft during the take-off period. Twelve birds were equipped with heart rate monitors and placed in individual cages besides an active runway. Measurements were taken on the maximum heart rate of the bird and on the distance that the plane was from the bird when the maximum heart rate occurred. The resulting data are given below.

(a) Find the prediction equation using distance as the predictor variable and maximum heart rate as the response variable.

(b) Construct an ANOVA table and test the hypothesis that the slope is zero.

(c) Find a confidence interval on the slope.

(d) Comment on the validity of the assumption that the predictor variable is measured without error.

Maximum Heart Rate (beats/min)	Distance from Bird (ft)
287	1816
283	1588
301	1410
293	1192
298	1012
295	915
313	809
311	603
301	469
314	398
327	201
369	0

14 Reconsider the data given in Exercise 13. Using a point plot, box plot, and histogram of the maximum heart rates, determine if any observations in the data are outliers.

15 An automobile manufacturer is interested in predicting the emissions of one of its vehicle models at 100,000 miles of usage. This is needed to determine if its fleet will meet established emissions standards at that time point. Unfortunately, the vehicle model is fairly new and few high-mileage vehicles are available. The manufacturer decides to locate 31 representative vehicles, test them for emissions, and record their mileage. The objective is to obtain a prediction equation of emissions as a function of mileage and then predict the emissions value at 100,000 miles. The resultant data are given below for a coded emissions value for total hydrocarbons (THC).

(a) Find a prediction equation using mileage as the predictor variable and THC as the response variable.

(b) Construct an ANOVA table and test the hypothesis that the slope is zero.

(c) Comment on the goodness of fit of this model.

(d) Predict the THC value at 100,000 miles, and present a 90% prediction interval for that value.

Mileage (miles)	Coded THC
54093	1.332
48796	1.196
63461	1.246
71541	1.544
52040	1.183
60836	1.171
62800	1.355
46063	1.099
28192	1.083
45343	1.151
43347	1.144
30846	1.084
25140	1.082
58078	1.121
25008	1.105
39107	1.136
35045	1.150
20290	1.053
43248	1.139
19758	1.062
28736	1.058

Mileage (miles)	Coded THC
26268	1.086
26769	1.050
26069	1.061
48872	1.168
68247	1.391
65349	1.351
58782	1.384
67765	1.295
77592	1.865
66312	1.530
71218	1.412
74000	1.700

16 Examine a scatter plot of the data given in Exercise 15 and include a loess smooth. Apply any needed transformation, and refit the prediction equation. What do you predict for the THC emissions when the mileage is 100,000 miles.

17 An engineer is designing an electronic controller for an engine. To find the proper settings to maximize the engine operation, he collects data on engine throttle as a function of engine torque, with the ultimate goal of obtaining a prediction equation for engine throttle. The data are given below for 24 engine settings.

(a) Find a prediction equation using torque as the predictor variable and throttle as the response variable.

(b) Construct an ANOVA table, and test the hypothesis that the slope is zero.

(c) Comment on the goodness of fit of this model.

Throttle	Torque
0.113	−29
1.221	179
2.263	325
3.334	385
4.365	407
5.510	411
0.114	−87
1.213	44
2.271	229
3.329	452
4.375	642

Throttle	Torque
5.480	754
0.116	−99
1.207	−7
2.275	162
3.324	375
4.390	559
5.450	670
0.121	−111
1.216	−47
2.277	100
3.330	282
4.385	505
5.510	577

18 A study is conducted to determine the relationship between emissions and fuel properties for a diesel engine. Fifteen fuels were used in an engine that ran on a dynamometer for a fixed period of time. The results for the NO_x emission and the density fuel property values are given below.

(a) Find a prediction equation using density as the predictor variable and NO_x as the response variable.

(b) Construct an ANOVA table, and test the hypothesis that the slope is zero.

(c) Comment on the goodness of fit of this model.

NO_x	Density
2.572	0.8481
2.338	0.8279
2.383	0.8282
2.373	0.8287
2.446	0.8593
2.441	0.8595
2.498	0.8600
2.423	0.8285
2.408	0.8289
2.455	0.8293
2.643	0.8599
2.634	0.8602
2.659	0.8605
2.306	0.8291
2.547	0.8569

19 Examine a scatter plot of the data given in Exercise 18 and include a loess smooth. What do you notice that is peculiar about this plot? What do you think might be the cause?

20 A meetings planner for a professional society is interested in obtaining a regression equation to predict the number of technical papers that will be presented at an annual meeting based on the number of registrants. Such an equation is useful as it helps the planner determine the number of rooms needed for the presentations. Six years of past data are available and are given below.

(a) Find a prediction equation using registration count as the predictor variable and number of papers presented as the response variable.

(b) Construct an ANOVA table, and test the hypothesis that the slope is zero.

(c) Comment on the goodness of fit of this model.

(d) Suppose the meetings planner finds that there are 4,088 registrations for the meeting for this year. Predict the average number of papers expected to be presented at this meeting, and compute a 95% confidence interval for that value. Comment on whether it would be better to use this interval for planning purposes or to construct a prediction interval for the actual number of papers to be presented.

Registrants	Number of Papers
3760	1202
3262	1101
4158	1216
3689	1142
3362	932
4518	1306

CHAPTER 15

Linear Regression with Several Predictor Variables

The methods presented in Chapter 14 for regression on a single predictor variable can be generalized to the regression of a response variable on several predictors. Due to interrelationships among the predictor variables, both the fitting of multiple regression models and the interpretation of the fitted models require additional specialized techniques that are not ordinarily necessary with single-variable models. In this chapter we concentrate on the following topics:

- *least-squares parameter estimation,*
- *inferential techniques for model parameters,*
- *interaction effects for quantitative predictors, and*
- *polynomial regression models.*

A multiple linear regression analysis involves the fitting of a response to more than one predictor variable. For example, an experimenter may be interested in modeling automobile emissions (y) as a function of several fuel properties, including viscosity (x_1), cetane number (x_2), and a distillation temperature (x_3). Additional complications are introduced into such an analysis beyond the single-variable analyses discussed in the last chapter, because of the interrelationships that are possible among the predictors. As with ANOVA models, two or more predictors can have synergistic effects on the response, leading to the need to include joint functions of the several predictors in order to model the response adequately.

A benefit of the modeling of a response as a function of several predictor variables is flexibility for the analyst. A wider variety of response variables can be satisfactorily modeled with multiple regression models than can be

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

with single-variable models. In a single-variable analysis one is confined to using functions of only one predictor. In multiple regression analyses, different individual and joint functional forms for each of the predictors is permitted. Direct comparisons of alternative choices of predictors also can be made.

In this chapter we begin a comprehensive development of the use of multiple-regression modeling methodology. The first two sections of this chapter examine, respectively, least-squares parameter estimation, and statistical inference on model parameters. The last two sections detail issues of special concern to the widespread use of interaction terms (products of quantitative predictors) and polynomial terms (powers and products of predictors).

15.1 LEAST SQUARES ESTIMATION

Multiple linear regression models can be defined as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i, \qquad i = 1, 2, \ldots, n. \tag{15.1}$$

This model is a "linear" regression model because the unknown coefficients appear in linear forms, that is, as additive constants or multipliers of the values of the predictor variables. The predictor variables x_j may be functions of other variables so long as there are no unknown constants needed to determine their values. The predictor variables may be intrinsically numerical or they may be categorical variables. The predictor variables can be powers or products of other predictors, for example, $x_3 = x_1 x_2$.

As with the simple linear regression model described in Chapter 14, several assumptions usually accompany this model. These assumptions are listed in Table 14.1 and discussed in Chapter 18. None of these model assumptions need be valid to calculate least-squares estimates of the model parameters. The model assumptions are of critical importance if one wishes to make statistical inferences on the true model parameters or on the distribution of the model errors.

15.1.1 Coefficient Estimates

As observed in Chapter 14, the model in equation (15.1) can be written similarly to an ANOVA model:

$$y_i = \mu_i + e_i, \qquad i = 1, 2, \ldots, n, \tag{15.2}$$

where the means, or *expected values*, of the response variables y_i are given by

$$\mu_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \tag{15.3}$$

To estimate these n means, we insert estimates of the coefficients into the right side of equation (15.3):

$$m_i = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p. \tag{15.4}$$

Equation (15.4) can be used to estimate not only the means μ_i of the response variables but also the actual responses y_i. When used for this latter purpose the resulting estimates are termed *predicted* responses, notationally indicated by replacing m_i with $\hat{y}_i$, and equation (15.4) is called a *prediction equation.* Again these comments parallel those given in Chapter 14 for linear regression with a single predictor variable.

One method of obtaining the coefficient estimates in equation (15.4) is to use the principle of least squares. The residuals between the observed and predicted responses are

$$r_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_p x_{ip}. \tag{15.5}$$

The sum of the squared residuals, SSE, is given by

$$\mathrm{SS}_E = \sum r_i^2 = \sum (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_p x_{ip})^2. \tag{15.6}$$

The coefficient estimates, $b_0, b_1, \ldots, b_p$, that minimize SS_E are termed the least-squares coefficient estimates (see Exhibit 15.1).

EXHIBIT 15.1 LEAST-SQUARES ESTIMATORS FOR MULTIPLE LINEAR REGRESSION MODELS

1. Assume a multiple linear regression model of the form (15.1).

2. Least-squares estimators $b_0, b_1, \ldots, b_p$ minimize the sum of the squared residuals,

$$\mathrm{SS}_E = \sum (y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip})^2.$$

The process of minimizing SS_E and obtaining coefficient estimates involves the differentiation of equation (15.6) with respect to each of the $p + 1$ unknown regression coefficients. The $p + 1$ derivatives are then set equal to zero, and the least-squares estimates are obtained by simultaneously solving the equations. We leave the algebraic details as an exercise for the interested reader.

The $p + 1$ equations that must be simultaneously solved to find the least-squares estimators are termed *normal equations*. The normal equations for a model having $p = 2$ predictor variables are

$$\begin{aligned} \bar{y} &= b_0 + b_1\bar{x}_1 + b_2\bar{x}_2, \\ \sum x_{i1}y_i &= b_0 \sum x_{i1} + b_1 \sum x_{i1}^2 + b_2 \sum x_{i1}x_{i2}, \\ \sum x_{i2}y_i &= b_0 \sum x_{i2} + b_1 \sum x_{i1}x_{i2} + b_2 \sum x_{i2}^2. \end{aligned} \tag{15.7}$$

There are three equations and three unknowns in (15.7). One can simultaneously solve them and obtain the unique least-squares coefficient estimates for β_0, β_1, and β_2. Ordinarily, the existence of a unique solution for the least-squares estimators requires that the sample size n be greater than the number p of predictor variables.

As the number of predictor variables increases, the solutions to the normal equations are ordinarily obtained using computer software. The least-squares fits presented in this chapter are all obtained using computer algorithms that solve the normal equations. The algebraic solutions to the normal equations are most easily expressed in matrix notation. The interested reader is referred to the appendix to this chapter for the algebraic solutions.

Least-squares estimators also can be calculated for models that do not have an intercept term β_0. Whether to include an intercept term in the model is an option with most computer regression algorithms. When one is uncertain about including an intercept term, it should be kept in the model and then tested for statistical significance (see Section 15.2).

15.1.2 Interpreting Least-Squares Estimates

The least-squares coefficient estimates for the regression model in (15.1) have a slightly modified interpretation from that given for the slope estimate in Chapter 14. The estimates generally are said to measure the change in the predicted response variable due to a unit change in one predictor variable while all remaining predictor variables are held constant. For this reason the estimates often are termed *partial* regression coefficient estimates.

Because of interrelationships among predictor variables, it is not always possible to change one predictor variable while holding the others fixed. The magnitudes of the coefficient estimates themselves can depend heavily on which other predictor variables are included in the regression model. Adding or deleting predictor variables can cause the estimated coefficients to drastically change in magnitude as well as sign. These problems become acute when severe collinearity is present among the predictor variables (see Section 15.4 and 19.4).

EXHIBIT 15.2 INTERPRETATION OF LEAST-SQUARES ESTIMATES

The least-squares estimates b_j measure the change in the predicted response $\hat{y}_i$ associated with a unit change in x_j after adjusting for its common variation with the other predictor variables.

An explicit interpretation of the least-squares coefficient estimates, one that takes into account the above difficulties, is given in Exhibit 15.2. According to this interpretation, if one desires to determine the effect on $\hat{y}$ due to changes in x_j, the predictor variable values for all predictor variables must be specified. The net changes in $\hat{y}$ due to the change in x_j are then a function of the changes not only of x_j but also of the specific changes that occur in the other predictor variables. Only if x_j does not vary systematically with the other predictor variables can the usual interpretation that b_j measures the effect of a unit change in x_j be considered appropriate.

The specific adjustment referred to in the above interpretation for least-squares estimators can be obtained as follows. Just as linear regression can be used to determine a linear relationship among the response and the predictor variables, it also can be used to relate the predictors to one another. Regress x_j on the other $p - 1$ predictor variables and a constant term. Denote the residuals from this fit by r^*:

$$r_i^* = x_{ij} - \hat{x}_{ij}.$$

The least-squares estimate b_j measures the change in the response due to a unit change in r^*, not x_j. Note that if x_j cannot be well predicted by the other predictor variables, r^* is approximately equal to $x_j - \bar{x}_j$ and hence b_j can be interpreted in the usual manner as measuring the change in $\hat{y}$ associated with a unit change in x_j.

To more clearly demonstrate this important interpretation of least-squares coefficient estimates for multiple regression models, consider the following illustration. An experiment was conducted to study the effects of ambient temperature on tire treadwear for four brands of tires, one of which was used as a control. Two convoys, each consisting of four cars of the same model, were driven day and night using replicate sets of tires during two seasons of the year, winter and summer. Rotation of the vehicles in the convoy and of tire brands on the vehicles ensured that no systematic bias would be incurred on any tire brand due to the experimental procedure. Various forms of randomization were also included in the experimental design to eliminate systematic vehicle and driver biases on the comparison of brand effects. Thirty-two tires of each of the four tire brands were exposed to 8,000 miles of wear. One response

TABLE 15.1 Tire Treadwear Rates for Tire Brand C

Relative Wear	Temperature (°F)	Wet Miles (mi)
2.25847	53.2	388
2.19915	53.2	388
2.19068	53.2	388
1.99153	53.2	388
2.27837	53.3	438
2.22698	53.3	438
2.03854	53.3	438
2.05139	53.3	438
2.68891	66.0	58
2.63003	66.0	58
2.59078	66.0	58
2.58587	66.0	58
2.35556	70.3	7
2.41333	70.3	7
2.34222	70.3	7
2.42667	70.3	7
2.15361	76.9	28
2.20181	76.9	28
2.22892	76.9	28
2.16867	76.9	28
2.09884	78.4	25
2.25000	78.4	25
2.04360	78.4	25
2.08721	78.4	25
2.07979	88.1	275
2.23404	88.1	275
2.07713	88.1	275
2.21011	88.1	275
2.01934	89.6	324
2.26796	89.6	324
2.10221	89.6	324
2.01105	89.6	324

of interest in this study was the relative wear rate, obtained by dividing the average wear rate (treadwear loss in miles per 1,000 miles of exposure) of a test tire by the average wear rate of the control tire.

Table 15.1 contains a portion of the data from this experiment for one tire brand, C. Data on the wear rate (y), temperature, and the number of miles traveled on wet pavement are displayed in the table. A least-squares fit

of y to x_1 (temperature/100), x_2 (miles/100), and $x_3 = x_1x_2$ resulted in the following fit:

$$\hat{y} = 5.29 - 4.04x_1 - 0.80x_2 + 1.06x_1x_2. \tag{15.8}$$

Consider now the interpretation of one of the estimated regression coefficients, the coefficient for temperature.

Figure 15.1 shows the relative wear plotted against the temperature residuals r^*. The temperature residuals were obtained by regressing temperature x_1 on a constant term and on the other two predictors, x_2 and $x_3 = x_1x_2$. The slope of the fitted line in Figure 15.1 is the least-squares estimate b_1 from the fit to all three predictor variables, -4.04 in Equation (15.8).

The magnitudes of least-squares estimates often are used as indicators of the importance of the predictor variables. Because predictor variables generally are measured on different scales, it is inappropriate to compare coefficient magnitudes directly. The concern about systematic variation among the predictors is another reason why regression coefficients should not be directly compared. In its extreme form, collinearities can render such comparisons completely meaningless — worse yet, erroneous.

If one is concerned that such comparisons suffer from possible misinterpretation due to systematic variation among the predictor variables, *standardized* coefficient estimates, also termed *beta-weight coefficients*, should be used. These standardized estimates, $\hat{\beta}_j$, remove the scaling problem and produce

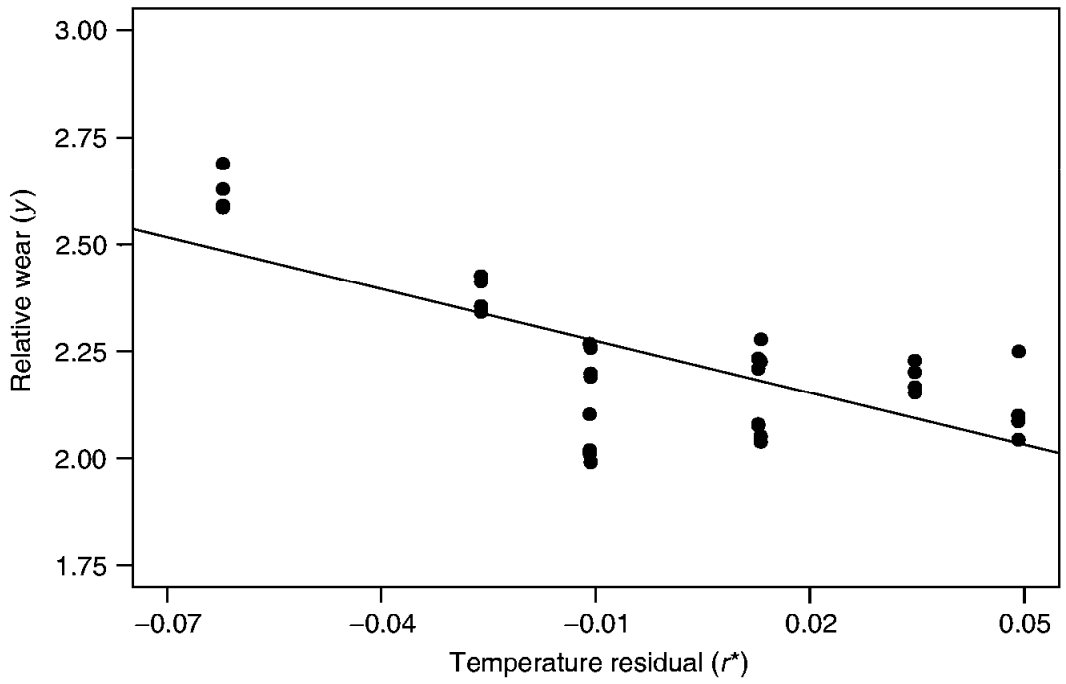


Figure 15.1 Relative wear versus temperature residuals.

dimensionless estimates that are directly comparable. They are defined as

$$\hat{\beta}_j = b_j (s_{jj}/s_{yy})^{1/2}, \tag{15.9}$$

where $s_{yy} = (n-1)^{-1} \sum (y_i - \bar{y})^2$ and $s_{jj} = (n-1)^{-1} \sum (x_{ij} - \bar{x}_j)^2$.

15.2 INFERENCE

Inferences on multiple linear regression models are needed to assess not only the overall fit of the prediction equation but also the contributions of individual predictor variables to the fit. In most of the inferential procedures discussed in this section we impose the model assumptions listed in Table 14.1, including the assumption that the error terms are normally distributed with a constant standard deviation. Investigating the validity of these assumptions for specific data sets is discussed in Chapter 18.

15.2.1 Analysis of Variance

Many different statistical measures are available for assessing the adequacy of the fitted model. While we stress the need to perform a comprehensive regression analysis (PISEAS) before claiming a fit is adequate, we examine below a few of the measures commonly reported with regression analyses. These generally are based on the use of the ANOVA table and the estimated regression coefficients.

The derivation of an ANOVA table for a multiple regression analysis is similar to the derivation for a single predictor variable in Chapter 14. The total and error sum of squares can again be expressed as

$$\text{TSS} = \sum (y_i - \bar{y})^2, \quad \text{SS}_E = \sum (y_i - \hat{y}_i)^2.$$

The regression sum of squares is

$$\text{SS}_R = \sum (\hat{y}_i - \bar{y})^2.$$

In multiple regression models there are p degrees of freedom for the sum of squares due to regression, because p coefficients, namely $\beta_1, \beta_2, \ldots, \beta_p$, must be estimated to obtain the regression sum of squares. One can show this

TABLE 15.2 Symbolic ANOVA Table for Multiple Regression Models

Source of Variation	df	Sum of Squares	Mean Squares	F-Statistic
Regression	p	SS_R	$MS_R = SS_R/p$	$F = MS_R/MS_E$
Error	$n - p - 1$	SS_E	$MS_E = SS_E/(n - p - 1)$	
Total	$n - 1$	TSS		

algebraically (see the exercises). A symbolic ANOVA table for a regression model having p predictor variables is shown in Table 15.2.

One useful measure of the adequacy of the fitted model is the estimated error standard deviation, $s_e = (MS_E)^{1/2}$, where $MS_E = SS_E/(n - p - 1)$. A small value for this statistic indicates that the predicted responses closely approximate the observed responses, while a large value may indicate either a large random error component or improper specification of the model.

The coefficient of determination, R^2, is another extensively used measure of the goodness of fit of a regression model. It can be defined in many similar ways, and because of this it is sometimes used inappropriately. Care should be exercised when using R^2 values to compare fits between (a) models with and without an intercept term, (b) models in which the response variable is not in exactly the same functional form, and (c) linear and nonlinear regression models. The references at the end of this chapter should be consulted regarding the appropriate calculation of R^2 in some of the alternatives to multiple linear regression models.

A preferred choice for calculating R^2 is as follows:

$$R^2 = 1 - \frac{SS_E}{TSS}. \tag{15.10}$$

For least-squares estimates of the model parameters the value of R^2 lies between 0 and 1; the closer it is to 1, the closer the predicted responses are to the observed responses. The coefficient of determination for models fitted by least squares can also be interpreted as the square of the ordinary correlation coefficient between the observed and the predicted responses [see Equation (14.13)].

The coefficient of determination is often adjusted to take account of the number of observations and the number of predictor variables. This adjustment is made because R^2 can be arbitrarily close to 1 if the number of predictor variables is too close to the number of observations. In fact, $R^2 = 1$ if $n = p + 1$ and no two responses have exactly the same set of predictor-variable values, regardless of whether there is any true relationship among the response

TABLE 15.3 ANOVA Table for Tire-Treadwear Data

Source of Variation	df	Sum of Squares	Mean Squares	F-Statistic	p-Value
		(a) *No Interaction*			
Regression	2	0.378	0.189	7.60	0.002
Error	29	0.720	0.025		
Total	31	1.098			
		(b) *Interaction Added*			
Regression	3	0.835	0.278	29.58	0.000
Error	28	0.263	0.009		
Total	31	1.098			

and the predictor variables, that is, regardless of whether equation (15.1) is the correct model. The adjusted R^2 is calculated from the following formula:

$$R_a^2 = 1 - a\frac{\text{SS}_E}{\text{TSS}}, \tag{15.11}$$

where $a = (n-1)/(n-p-1)$. This adjusted coefficient of determination is always less than Equation (15.10), but the difference between the two is usually minor if n is sufficiently large relative to p. When n is not much larger than p, the adjusted coefficient of determination should be used.

It is possible to find data sets where the R^2 value is near 1 but the estimated standard deviation is large. This usually is the result of model misspecification, although large error variability also might be a cause. Hence, caution should be used in relying on a single measure of the fit, such as R^2 values.

The ANOVA table for the tire-treadwear data in Table 15.1 using the fitted model without an interaction term, is shown in Table 15.3(a). The estimated model standard deviation is $s_e = 0.158$, and the coefficient of determination of the fitted model is $R^2(\times 100\%) = 34.4\%$, $R_a^2 = 29.9\%$. Addition of the interaction term as in equation (15.8) improves the fit dramatically [see Table 15.3(b)]: $s_e = 0.097$, $R^2(\times 100\%) = 76.0\%$, $R_a^2 = 73.4\%$. Note that the adjusted R_a^2 values do not differ appreciably from the unadjusted values, because the number of observations, $n = 32$, is sufficiently large relative to the number of predictors in the model, $p = 2$ or 3.

15.2.2 Lack of Fit

The ability to determine when a regression fit adequately models a response variable is greatly enhanced when repeat observations for several combinations

of the predictor variables are available. The error sum of squares can then be partitioned into a component due to *pure* error and a component due to *lack-of-fit* error:

$$SS_E = SSE_P + SSE_{LOF}. \tag{15.12}$$

The pure error sum of squares, SSE_P, is computed using the responses at the repeat observations:

$$SSE_P = \sum_{i=1}^{m} \left(\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\bullet})^2 \right). \tag{15.13}$$

In Equation (15.13), we have temporarily changed to a double subscript for clarity. It is only in the calculation of SSE_P that this is needed. We assume that there are m combinations of values of variables with repeat observations. For the ith combination, n_i repeats are available, the average of which is denoted by $\bar{y}_{i\bullet}$. The pure error sum of squares is identical to the numerator of the pooled variance discussed in Exhibit 3.4 and Equation (6.11) for the estimation of a common variance. The total number of degrees of freedom for this estimate of error is

$$f_P = \sum_{i=1}^{m} (n_i - 1) = q - m, \tag{15.14}$$

where $q = n_1 + n_2 + \cdots + n_m$. The lack-of-fit error sum of squares, SSE_{LOF}, is obtained as the difference between SS_E and SSE_P, with $f_{LOF} = (n - p - 1) - (q - m)$ degrees of freedom.

The test procedure for checking model specification is based on comparing the lack-of-fit error mean square with the pure error mean square. The symbolic

TABLE 15.4 Symbolic ANOVA Table for Lack-of-Fit Test

Source of Variation	df*	Sum of Squares	Mean Squares	F-Statistic
Regression	p	SS_R	MS_R	$F = MS_R/MS_E$
Error	$n - p - 1$	SS_E	MS_E	
Lack of fit	f_{LOF}	SSE_{LOF}	MSE_{LOF}	$F = MS_{LOF}/MSE_P$
Pure	f_P	SSE_P	MSE_P	
Total	$n - 1$	TSS		

* $f_P = q - m$, $f_{LOF} = n - p - 1 - f_P$.

ANOVA table in Table 15.2 is expanded in Table 15.4 to include this partitioning. This table includes the lack-of-fit F-statistic, $F = \text{MSE}_{\text{LOF}}/\text{MSE}_{\text{P}}$, to test the hypothesis that the model is correctly specified versus the alternative that the model is misspecified. The degrees of freedom for this F statistic are $\nu_1 = f_{\text{LOF}}$ and $\nu_2 = f_{\text{P}}$.

A statistically significant lack-of-fit F-statistic implies that the terms in the model do not capture all of the assignable-cause variation of the response variable. Because the pure error mean square measures the uncontrolled error variation only through calculations on repeat observations, a large F-ratio indicates that the lack-of-fit error mean square is measuring variation that exceeds that due to uncontrollable repeat-test error. When a large F-ratio occurs, one should consider alternative model specifications, including transformations of the response and predictors, additional polynomial terms for the predictors, etc.

The treadwear data in Table 15.1 have $m = 8$ values of temperature and wet miles for which there are 4 responses each. Hence, it is possible to test for lack of fit of the proposed model. The partitioned ANOVA table is given in Table 15.5 for a regression fit containing only the linear terms involving the two predictor variables [e.g., Table 15.3(a)]. The lack-of-fit F-statistic is statistically significant ($p < 0.001$). When the test for lack of fit is rerun after adding the interaction term, the F-statistic is much smaller, $F = 2.39(0.05 < p < 0.10)$.

An investigator, whenever possible, should plan to include repeat predictor-variable values in a regression data base. This will facilitate the use of the above lack-of-fit test. When it is not possible to collect repeat points, nearby points (that is, *nearest neighbors*) can be used to perform an approximate test for misspecification. It should be noted that model misspecification resulting from the exclusion of useful predictor variables may not be detected by the lack-of-fit test. Its sensitivity is directed mainly to reexpressions of current variables in the model.

TABLE 15.5 ANOVA Table for Lack of Fit of Treadwear Data

Source of Variation	df	Sum of Squares	Mean Squares	F-Value	p-Value
Regression	2	0.378	0.189	7.60	0.002
Error	29	0.720	0.025		
Lack of fit	5	0.532	0.106	13.56	0.000
Pure	24	0.188	0.008		
Total	31	1.098			

15.2.3 Tests on Parameters

The F-statistic in the ANOVA table in Table 15.2 allows one to simultaneously test that all $\beta_j = 0$ versus the alternative that at least one β_j is not zero. Specifically, to test

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad \text{vs} \quad H_a: \text{at least one} \quad \beta_j \neq 0,$$

use $F = \text{MS}_R/\text{MS}_E$ as the test statistic, with $\nu_1 = p$ and $\nu_2 = n - p - 1$ degrees of freedom. As in Chapter 14 we recommend that a large significance level, say $\alpha = 0.25$, be used in this test in order to reduce the chance of incorrectly deleting a useful predictor variable.

The F-statistic in the ANOVA table for the tire-treadwear data [Table 15.3(b)] is highly significant ($p < 0.001$). This indicates that at least one of the three predictor variables (temperature, wet miles, and their interaction) is useful in predicting the relative wear rate of this brand of tire.

Testing hypotheses on individual regression coefficients often is of primary interest to an experimenter performing a regression analysis. As with the single-variable model in Chapter 14, a t-statistic can be constructed for testing

$$H_0: \beta_j = c \quad \text{vs} \quad H_a: \beta_j \neq c$$

for some specified constant c using the general procedures given in Section 6.3. The test statistic used for this purpose is

$$t = \frac{b_j - c}{s_e c_{jj}^{1/2}}, \tag{15.15}$$

where $s_e = (\text{MS}_E)^{1/2}$ is the estimated error standard deviation and

$$c_{jj} = [(n-1)s_j^2(1 - R_j^2)]^{-1}. \tag{15.16}$$

In Equation (15.16), s_j^2 is the sample variance of the n values of the jth predictor variable and R_j^2 is the coefficient of determination for the regression of x_j on the constant term and the $p - 1$ other predictor variables.

The t-statistics (15.15) with $c = 0$ are used to test the statistical significance of the individual model parameters; that is, the usefulness of x_j as a predictor of the response variable. Note that because β_j is a partial regression coefficient and both b_j and c_{jj} are functions of the values of the other predictor variables, this test determines the importance of the jth predictor variable only *conditionally*, that is, conditioned on the other predictor variables being in the model. Thus, it can be considered a conditional test and should not be interpreted as a determinant of the significance of the jth predictor variable without regard to the presence or absence of the other predictor variables.

The above approach also can be used to test the significance of the intercept term. To test $H_0: \beta_0 = c$ vs $H_a: \beta_0 \neq c$ we use a t-statistic similar in form to Equation (15.15). The appendix to this chapter outlines the algebraic details for the formation of this t-statistic. Most regression computer programs provide the t-statistic for testing whether the intercept or the individual regression coefficients are zero.

Table 15.6 contains additional information on the regression of treadwear on temperature, wet miles, and their interaction. The intercept term is significantly different from zero, and each individual predictor variable contributes significantly to the fits, given that the other two predictor variables are also included in the model.

The contention that t-tests on individual parameters are conditional tests can be appreciated by relating the t-statistic to an equivalent F-statistic derived using the principle of reduction in sums of squares, described in Section 8.1.

Consider the full regression model in Equation (15.1) and a reduced model containing any subset of $k < p$ predictor variables. Denote the full-model (M_1) error sum of squares by SSE_1 and the reduced-model (M_2) error sum of squares by SSE_2. Denote the reduction in error sum of squares resulting from the fit of the additional terms in the full model by

$$R(M_1|M_2) = \text{SSE}_2 - \text{SSE}_1. \tag{15.17}$$

There are $p - k$ more predictor variables in the full model than the reduced one. Therefore, the F-statistic for determining the statistical significance of this subset of predictor variables is

$$F = \frac{\text{MSR}(M_1|M_2)}{\text{MSE}_1}, \tag{15.18}$$

where $\text{MSR}(M_1|M_2) = R(M_1|M_2)/(p-k)$. Under the null hypothesis that the $p - k$ additional predictor variables in the full model have regression

TABLE 15.6 Summary Statistics for Tire-Treadwear Regression

Variable	Coefficient Estimate	t Statistic*	95% Confidence Interval
Intercept	5.288	13.94	(4.512, 6.064)
Temperature†	−4.044	−7.79	(−5.102, −2.982)
Wet miles‡	−0.801	−7.52	(−1.018, −0.583)
Temperature × wet miles	1.061	6.97	(0.539, 1.372)

*All are statistically significant ($p < 0.001$).
†Fahrenheit temperature/100.
‡Distance in 10^{-2}mi.

coefficients equal to zero, this statistic has an F-distribution with $\nu_1 = p - k$ and $\nu_2 = n - p - 1$ degrees of freedom. This test is called a "partial F-test" in many textbooks, since it measures the contribution of a subset of predictor variables conditioned on other predictor variables being in the model. If $k = p - 1$, the F-statistic (15.18) is the square of the t-statistic from the full model corresponding to the term left out of the reduced model.

To apply this principle reconsider the two models for the tire-treadwear data, that is, the fits with and without the interaction term. Using the two error sums of squares from Table 15.3(a) and (b), the reduction in error sum of squares is

$$R(M_1|M_2) = 0.720 - 0.263 = 0.457,$$

and the corresponding partial F-statistic is

$$F = \frac{0.457/1}{0.263/28} = 48.65.$$

Because this statistic is highly significant ($p < 0.001$), the temperature-by-wet-miles interaction is a statistically significant predictor of wear rate in addition to the contributions of the individual linear terms for the two variables. The equivalence of this procedure and the use of the t-statistic to test the significance of the interaction term is readily established, since $t = F^{1/2} = (48.65)^{1/2} = 6.97$ is the value of the interaction t-statistic in Table 15.6.

15.2.4 Confidence Intervals

Confidence intervals for the regression coefficients in Equation (15.1) can be constructed using the same type of procedures discussed in Sections 14.5 and 2.4. A $100(1 - \alpha)\%$ confidence interval for β_j is given by

$$b_j - t_{\alpha/2} s_e c_{jj}^{1/2} \leqslant \beta_j \leqslant b_j + t_{\alpha/2} s_e c_{jj}^{1/2}, \tag{15.19}$$

where $t_{\alpha/2}$ is a two-tailed $100\alpha\%$ t critical value having $n - p - 1$ degrees of freedom. A $100(1 - \alpha)\%$ confidence interval for β_0 can be defined in a similar fashion (see the appendix to this chapter). 95% confidence intervals for the coefficients in the tire treadwear data are shown in the last column of Table 15.6.

Simultaneous confidence intervals for all the coefficients in equation (15.1) cannot be computed using the individual coefficient intervals for β_0 and the β_j. The individual intervals are useful for estimating ranges for the individual coefficients, but they ignore the systematic variation of the predictor variables and consequent correlation among the coefficient estimators. Even if the estimated coefficients were statistically independent, the construction of several

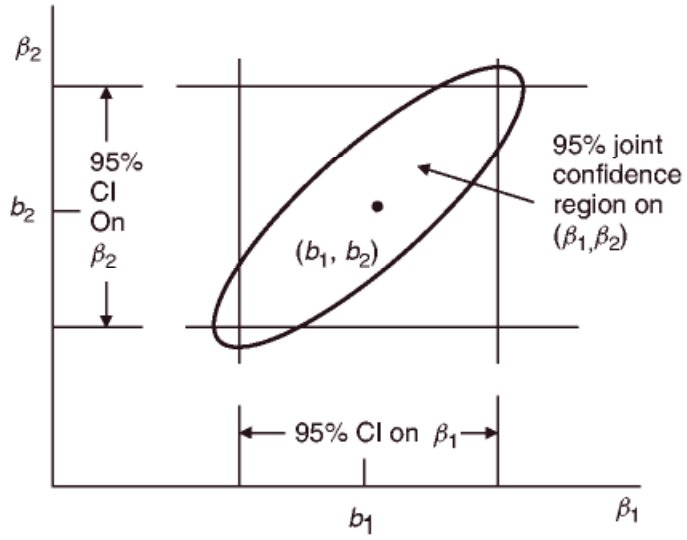


Figure 15.2 Schematic comparison of a typical (rectangular) confidence region formed from individual 95% confidence intervals (CI) with the true joint (elliptical) 95% confidence region.

individual confidence intervals would not result in an overall confidence region with the stated confidence. The reasons for this are similar to those discussed in Section 6.4 and are related to the distinction between experimentwise and comparisonwise error rates. Thus, the chance that one particular interval covers the corresponding regression coefficient is as stated, $1 - \alpha$, but the probability that all the intervals simultaneously cover all the regression coefficients can be much less than $1 - \alpha$.

Figure 15.2 illustrates the difference between joint and individual confidence statements. The individual $100(1 - \alpha)\%$ interval for β_1 and β_2 create a rectangular region, while the simultaneous interval is elliptical. The stronger the systematic variation between the two involved predictor variables, the narrower will be the joint confidence ellipsoid. The algebraic details of the construction of simultaneous confidence regions are outlied in the appendix to this chapter.

15.3 INTERACTIONS AMONG QUANTITATIVE PREDICTOR VARIABLES

In Chapter 5, interactions were defined as joint factor effects in an ANOVA model. Interactions have a similar interpretation in terms of the joint effects of predictor variables in regression models. Most often interaction terms are formed as products of two or more predictor variables, although they can be specified in any manner felt to be reasonable for the model under investigation.

To illustrate geometrically the effects of adding interaction terms among quantitative predictors to a regression model, consider again the tire treadwear data in Table 15.1. In the last two sections, regression fits with and without an interaction term were discussed. In Section 15.2.2, significant lack of fit results from the exclusion of an interaction term between temperature and wet miles, but the lack of fit test was not strongly significant ($0.05 < p < 0.10$) when the interaction term was added to the fit. We now wish to demonstrate geometrically the effect of adding an interaction term. The interpretation parallels, in a continuous manner, the interpretation that was given in Section 5.2 for interactions in models for designed experiments involving factor levels.

The least-squares fit of treadwear (y) to temperature (x_1 = temperature/100) and the number of miles traveled on wet pavement (x_2 = miles/100) is

$$\hat{y} = 2.75 - 0.55x_1 - 0.06x_2. \tag{15.20}$$

The graph of this fitted model is a plane in a three-dimensional data space, Figure 15.3.

There is no interaction term between temperature and wet miles in Equation (15.20). Graphically one can see from Figure 15.3 that a change in temperature (x_1) for a fixed value of wet miles (x_2) produces the same change in the relative wear rate regardless of the value of wet miles used (and vice versa). Thus, the fitted model is a plane in three dimensions. The contour

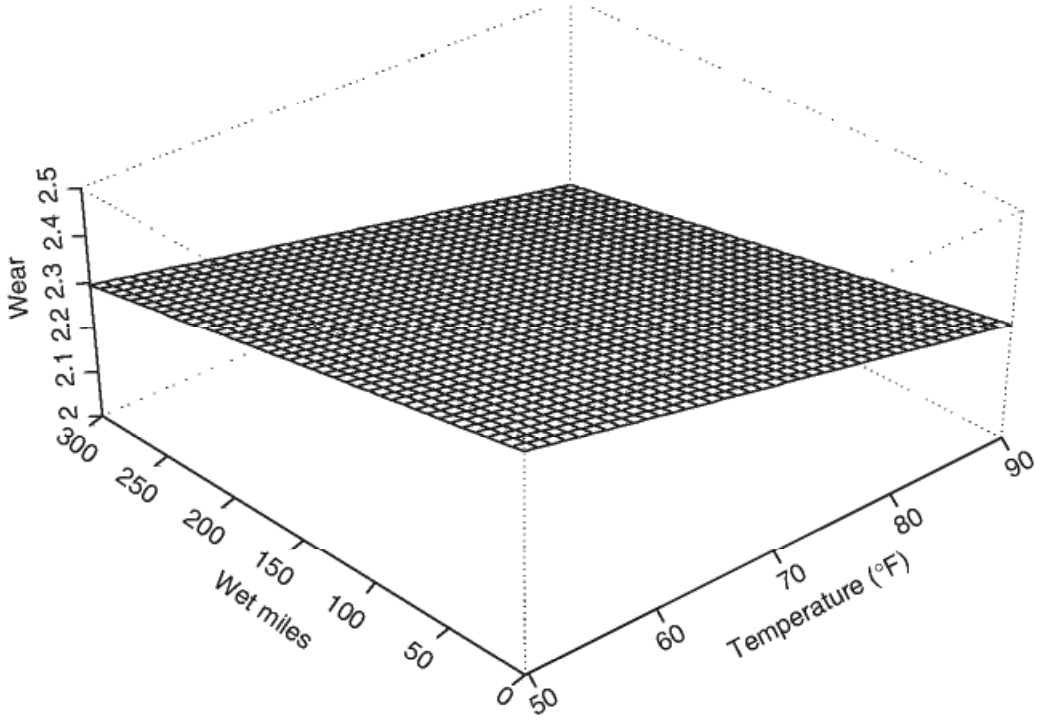


Figure 15.3 Tire-treadwear fit without interaction.

lines (curves representing constant values of the fitted response variable) of the surface of the plane are parallel for fixed values of x_1 as well as for fixed values of x_2. Models of this type are called *first-order* models, because the exponent of each x-term is one.

The response surface corresponding to Equation (15.8) is plotted in Figure 15.4. Temperature and wet miles interact in Equation (15.8) because of the presence of the product term. The change in the relative wear rate corresponding to a specified change in temperature depends on how many wet miles the vehicle is driven. Consequently, the contour lines on the surface of Figure 15.4 are not parallel. Models of this type are a special type of *second-order* regression models, because the sum of the exponents of the highest-order term (x_1x_2) is two. A complete second-order model would contain linear, cross-product, and pure quadratic (x_1^2 and x_2^2) terms.

Rearranging the terms in the fit (15.8) allows a direct parallel to be established between the interpretation of interaction terms for quantitative predictors and that used for factors in Section 5.2. Rewrite Equation (15.8) as

$$\hat{y} = 5.29 - (4.04 - 1.06x_2)x_1 - 0.80x_2 \tag{15.21}$$

or, equivalently, as

$$\hat{y} = 5.29 - 4.04x_1 - (0.80 - 1.06x_1)x_2 \tag{15.22}$$

In Equation (15.21), the effect of temperature on wear (that is, the slope of temperature) depends on the number of wet miles traveled. Similarly, in

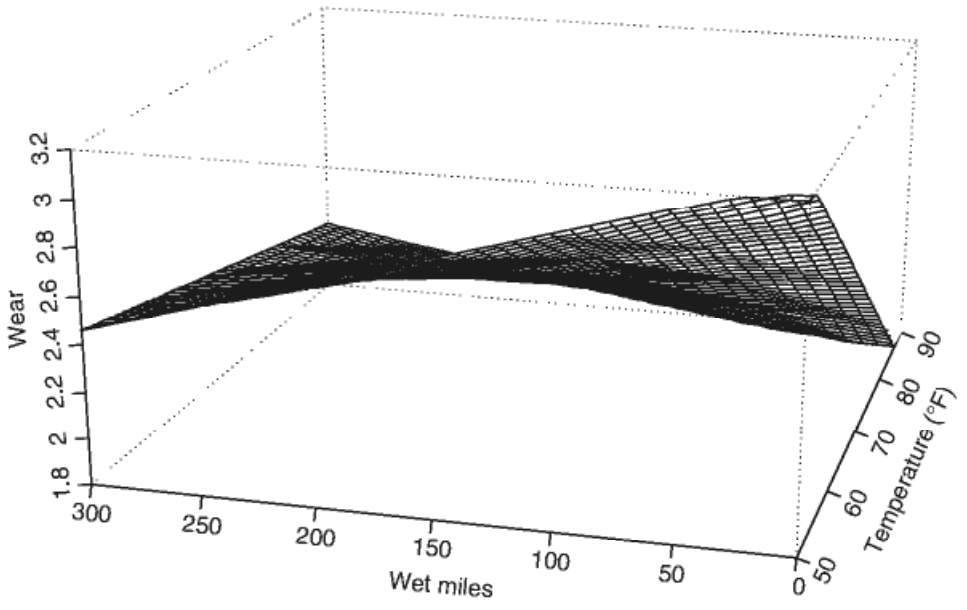


Figure 15.4 Tire-treadwear fit with interaction.

Equation (15.22), the effect of wet miles on wear depends on the temperature. In both forms of the fit (15.8), the effect of one quantitative predictor on the response depends on the value of the other predictor. This is precisely the definition of an interaction that was given for factors in Exhibit 5.3. The twist apparent in Figure 15.4 is a rotation of straight lines, as indicated in equations (15.21) and (15.22).

Scatterplots are often useful in determining whether an interaction term might be beneficial in a regression model. If one of the variables is categorical, a labeled scatterplot of y versus the quantitative variable might show different slopes for some values of the categorical variable. If both predictors are quantitative, a plot of the residuals from a fitted model without the interaction term [e.g., Equation (15.20)] versus the product x_1x_2 might show a strong linear trend. If so, an interaction term in Equation (15.8) would be appropriate. If the plot is nonlinear, an interaction component to the model is still suggested, but it might be a nonlinear function of the product.

Ordinarily, one should not routinely insert products of all the predictors in a regression model. To do so might create unnecessary complications in the analysis and interpretation of the fitted models due to collinear predictor variables (Section 15.4). The purpose of including interaction terms in regression models is to improve the fit either because theoretical considerations require a modeling of the joint effects or because an analysis of regression data indicates that joint effects are needed in addition to the linear terms of the individual variables.

15.4 POLYNOMIAL MODEL FITS

A necessary requirement for regression modeling and a key component of the PISEAS comprehensive regression analysis approach (Chapter 14) is the specification of the functional relationship between the response and the predictor variables. If this relationship is known to the experimenter on the basis of theoretical arguments or previous empirical results, it should be used; however, frequently the model is not known prior to the data analysis. In many important problems in engineering and science the underlying mechanism that generates the data is not well understood, due to the complexity of the problem and to lack of sufficient theory. In these cases polynomial models often can provide adequate approximations to the unknown functional relationship.

The polynomial models discussed in this chapter are defined in Exhibit 15.3. Many theoretical models are polynomial. For example, the physical law describing the volume expansion of a rectangular solid is a third-order polynomial in temperature t given by

$$V = L_0W_0H_0(1 + 3\beta t + 3\beta^2t^2 + \beta^3t^3),$$

where L_0, W_0, and H_0 are the dimensions of the solid at 0°C and β is the coefficient of linear expansion. Another example is the second-order relationship between temperature T, pressure P, and volume V in ideal gas laws:

$$RT = PV,$$

where R is a constant that depends on the type of gas.

EXHIBIT 15.3 POLYNOMIAL MODELS

- A polynomial model for p predictor variables has the form

$$y = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \cdots + \beta_m t_m + e, \tag{15.23}$$

  where the jth variable, t_j, is either a single predictor variable or a product of at least two of the predictors; each variable in t_j can be raised to a (positive) power.
- The order of a polynomial model is determined by the maximum (over the m terms) of the sum of the powers of the predictor variables in each term of the model.

  (a) A first-order model is of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e. \tag{15.24}$$

  (b) A complete second-order model (also called a quadratic model) is of the form

$$y = \beta_0 + \sum_i \beta_i x_i + \sum\sum_{i<j} \beta_{ij} x_i x_j + \sum_i \beta_{ii} x_i^2 + e. \tag{15.25}$$

  Note that for a first-order model $m = p$ and for a second-order model $m = p(p+3)/2$.

Our primary concern in this section is the use of polynomial models to approximate unknown relationships between responses and predictor variables. A motivation for this is contained in calculus-based formulations of Maclaurin and Taylor series expansions of functions: any suitably well-behaved function of a mathematical variable x can be written as an infinite sum of terms involving increasing powers of x. A Taylor series expansion often is used to approximate a complicated function by expressing it as a polynomial, perhaps containing an infinite number of terms, and retaining only a few low-order terms of the series. The number of terms necessary to give an adequate approximation depends on the complexity of the function, the range of interest of x, and the use of the approximation.

The polynomial models considered in this chapter exhibit the same functional form in the predictor variables as the corresponding Taylor series approximation. The unknown coefficients of the powers of the predictor variables [β_j, $j = 1, 2, \ldots, m$ in Equation (15.23)] are estimated as described in Section 15.1.

In many respects a polynomial model in several predictor variables can be viewed as a multidimensional french curve. A french curve is a drawing instrument that is used to draw a smooth curve through points that do not fall on such regular curves as an ellipse or a circle. In the same way, polynomial models are used to smooth response data over a region of the predictor variables. As higher-order terms are included in the model, the "french curve" acquires more twists and bends. An important operating principle for determining the order of the model to fit to experimental data (that is, how elaborate a french curve should be used) is that of parsimony.

One should start with the simplest model warranted by what is known about the physical mechanism under study and by suggestions obtained from plots of the response variable versus the predictor variables. If a lack-of-fit test (Section 15.2.2) or a residual analysis (Section 18.2) indicates that a proposed model is an inadequate approximation to the observed responses, one can either add the next higher-order terms into the model or investigate nonlinear models. In many experimental situations, a first- or second-order polynomial is adequate to describe a response.

The data shown in Table 15.7 are 37 observations from a synthetic-rubber process. The data were collected to investigate the relationship between the

TABLE 15.7 Synthetic-Rubber Process Data: Production Rates and Solvent Weights

Rate (lb/hr)	Solvent Weight (%)	Rate (lb/hr)	Solvent Weight (%)	Rate (lb/hr)	Solvent Weight (%)
400	0.02	450	0.03	460	0.06
480	0.01	480	0.03	640	0.03
760	0.06	890	0.09	1020	0.16
1090	0.16	1100	0.12	1100	0.10
1120	0.14	1140	0.12	1140	0.14
1150	0.11	1140	0.13	1150	0.13
1190	0.19	1220	0.17	1280	0.12
1290	0.12	1290	0.15	1290	0.19
1320	0.13	1320	0.16	1320	0.21
1330	0.25	1340	0.13	1350	0.20
1380	0.16	1380	0.23	1410	0.30
1420	0.43	1610	0.41	1660	0.40
1750	0.26				

weight (in percent) of a solvent and the corresponding production rate of the rubber process. Figure 15.5 shows a plot of the solvent content versus the production rate. The plot indicates a strong linear component in the relationship, with an indication of possible curvature. A linear regression model was fit to the data. The ANOVA table is presented in Table 15.8. Note that there is evidence of lack of fit with the straight-line model. A quadratic model provides an improved fit, as will be demonstrated below.

Although polynomial models are useful tools for regression analysis, their use has some potential drawbacks. One such drawback is that to model some response variables satisfactorily, a complicated polynomial model may be needed, when simpler nonlinear models (Chapter 18) or a linear model

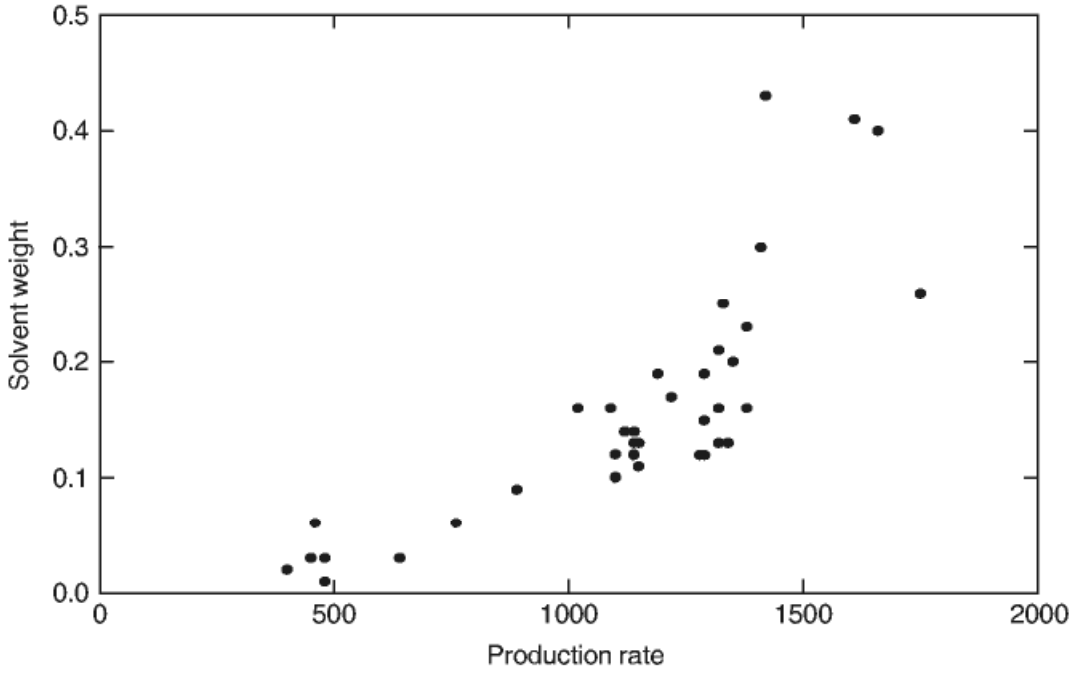


Figure 15.5 Scatterplot of solvent weights vs. production rates for a synthetic-rubber process.

TABLE 15.8 ANOVA Table for the Synthetic-Rubber Process Data

Source	df	SS	MS	F	p-Value
Regression	1	0.2433	0.2433	64.0	0.000
Error	35	0.1329	0.0038		
Lack of fit	25	0.1239	0.0050	5.52	0.004
Pure	10	0.0090	0.0009		
Total	36	0.3762			

with one or more predictors reexpressed may be equally satisfactory. This is especially likely when polynomial models are used routinely without proper thought about the physical nature of the system being studied.

The relationship between volume V and pressure P in the expansion of gases at a constant temperature is known to follow Boyle's law, derivable from the kinetic theory of gases:

$$PV = k, \qquad \text{or} \quad V = kP^{-1},$$

where k depends primarily on the mass of the gas and on the constant temperature. If one were to attempt to fit the volume with a polynomial (positive powers) in pressure, at least a quadratic or cubic polynomial would be needed to fit the data adequately over a reasonable range of pressure values. A simpler model is the inverse relationship shown above.

A second risk in fitting polynomial models is the temptation to use them to extrapolate outside the experimental region. Remembering the french-curve analogy, the fitted polynomial does no more than describe the response in a smooth manner over the region where data were obtained. Like a french curve, polynomial models can be used to extrapolate past the range of the data, but the behavior of the actual system may be entirely different.

Another peril of polynomial models is collinear predictor variables. In mathematics courses, notably linear algebra, two lines are said to be collinear if they have the same intercept and the same slope. Three lines are said to be coplanar if they all lie in a two-dimensional plane. Four or more lines are said to be coplanar if they reside in a subspace of dimension at most one less than the number of lines ("coplanar" is used even though the subspace may be of dimension greater than two). In each of these instances the lines are said to be linearly dependent. Rather than distinguish collinear from coplanar situations, we refer to all of these situations as being collinear.

In regression analyses, collinear predictor variables supply redundant information (see Exhibit 15.4). Frequently variables are not exactly redundant but are extremely close to being redundant. The closer the linear dependence among two or more of the predictor variables is to being exact, the stronger the redundancy among the variables.

EXHIBIT 15.4

Collinear Predictor Variables. Two or more predictor variables are collinear if there exists an approximate linear dependence among them.

The collinearity problem is not confined to polynomial models, but they can display it in special ways. A common way in which collinearities arise with polynomial models is through the failure to standardize the predictor

variables before forming products and powers (see Exhibit 15.5). The form of standardization given in Exhibit 15.5 is sometimes called the *normal-deviate* standardization because of its similarity to the formation of standard normal variables. An equivalent form of standardization is

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{d_j}, \tag{15.26}$$

where $d_j = (n-1)^{1/2} s_j$. This form of standardization is usually called *correlation-form* standardization, because

$$\sum_{i=1}^{n} z_{ij} z_{ik} = r_{jk},$$

the correlation coefficient for the observations on x_j and x_k.

EXHIBIT 15.5

Standardized Predictor Variable. A standardized predictor variable, denoted z_j, is obtained from the corresponding raw predictor variable x_j by subtracting the average $\bar{x}_j$ from each observation and then dividing the result by the standard deviation s_j:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}. \tag{15.27}$$

These forms of standardization result in the standardized variables having an average of zero and standard deviations of 1 or $(n-1)^{1/2}$, respectively. The normal-deviate standardization scales predictor variables so they have a range of approximately -2.5 to $+2.5$, regardless of the original units. One major benefit of standardization is that collinearity problems may be lessened, perhaps even eliminated, simply by standardizing prior to forming products and powers of the predictor variables. A second benefit is that inaccuracies in computations due to roundoff error are lessened when computing the regression coefficients.

Two other benefits are derived from standardizing the predictors. First, the magnitude of the regression coefficients are comparable and can be used to assess the relative effects of the predictor variables on the response. If the response and the predictor variables are all standardized, the resulting coefficient estimates are the beta weights of Equation (15.9). Second, the constant term has a meaningful interpretation when fitting response surfaces. It is the estimated value of the response at the centroid (center) of the experimental region.

A straightforward example provides a quantitative illustration of the collinearity problems that can arise with polynomial models and the potential benefits of standardization. Suppose one wishes to model a response variable as a polynomial function of a single predictor variable. Suppose further that the response variable is observed at values of the predictor variable from 1 to 5 in increments of 0.1:

$$x_i = 1 + 0.1(i - 1), \qquad i = 1, 2, \ldots, 41.$$

If this predictor is not standardized, the correlations between pairs of the linear, quadratic, and cubic powers of x are extremely highly correlated as shown in Table 15.9.

The average of the 41 values of x is 3.0, and the standard deviation is 1.198. Standardized values of the predictor variable are then obtained from the following equations:

$$z_i^k = \left(\frac{x_i - 3.0}{1.198}\right)^k, \qquad k = 1, 2, 3.$$

Correlations between pairs of these standardized powers are also shown in Table 15.9. Observe that the correlations between the linear and quadratic and between the quadratic and cubic powers are zero. Very high correlation remains between the linear and cubic terms.

Because many response-surface models are at most quadratic functions of several variables, this illustration demonstrates the potential for reduction, even elimination, of collinearity problems with such polynomial fits. The illustration also indicates that collinearity problems are not always eliminated by standardization. The benefits of standardization are dependent on the polynomial terms used in the model and the range of values of the predictors. Note, however, that one can always calculate correlation coefficients as one indicator of whether collinearities may remain a serious problem after standardization.

TABLE 15.9 Correlation Coefficients for Powers of x Equally Spaced in Increments of 0.1 from 1 to 5

Raw Predictors		Standardized Predictors	
Variables	Correlation	Variables	Correlation
x, x^2	0.985	z, z^2	0.000
x, x^3	0.951	z, z^3	0.917
x^2, x^3	0.990	z^2, z^3	0.000

Table 15.10 contains two fits to the solvent data listed in Table 15.7. Table 15.10(a) summarizes a quadratic fit when the production rate is not standardized, and Table 15.10(b) provides a similar summary for the standardized production rate. The regression equations produce the same ANOVA table (see Table 15.11) and identical predicted values. For these reasons the fitted models are considered to be equivalent.

Interpreting the coefficients in Table 15.10(a) is troublesome. The estimated regression coefficient for the linear term is negative and not statistically significant; yet the trend in Figure 15.5 is clearly increasing in the production rate. On standardizing the production rate prior to forming the linear and quadratic terms, the interpretation is more reasonable. Both the linear and the quadratic coefficient estimates are positive and statistically significant. Although the coefficients for the nonstandardized fit are obtainable from the standardized coefficient estimates (see the exercises), the reasonable interpretation of the standardized estimates render this fit more appealing.

TABLE 15.10 Summary of a Quadratic Fit to the Synthetic-Rubber Data

Model Term	Coefficient Estimate	t-Value
(a) *Raw Rate Values*		
Intercept	0.0558	0.82
Linear (rate)	-0.0147×10^{-2}	−1.05
Quadratic (rate^2)	0.0193×10^{-5}	2.80
(b) *Standardized Rate Values*		
Intercept	0.1358	11.13
Linear (rate)	0.0995	8.85
Quadratic (rate^2)	0.0229	2.80

TABLE 15.11 ANOVA Table for Quadratic Fit to Synthetic-Rubber Data

Source	df	SS	MS	F	p-Value
Regression	2	0.2682	0.1341	42.2	0.000
Error	34	0.1080	0.0032		
Lack of Fit	24	0.0990	0.0041	4.59	0.008
Pure	10	0.0090	0.0009		
Total	36	0.3762			

The differences in the coefficient estimates for the nonstandardized and standardized fits can be explained by referring to the interpretation of least-squares coefficient estimates in Section 15.1. The linear-coefficient estimate can be obtained by regressing the solvent weights y on the residuals r^* from a least-squares fit of the production-rate values x on the squares of these values x^2; that is, we regress y on $r^* = x - \hat{x}$, where $\hat{x} = a_1 + a_2 x^2$ and a_1 and a_2 are least-squares estimates from the regression of x on x^2. The high correlation between x and $x^2 (r = 0.98)$ results in all the residuals r^* being close to zero. The regression of y on r^* thus yields a nonsignificant estimated coefficient, which happens to be negative.

Standardizing the predictor variable prior to forming the quadratic term produces standardized linear and quadratic variables that are less correlated $(r = -0.55)$. Consequently the regression of y on r^* is better able to indicate the effect of the linear term of the model. This is why the estimated coefficient is positive and statistically significant.

Significant lack of fit (see Table 15.11) is detected in the ANOVA table for both of the quadratic fits shown in Table 15.10. A reasonable question to ask is whether a cubic term will help better describe the relationship between solvent content and production rate. The answer is no. This lack of fit is caused by poorly fitting the observations corresponding to high solvent contents, in particular observations 34 and 37. The poor fitting of these two points is symptomatic of the real problem with this fit: the increasing variability of the solvent weights with the magnitude of the weights. A transformation of the solvent weights that satisfactorily accounts for both its apparent nonlinear relationship with the production rate and its nonconstant variability is presented in Section 18.3.

APPENDIX: MATRIX FORM OF LEAST-SQUARES ESTIMATORS

The multiple linear regression model (15.1) can be represented in matrix form as follows:

$$\mathbf{y} = X_c \boldsymbol{\beta}_c + \boldsymbol{e}, \tag{15.A.1}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix},$$

$$X_c = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{i1} & x_{i2} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix},$$

$$\boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{pmatrix},$$

and $\boldsymbol{\beta}_c = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$. We use the notation $\mathbf{z}'$ to denote the transpose of a vector (similarly for a matrix).

The subscript c used with the X-matrix and the vector of regression coefficients is intended to stress that the constant term is included in the model; consequently, β_0 appears in the coefficient vector $\boldsymbol{\beta}_c$, and the first column of X_c is a column of ones. In subsequent chapters properties of the least-squares estimators of the coefficients of the nonconstant predictor variables $x_1, x_2, \ldots, x_p$ will be discussed. When such discussions occur, of interest will be the X-matrix without a column of ones and the $\boldsymbol{\beta}$-vector without the constant β_0. These quantities will be denoted by dropping the subscript c.

Using the model (15.A.1), the general expression for the normal equations in matrix form is

$$X_c'X_c\mathbf{b}_c = X_c'\mathbf{y}. \tag{15.A.2}$$

The solutions of the normal equations, the least-squares estimators for multiple linear regression models, are obtained by multiplying both sides of equation (15.A.2) by the matrix inverse of $X_c'X_c$, denoted $(X_c'X_c)^{-1}$:

$$\mathbf{b}_c = (X_c'X_c)^{-1}X_c'\mathbf{y}. \tag{15.A.3}$$

The elements of the vector $X_c'\mathbf{y}$ are the sums of the cross-products of the elements of the columns of X and the observations in $\mathbf{y}$. The diagonal elements of $X_c'X_c$ are the sums of the squares of the elements in the columns of X_c, and the off-diagonal elements are the sums of the cross-products of the elements in the columns of X_c.

The t-statistic and confidence intervals for the intercept parameter and the regression coefficients are based on the least-squares estimators of

$\boldsymbol{\beta}_c$, equation (15.A.3), and the matrix $C = (X_c'X_c)^{-1}$. Denote the diagonal elements of this matrix by $c_{00}, c_{11}, \ldots, c_{pp}$. The c_{jj}-values used in Equations (15.15) and (15.19) are the last p diagonal elements of C. To test hypotheses and form confidence intervals on β_0, these same equations can be used when b_0 is substituted for b_j and when c_{00}, the first diagonal element of C, is inserted for c_{jj}.

The formula for the $100(1-\alpha)\%$ confidence region for $\boldsymbol{\beta}_c$ is

$$\frac{(\mathbf{b}_c - \boldsymbol{\beta}_c)'X_c'X_c(\mathbf{b}_c - \boldsymbol{\beta}_c)}{(p+1)\text{MS}_E} < F_\alpha, \tag{15.A.4}$$

where F_α is the upper-tail $100\alpha\%$ point of an F-distribution with $\nu_1 = p+1$ and $\nu_2 = n-p-1$ degrees of freedom. The references at the end of this chapter include detailed discussions on these types of intervals and how to construct them. They also discuss the construction of simultaneous confidence regions for the regression coefficients in $\boldsymbol{\beta}$, that is, excluding the constant term.

Confidence intervals for mean responses [that is, Equation (15.3)] and prediction intervals for future responses can also be constructed for fixed values of the predictor variables. A $100(1-\alpha)\%$ confidence interval for the mean response μ, at a fixed point $\mathbf{u}' = (1, u_1, \ldots, u_p)$, where u_j is the value of the jth predictor variable for which the mean response is desired, is

$$\hat{y} \pm t_{\alpha/2}\{\text{MS}_E[\mathbf{u}'(X_c'X_c)^{-1}\mathbf{u}]\}^{1/2}, \tag{15.A.5}$$

where $\hat{y} = \mathbf{b}_c'\mathbf{u} = b_0 + \sum b_j u_j$. To construct a $100(1-\alpha)\%$ prediction interval for a future response y_f, we modify Equation (15.A.5) as follows:

$$\hat{y} \pm t_{\alpha/2}\{\text{MS}_E[1 + \mathbf{u}'(X_c'X_c)^{-1}\mathbf{u}]\}^{1/2}, \tag{15.A.6}$$

where, again, $\hat{y} = \mathbf{b}_c'\mathbf{u}$.

Simultaneous confidence and prediction intervals corresponding to the single intervals (15.A.5) and (15.A.6) also can be constructed. These have forms similar to the joint region defined in equation (15.A.4) for the coefficient parameters. The references at the end of this chapter include more details on these procedures.

Note that the uncertainty limits in the confidence intervals (15.A.5) and (15.A.6) depend on the value of $\mathbf{u}$ through the term $\mathbf{u}'(X_c'X_c)^{-1}\mathbf{u}$. While it is not obvious, the limits are smallest when $\mathbf{u}' = (1, \bar{x}_{\bullet 1}, \ldots, \bar{x}_{\bullet p})$, where $\bar{x}_{\bullet j}$ is the average of the predictor values for x_j. The magnitude of $\mathbf{u}'(X_c'X_c)^{-1}\mathbf{u}$ increases as $\mathbf{u}'$ deviates from $(1, \bar{x}_{\bullet 1}, \ldots, \bar{x}_{\bullet p})$.

REFERENCES

Text References

The texts referenced at the end of Chapter 14 also provide coverage of the topics discussed in this chapter. These texts rely on matrix algebra in the discussions of multiple linear regression.

Useful texts for simultaneous confidence and prediction limits include those by Draper and Smith and by Montgomery and Peck (also referenced in Chapter 14). An advanced text that includes a chapter on this topic is

Seber, G. A. F. (1977). *Linear Regression Analysis*, New York: John Wiley & Sons, Inc.

A discussion of various definitions of the coefficient of determination, including recommendations for which ones to use, is found in the following article:

Kvalseth, T. O. (1985). "Cautionary Note About R^2," *The American Statistician*, **39**, 279–285.

Data References

The treadwear data were taken from the following report:

R. N. Pierce, R. L. Mason, K. E. Hudson, and H. E. Staph (1985). "An Investigation of a Low-Variability Tire Treadwear Test Procedure and of Treadwear Adjustment for Ambient Temperature," Report No. SwRI EFL-7928-1, Southwest Research Institute, San Antonio, TX.

The data in Exercise 2 are extracted from

Hare, C. T. (1977). "Light Duty Diesel Emission Correction Factors for Ambient Conditions," Final Report to the Environmental Protection Agency under Contract No. 68-02-1777, Southwest Research Institute, San Antonio, TX.

The data in Exercise 3 were provided by Professor Ladislav P. Novak, Department of Anthropology, Southern Methodist University.

The data in exercise 9 and 11, respectively, are taken from

Montemayor, A. F., Owens, E. C., Buckingham, J. P., Jung, P. K., and Giannini, R. M. (1985). "Fuel Property Effects on the Cold Startability of Navy High-Speed Diesel Engines," Interim Report, U.S. Army Contract No. DAAK70-85-C-0007.

Montemayor, A. F. and Owens, E. C. (1985). "Fuel Property Effects on the Unaided Cold Starting of a Two-Cycle Diesel Engine," Interim Report, U.S. Army Contract No. DAAK70-85-C-0007.

EXERCISES

1 Algebraically show that the least-squares coefficient estimators minimize the error sum of squares in equation (15.6).

2 The following data are taken from an experiment designed to investigate the effects of three environmental variables on exhaust emissions of light-duty diesel trucks. Investigate regression fits to the nitrogen oxides (NO_x) data listed below. Do any of your fits to the data appear to be satisfactory? The researchers expected that humidity would have a substantial negative effect on NO_x emissions and that temperature might not be an important predictor. Does your analysis tend to confirm or contradict these expectations?

NO_x (ppm)	Humidity (%)	Temperature ($^\circ F$)	Barometric Pressure (in. Hg)
0.70	96.50	78.10	29.08
0.79	108.72	87.93	28.98
0.95	61.37	68.27	29.34
0.85	91.26	70.63	29.03
0.79	96.83	71.02	29.05
0.77	95.94	76.11	29.04
0.76	83.61	78.29	28.87
0.79	75.97	69.35	29.07
0.77	108.66	75.44	29.00
0.82	78.59	85.67	29.02
1.01	33.85	77.28	29.43
0.94	49.20	77.33	29.43
0.86	75.75	86.39	29.06
0.79	128.81	86.83	28.96
0.81	82.36	87.12	29.12
0.87	122.60	86.20	29.15
0.86	124.69	87.17	29.09
0.82	120.04	87.54	29.09
0.91	139.47	87.67	28.99
0.89	105.44	86.12	29.21

3 The following data are physical measurements (all in cm) taken on female applicants to the police department of a metropolitan police force. They are a portion of a much larger data base that was compiled to investigate physical requirements for police officers. Fit a linear regression model to these data, using the overall height as the response variable. Assess the adequacy of the fit.

Measurements of Female Police Department Applicants

Applicant	Overall Height	Sitting Height	Upper Arm Length	Forearm Length	Hand Length	Upper Leg Length	Lower Leg Length	Foot Length
1	165.8	88.7	31.8	28.1	18.7	40.3	38.9	6.7
2	169.8	90.0	32.4	29.1	18.3	43.3	42.7	6.4
3	170.7	87.7	33.6	29.5	20.7	43.7	41.1	7.2
4	170.9	87.1	31.0	28.2	18.6	43.7	40.6	6.7
5	157.5	81.3	32.1	27.3	17.5	38.1	39.6	6.6
6	165.9	88.2	31.8	29.0	18.6	42.0	40.6	6.5
7	158.7	86.1	30.6	27.8	18.4	40.0	37.0	5.9
8	166.0	88.7	30.2	26.9	17.5	41.6	39.0	5.9
9	158.7	83.7	31.1	27.1	18.3	38.9	37.5	6.1
10	161.5	81.2	32.3	27.8	19.1	42.8	40.1	6.2
11	167.3	88.6	34.8	27.3	18.3	43.1	41.8	7.3
12	167.4	83.2	34.3	30.1	19.2	43.4	42.2	6.8
13	159.2	81.5	31.0	27.3	17.5	39.8	39.6	4.9
14	170.0	87.9	34.2	30.9	19.4	43.1	43.7	6.3
15	166.3	88.3	30.6	28.8	18.3	41.8	41.0	5.9
16	169.0	85.6	32.6	28.8	19.1	42.7	42.0	6.0
17	156.2	81.6	31.0	25.6	17.0	44.2	39.0	5.1
18	159.6	86.6	32.7	25.4	17.7	42.0	37.5	5.0
19	155.0	82.0	30.3	26.6	17.3	37.9	36.1	5.2
20	161.1	84.1	29.5	26.6	17.8	38.6	38.2	5.9
21	170.3	88.1	34.0	29.3	18.2	43.2	41.4	5.9
22	167.8	83.9	32.5	28.6	20.2	43.3	42.9	7.2
23	163.1	88.1	31.7	26.9	18.1	40.1	39.0	5.9
24	165.8	87.0	33.2	26.3	19.5	43.2	40.7	5.9
25	175.4	89.6	35.2	30.1	19.1	45.1	44.5	6.3
26	159.8	85.6	31.5	27.1	19.2	42.3	39.0	5.7
27	166.0	84.9	30.5	28.1	17.8	41.2	43.0	6.1
28	161.2	84.1	32.8	29.2	18.4	42.6	41.1	5.9
29	160.4	84.3	30.5	27.8	16.8	41.0	39.8	6.0
30	164.3	85.0	35.0	27.8	19.0	47.2	42.4	5.0
31	165.5	82.6	36.2	28.6	20.2	45.0	42.3	5.6
32	167.2	85.0	33.6	27.1	19.8	46.0	41.6	5.6
33	167.2	83.4	33.5	29.7	19.4	45.2	44.0	5.2

4 Refer to Exercise 2. Define second-order cross-product and quadratic terms for each of the three predictor variables in the nitrogen oxides data set. Do so without centering or scaling the raw humidity, temperature, or barometric-pressure variables. Calculate correlations between all pairs of the nine second-order terms. Are there any obvious collinearities among

the second-order terms? Comment on the advisability of including all the second-order terms in a model for the nitrogen oxide emissions.

5 Calculate least-squares coefficient estimates for the complete nine-term second-order fit to the nitrogen oxides emissions using the untransformed second-order terms for Exercise 4. Compare the least-squares coefficient estimates of the linear parameters from this fit with those from a fit using only the linear terms. How have the coefficient estimates changed? Are the effects of the collinearities apparent? If so, how? Which fit would you prefer to use? Why?

6 Redo Exercises 4 and 5 using standardized predictor variables. How do the conclusions from the two previous exercises change?

7 Refer to the data in Exercise 3. Two indices that are often important in studies of this type are the brachial index and the tibiofemural index. The brachial index is the ratio of the upper-arm length to the forearm length. The tibiofemural index is the ratio of the upper-leg length to the lower-leg length. Fit a regression model to the overall height using the seven original predictors and the two indices. Compare the coefficient estimates of the nine-predictor fit with those of the fit to the seven original variables. Examine the t-statistics for the two fits. Are there any important differences in the fits, especially among the forearm and leg predictors?

8 Consider the algebraic form of a first-order multiple linear regression model, equation (15.24). Use suitable transformations of the predictor variables and the regression coefficients to rewrite the original model in terms of the standardized predictors (15.26). How do the original model parameters relate to the parameters in the transformed model? Assume that least-squares estimates are available for the original or for the transformed model. The least-squares estimates for the two models satisfy the same relationships as the two sets of model parameters. Express the least-squares estimates for the parameters of the transformed model as a function of the least-squares estimates of the original model parameters.

9 A study was conducted to determine the effects of various fuel properties on the startability of diesel engines used in Navy ships. Each engine was placed in a refrigerated box to simulate cold-starting conditions and tested using a variety of fuels. A temperature of 5° C was chosen as representative of actual engine below-deck operating conditions. The response variable of interest was the time required to start the engine (in seconds) in the cold environment. The predictor variables included the cranking speed (in rpm), and various diesel fuel properties including the pour point (in °C), the aniline point (in °C), and the cetane number. The data are given below for 29 observations on a specific engine using 11 different fuels.

(a) Fit a multiple regression model to this data and present the resulting prediction equation.

(b) Test for the significance of the overall regression.

(c) Calculate and interpret the R^2 value and the estimated error standard deviation.

(d) Use pairwise correlations to determine if any strong collinearities exist among the four predictor variables.

(e) Test hypotheses on the individual regression coefficients. Determine which predictor variables are most useful in predicting the engine start time. Comment on your choice of a significance level in terms of the importance of Type I and Type II errors.

Fuel	Start Time	Cranking Time	Pour Point	Aniline Point	Cetane Number
A	26.5	181	−17	66.1	48.7
A	15.7	198	−17	66.1	48.7
A	8.4	206	−17	66.1	48.7
A	4.0	232	−17	66.1	48.7
A	8.3	216	−17	66.1	48.7
A	9.0	218	−17	66.1	48.7
A	5.0	218	−17	66.1	48.7
B	20.0	227	−13	62.7	46.2
B	17.2	227	−13	62.7	46.2
B	29.1	210	−13	62.7	46.2
B	22.8	194	−13	62.7	46.2
B	30.4	195	−13	62.7	46.2
C	38.3	194	−17	62.8	45.2
C	26.3	202	−17	62.8	45.2
C	33.0	193	−17	62.8	45.2
D	48.2	195	−19	55.6	43.6
D	31.3	200	−19	55.6	43.6
E	39.3	190	−14	63.0	47.0
E	35.8	193	−14	63.0	47.0
F	42.0	190	−14	60.5	44.9
F	39.1	186	−14	60.5	44.9
G	30.3	196	−17	61.7	45.6
H	23.4	203	−14	64.5	48.1
I	27.5	203	−16	61.9	46.3
I	30.1	206	−16	61.9	46.3
J	21.5	208	−16	64.6	48.4
J	21.8	206	−16	64.6	48.4
K	35.7	202	−21	53.6	41.8
K	32.9	201	−21	53.6	41.8

10 The data in Exercise 9 contain repeat observations on most of the fuels in the study. Using this additional information conduct a lack-of-fit test for the derived model. Interpret your result and determine if any predictor-variable re-expressions would be useful.

11 A study was run to determine a prediction equation for the minimum unaided starting temperature (MUST) in °C for a two-cycle diesel engine as a function of various fuel properties. Although 21 fuels were included in the project, MUST measurements were available on only 18 fuels due to various experimental problems. The fuels represented blended fuels as well as fuels with no additional additives in them. The MUST value was found by successively cooling the engine to a particular temperature and then attempting to start the engine. When two "start" and two "no-start" tests were completed that were no more than 2° C apart in temperature, the average of the two no-start and two start temperatures was used as the corresponding MUST value. Affecting the MUST values are potentially several fuel properties, including cetane number, kinematic viscosity at 40° C, distillation temperature in °C at 50% boil off, and the auto-ignition temperature in °C.

(a) Fit a multiple regression model to the MUST data given below using only cetane number and viscosity as the predictor variables.

(b) Test for the significance of the regression fit.

(c) Construct a 95% joint confidence region for the two regression coefficients and superimpose on it the individual confidence intervals. Interpret your results.

(d) What could be done to narrow the joint confidence region derived in part (c)?

Fuel	MUST	Kinematic Viscosity	Cetane Number	50% BP	Auto-ignition Temperature (°C)
1	−4.0	3.00	50	282.8	245
2	−1.0	1.50	45	218.0	250
3	−8.6	1.95	57	236.9	245
4	2.8	1.07	41	191.2	255
5	−9.4	1.56	47	236.2	265
6	1.0	0.78	35	162.6	245
7	−4.2	1.12	43	191.2	191
8	−7.6	1.39	49	177.8	185
9	−6.0	2.07	48	223.7	185
10	−9.8	2.57	60	238.3	179
11	9.0	0.76	28	166.2	202
12	12.0	3.74	35	306.1	210
13	−4.0	0.82	39	169.8	190

Fuel	MUST	Kinematic Viscosity	Cetane Number	50% BP	Auto-ignition Temperature (°C)
14	4.0	0.78	40	164.3	190
15	3.5	3.73	37	306.2	204
16	−9.4	1.46	42	212.8	185
17	−4.5	1.80	39	254.0	190
18	−4.5	1.49	44	218.6	183

12 Fit the MUST data given in Exercise 11 using the first three predictor variables, and again refit it using all four predictor variables. Compare the resulting two prediction equations. Does the 50% BP coefficient change from significant to nonsignificant between these two fits for any reasonable significance level? If so, explain what properties of the data cause this result. Select the fit you prefer and justify your answer.

13 Given the following data, fit a model to y using x_1, x_2, x_3, and x_4 as the predictor variables, and test for significance of the regression. Compute the t-statistics for each regression coefficient and interpret the results.

y	x_1	x_2	x_3	x_4
16.20	1.1858	0.3019	1.2826	1.7109
16.81	1.0517	0.2667	1.2129	1.8275
14.41	1.3362	0.3337	1.2707	1.6452
14.37	1.3552	0.3291	1.2396	1.6818
16.96	1.1567	0.2471	1.1810	2.0106
15.22	1.2698	0.2485	1.1248	2.0722
12.49	1.1096	0.1819	0.9935	2.3053
10.16	1.3384	0.2898	1.1225	1.8793
23.01	0.6008	0.0640	1.4032	3.6711
23.70	1.9174	0.4184	1.4014	1.4909

14 Refit the data in Exercise 13 using only x_1, x_2, x_3, and then again using only x_2 and x_3. Compare the fits of these two models with the fit obtained in Exercise 13. Note the change in the R^2 values and adjusted R^2 values across these models. Do you think the change is due to the significance of the regression coefficient or to the small sample size? Explain your answer.

15 Algebraically prove that the square of the correlation coefficient between the observed and the predicted values of a response variable in a regression model equals the R^2 value obtained from the associated regression fit. Numerically confirm this result using the prediction equation derived in Exercise 13.

16 Fit a polynomial model to the following data using Y as the response variable and X_1 as the predictor variable by completing the following steps.

(a) Use a scatter plot of Y versus X_1 to determine the highest power of X_1 to use in the model.

(b) Fit the specified polynomial model and present the resulting prediction equation.

(c) Test for the overall significance of the regression, and for the significance of the individual powers of X_1. Comment on your choice of a significance level for these tests in terms of the importance of Type I and Type II errors.

Y	X_1	X_2	Y	X_1	X_2
10.00	1.07	1500	8.50	1.83	1410
2.00	3.00	4300	7.80	1.35	1710
8.00	1.51	1200	2.40	3.54	4030
3.00	2.02	3200	2.00	2.39	4000
5.00	2.20	2000	4.00	3.80	3000
6.00	1.90	1600	5.00	3.15	2800
10.00	3.11	1200	3.80	2.39	4000
10.00	1.83	1200	2.75	3.70	4500
6.00	2.35	1300	4.44	3.15	3500
2.00	2.79	4000	9.00	0.97	1700
10.00	1.90	1100	10.00	1.10	1800
10.00	1.37	1500	7.30	1.90	1900
10.00	1.55	1400	9.00	1.50	1900
5.80	1.64	1650	8.00	3.10	2100
7.00	2.18	1750	3.70	2.70	4200
4.00	3.17	5000	4.70	3.00	4500
10.00	1.79	0	4.00	2.46	0
3.80	2.94	3700	6.00	2.00	0
4.50	3.48	3100	10.00	1.83	0
8.50	1.68	1350	10.00	1.88	0
8.50	1.58	1700	10.00	1.84	0
3.00	2.13	3200	3.00	1.88	0
5.30	3.10	2700	3.00	1.84	0
2.80	1.82	0	2.50	2.30	0
2.10	2.87	3830	2.00	2.40	0
7.90	1.91	2040	8.00	2.00	0
3.70	2.74	4030	2.00	3.95	0
6.90	1.51	2040			

17 Repeat Exercise 16, but use X_2 in place of X_1 in your polynomial model. Compare the two fits and determine which provides the better fit.

18 Using the data in Exercise 16, fit Y to a model containing X_1 and X_2. Plot the fitted model in a three-dimensional plane similar to the plot given in Figure 15.3. Add the square of X_2 to the model and refit it using this term as well as the linear terms in X_1 and X_2. Again plot the fitted model in a three-dimensional plane. Compare the resultant plots and prediction models, and state any conclusions about which fit might be preferred.

19 Compare the three-variable fit derived in Exercise 18 involving X_1, X_2, and $X_2{}^2$ with the single-variable fit derived in Exercise 16 using X_1 alone. Use the F-statistic in Equation (15.18) to make the comparison. State a conclusion about the usefulness of adding both the linear and the quadratic terms in X_2.

20 Given the data below, fit a model to Y using X_1, X_2, and $X_2{}^2$ as the three predictor variables. Plot the fitted model in a three-dimensional plane similar to the plot given in Figure 15.3. Using the individual t-statistics for the regression coefficients, determine which predictor terms can be deleted from the model.

Y	X_1	X_2
0	650	−35
20	650	340
40	650	550
60	650	580
80	650	588
100	650	591
0	800	−50
20	800	300
40	800	514
60	800	648
80	800	664
100	800	667
0	1200	−102
20	1200	173
40	1200	529
60	1200	980
80	1200	1264
100	1200	1280

Y	X_1	X_2
0	1600	−131
20	1600	60
40	1600	411
60	1600	678
80	1600	951
100	1600	1035
0	1950	−159
20	1950	42
40	1950	365
60	1950	560
80	1950	800
100	1950	870

CHAPTER 16

Linear Regression with Factors and Covariates as Predictors

Many regression analyses involve categorical predictors. For example, one might wish to include as predictors factors from a designed experiment along with quantitative covariates that were measured but not controlled in an experiment. Through the proper definition of predictors to represent the controlled factors, regression models can be fit that are equivalent to the usual ANOVA models for designed experiments. Models that include both controlled factors and one or more quantitative covariates are often referred to as analysis of covariance (ANCOVA) models. In this chapter, the similarity between regression models and ANOVA models that relate a response variable to both experimental factors and covariates is developed. In particular, this chapter highlights

- *the equivalence of regression models and ANOVA models for designed experiments,*
- *the proper definition of quantitative predictors to represent categorical factor levels, and*
- *the analysis of data from completely randomized designs and from randomized complete block designs for models that include a single covariate.*

The first section of this chapter details the important definition of indicator variables and their relationship to factors and categorical predictors. With the appropriate definition of indicator variables to represent categorical predictor values, all the regression methods discussed in the previous two chapters are applicable to the analysis of the effects of the predictors on responses. The

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

last two sections of this chapter detail two specific applications that are common in experimental work: ANCOVA model fits for completely randomized experimental designs and for randomized complete block designs, each of which includes one quantitative covariate.

16.1 RECODING CATEGORICAL PREDICTORS AND FACTORS

Multiple linear regression models were defined in Chapter 15, Equation (15.1) as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i \qquad i = 1, 2, \ldots, n. \tag{16.1}$$

In the definition of model (16.1), each of the predictor values x_{ij} must be numerical. To include categorical predictors or qualitative factors in a regression model like Equation (16.1), the categorical predictors or qualitative factors must be recorded to have meaningful quantitative values. In such circumstances, indicator (dummy) variables are preferred to other possible quantitative predictors. The reminder of this section details the rationale for this preference.

16.1.1 Categorical Variables: Variables with Two Values

Categorical variables were used extensively in Chapters 4 to 13, where they were defined as factors in designed experiments. Categorical variables are intrinsically nonnumerical. In a regression model, categorical predictor variables are any qualitative variables that could be important in the prediction of response-variable values. Categorical variables, in this context, can be obtained as controllable factors in designed experiments or as uncontrollable covariates in either designed or observational studies.

ANOVA models and the regression model (16.1) can be shown to be equivalent if categorical variables (factors) are assigned appropriate numerical codes. There are many numerical codes that can be used and that are equivalent to one another. Effects coding was used in Chapters 5 (e.g., Tables 5.5 to 5.7) and 7 to represent main effects and interactions. This type of coding can be used in a regression analysis. The following example introduces an alternative coding that clarifies the connection between ANOVA and regression models.

Recall the tire treadwear experiment about three tire brands in Section 15.1.2. Suppose that interest lies in a comparison of only two of the test-tire brands. If one wishes to account for the effects of temperature (x_1 = Fahrenheit temperature/100) and tire brand (x_2) on the relative wear rate (y), the following *indicator variable* for tire brand can be used:

$$x_2 = \begin{cases} 1 & \text{if brand A} \\ 0 & \text{if brand B.} \end{cases}$$

TABLE 16.1 Tire Treadwear Rates for Tire Brands A and B

Observation Number	Temperature (°F)	Relative Wear	
		Brand A	Brand B
1	66.0	1.17272	1.58489
2	66.0	1.43768	1.76644
3	66.0	1.16290	1.65358
4	66.0	1.51047	1.56035
5	70.3	1.00000	1.44000
6	70.3	1.33778	1.37333
7	70.3	0.98667	1.40000
8	70.3	1.28889	1.41778
9	53.2	1.15254	1.63983
10	53.2	1.11864	1.50000
11	53.2	1.01695	1.42797
12	53.2	1.32627	1.34322
13	53.3	1.07066	1.50321
14	53.3	1.38758	1.44754
15	53.3	1.05353	1.43897
16	53.3	1.19058	1.37045
17	88.1	0.97074	1.48138
18	88.1	1.25000	1.42819
19	88.1	0.98670	1.47606
20	88.1	0.80319	1.41223
21	89.6	0.94475	1.33149
22	89.6	0.81215	1.40055
23	89.6	0.94751	1.41436
24	89.6	1.32044	1.37569
25	78.4	1.33430	1.47093
26	78.4	0.87500	1.44477
27	78.4	1.20640	1.47674
28	78.4	1.22384	1.50872
29	76.9	1.36145	1.40060
30	76.9	1.34036	1.48494
31	76.9	1.22289	1.48193
32	76.9	1.21687	1.43675

A regression model fitted to the 64 observations shown in Table 16.1 for these two tire brands using the method of least squares (Section 15.1) resulted in the following fit:

$$\hat{y} = 1.68 - 0.30x_1 - 0.31x_2. \tag{16.2}$$

In a regression analysis the effect a predictor variable has on the response equals the difference in predicted responses for two values of the predictor variable(s). If one fixes the temperature variable at any value, the effect of the tire brands is the difference in $\hat{y}$ at $x_2 = 1$ and $x_2 = 0$. This difference is the estimated regression coefficient $b_2 = -0.31$. Thus, the effect of the tire brands is obtainable from the estimated coefficient for the tire-brand variable. With the definition of the tire-brand categorical variable given above, the estimated coefficient b_2 indicates that brand A results in a lower estimated relative wear rate than brand B. Note that the difference in predicted responses estimates the difference in two model means.

Another coding scheme that could be used to designate the tire brands is the effects coding that was used in Chapters 5 and 7:

$$x_2^* = \begin{cases} 1 & \text{if brand A,} \\ -1 & \text{if brand B} \end{cases}.$$

The resulting fitted prediction equation is

$$\hat{y} = 1.525 - 0.30x_1 - 0.155x_2^*.$$

By noting that $x_2 = (x_2^* + 1)/2$, the equivalence between the above fitted model and Equation (16.2) is readily established:

$$\begin{aligned} \hat{y} &= 1.525 - 0.30x_1 - 0.155x_2^* \\ &= 1.525 - 0.30x_1 - 0.155(2x_2 - 1) \\ &= 1.68 - 0.30x_1 - 0.31x_2. \end{aligned}$$

Other choices of values for the coding on x_2 result in changes in the estimates of the intercept and slope coefficients, as did these two choices for the tire-brand coding. The predicted responses, however, remain the same regardless of the coding used. In any coding, the predicted response at a specified temperature for tire brand A is 0.31 units lower than that for tire brand B.

When reporting a fitted regression equation involving categorical variables, it is generally preferable to list separate equations for each value of the categorical variable. This is done to preclude confusion over the use of the categorical variable and because ready comparisons may be made of the different equations. For the tire treadwear data, one simply inserts the two values for the categorical variable in equation (16.2):

$$\begin{aligned} \hat{y}_A &= 1.37 - 0.30x_1, \qquad \text{brand A,} \\ \hat{y}_B &= 1.68 - 0.30x_1, \qquad \text{brand B.} \end{aligned}$$

As mentioned above, the effect of the tire brands is the difference between the average relative wear rates for the two tires. This difference is the same as would be obtained in an ANOVA model for a factorial experiment in which two controllable factors, temperature and tire brand, were included at preselected levels. In this example, the reason the estimated slope coefficient for tire brands exactly equals the difference in the tire-brand averages, $\overline{y}_A - \overline{y}_B = -0.31$, is that x_2 was coded with the 0–1 convention and the two sets of tire brand measurements, y_{Aj} and y_{Bj}, were obtained from a test run on one vehicle. Because both measurements were taken on the same vehicle, the ambient temperature for each tire-brand measurement in a pair is the same. This experimental layout is identical to that of a factorial experiment in two factors. The major difference is that for this experiment the temperature values are not preselected; they are measured ambient values.

Had the ambient temperature values for each set of tire-brand measurements been different, the estimated coefficient for x_2 would not exactly equal the difference in the averages for the two tire brands. The amount of disagreement between the estimated slope coefficient (b_2) and the difference in the tire brand averages ($\overline{y}_A - \overline{y}_B$) would depend on how similar were the temperature values for the two sets of measurements, that is on how close the combinations of values of tire brands and temperatures were to the layout of a complete factorial experiment.

This example illustrates the equivalence of a two-factor ANOVA model and a regression model when one of the factors is categorical and has two levels. One can algebraically derive the equivalence between the estimated slope coefficient and the main effect for the categorical factor, but we leave that exercise to the interested reader. Note, however, that this discussion was restricted to a two-level factor and models without interaction terms. These results do not necessarily generalize, as we demonstrate in the next subsection.

16.1.2 Categorical Variables: Variables with More Than Two Values

When a categorical variable has more than two values, it is more difficult to isolate the influence of the predictor variable on the response. Care must be taken in model specification so that an ordering or a constant difference between categories is not arbitrarily imposed on the model when no such effect exists on the response variable. This can happen if an arbitrary assignment of values is made to the various categories.

As an example, consider extending the analysis of the above tire treadwear data to include all three test tire brands A, B, and C. One possible coding for brand type is

$$x_2 = \begin{cases} 1 & \text{if brand A,} \\ 2 & \text{if brand B,} \\ 3 & \text{if brand C.} \end{cases}$$

The resulting prediction equation is

$$\hat{y} = 0.76 - 0.30x_1 + 0.54x_2.$$

This fitted model is very similar to the equation obtained using only brands A and B, and the results appear plausible. A major interpretive problem with this fit is that the above definition of x_2 imposes an arbitrary ordering of the effects of the three tire brands. The estimated coefficient for x_2 implies that the estimated relative wear rate for brand B is 0.54 units greater than that for brand A, and the estimated relative wear rate for brand C is 0.54 units greater than that for brand B.

This estimated constant difference and fixed ordering is strictly the result of the coding chosen for x_2 and not necessarily of the relationship of brand type to mean treadwear. Using the above coding with any rearrangement of the tire brands will result in a similar interpretation: the effects of the tire brands will always be ordered according to the order specified in the definition of x_2, and the effects of the three tire brands will always be increasing or decreasing (depending on the sign of b_2) by a constant amount.

This problem can be overcome by using two indicator variables to designate the tire type. For example, let

$$x_2 = \begin{cases} 1 & \text{if brand A,} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad x_3 = \begin{cases} 1 & \text{if brand B,} \\ 0 & \text{otherwise.} \end{cases}$$

Using these predictor variables, tire brand C is indicated by $x_2 = x_3 = 0$. This specification imposes no ordering or fixed differences in tire-brand effects on treadwear. The resulting prediction equation for the treadwear data is

$$\hat{y} = 2.45 - 0.30x_1 - 1.08x_2 - 0.77x_3.$$

The arbitrariness of the fit when using a single categorical variable to specify three or more levels of a categorical factor is well illustrated with this example. While the signs on b_2 and b_3 in this fit are both negative, the latter estimated coefficient is almost half the size of the former one. The effect of tire brand on treadwear is that brand A decreases the estimated relative wear rate by 1.08 units over brand C, while brand B decreases the estimated relative wear rate by 0.77 units over brand C. The use of two predictor variables leads to correct conclusions about the influence of the tire brands on treadwear.

Many different specifications could be used to code a categorical variable with three or more values. One of the most common specifications of categorical variables in regression analyses is given in Exhibit 16.1.

EXHIBIT 16.1 CODING CATEGORICAL VARIABLES

If a categorical variable can take on any one of k different values, use $k - 1$ indicator variables of the form

$$x_j = \begin{cases} 1 & \text{if jth value,} \\ 0 & \text{otherwise} \end{cases} \tag{16.3}$$

for $k - 1$ values of the categorical variable. It is arbitrary which $k - 1$ values are chosen.

While it is arbitrary which of the values of the categorical variable are assigned to the indicator variables, there is sometimes a category that is preferable. Note that the kth category is identified by $x_1 = x_2 = \cdots = x_{k-1} = 0$. The estimated regression coefficient for x_j then measures the effect of changing from the kth category to the jth one. This was the interpretation used in the above example, and it is the interpretation used with all two-level predictor variables. Hence, if one of the values of the categorical variable represents a standard, it should represent the kth level.

The above tire-treadwear example illustrates a secondary use of categorical variables, namely in specifying an ANCOVA model (see Sections 16.2 and 16.3). An ANCOVA model is a regression model containing both quantitative variables and categorical variables. Often, the categorical variable is of most interest, but uncontrollable quantitative factors also affect the response. In such circumstances, ANCOVA models are defined and analyzed in a manner similar to that in the above example.

16.1.3 Interactions

In Chapter 15, interactions were interpreted by relating them to joint factor effects in an ANOVA model. Interactions have a similar interpretation in terms of the joint effects of predictor variables in regression models. Most often interaction terms are formed as products of two or more predictor variables, although they can be specified in any manner felt to be reasonable for the model under investigation.

Algebraically one can show that if a factor in an ANOVA model is represented in a regression model in terms of indicator variables as defined by equation (16.3), the ANOVA sum of squares for its main effect is obtainable as the total of the regression sums of squares for the effects represented by the individual indicator variables. There will be $k - 1$ such effects, one for each indicator variable, because there will be $k - 1$ of these terms in the regression model (16.1). Note too that the total number of degrees of freedom for these indicator variables, $k - 1$, is the same as for a main effect in an ANOVA model.

Continuing with this representation, an interaction between two categorical factors can be represented in regression models by products of the individual indicator variables for each factor. There are $(a-1)(b-1)$ such products for two factors having, respectively, a and b levels. The sum of squares for the interaction between the two factors would be obtained as the total of the individual regression sums of squares for the product terms. The total interaction degrees of freedom are $(a-1)(b-1)$, the number of degrees of freedom stated in Chapter 6 for two-factor interactions.

This development can be extended to any number of factors having an arbitrary number of levels and to interactions involving more than two factors. Thus, the assignment of sums of squares to main effects and interactions in ANOVA models results from a consideration of their expression as indicator variables in regression models.

16.2 ANALYSIS OF COVARIANCE FOR COMPLETELY RANDOMIZED DESIGNS

In many experimental situations, responses are affected not only by the controllable experimental factors, but also by uncontrollable variates. Models can be defined that relate the response to both the controllable and the uncontrolled variates, and the influence of all the variates on the response can be investigated. To distinguish controlled experimental factors from uncontrolled experimental variables, the latter are called *covariates*. (see Exhibit 16.2).

EXHIBIT 16.2

Covariate. A covariate is an uncontrolled experimental variable that influences the response but is itself unaffected by the experimental factors.

It is important to note that the definition in Exhibit 16.2 requires that the covariate be unaffected by the experimental factors. If covariates are so affected, they become additional responses and the analytic procedures in this chapter are inappropriate. Covariates are clearly unaffected by the experimental factors if the covariates can be measured on experimental units prior to the assignment of the factor–level combinations. In other cases, the covariate can only be measured at the same time the response is measured. In this latter situation it may be more difficult, but equally important, to determine whether the experimental factors affect the covariates.

If covariates are excluded from the analysis of the response variable, estimates of the factor effects are biased. When covariates are ignored in an analysis, tests for the statistical significance of the factors also will suffer from a loss of power. In other words, if covariates are ignored in the analysis,

their influences on the response will masquerade as experimental error. This leads to less sensitive factor comparisons, because the estimate of experimental error is inflated by the effects of the covariate. The analysis of experimental data when both covariates and factor variables are present is termed *analysis of covariance* (ANCOVA).

In Table 5.4, data were presented on the lifetimes of cutting tools used with lathes. Primary interest focused on the comparison of two different types of cutting tools, labeled type A and type B for simplicity. As is clearly evident from Figure 5.2, however, the tool lifetimes in Table 5.4 cannot be directly compared using either a two-sample t-test or a one-factor analysis of variance. This is because the lifetimes for each type of cutting tool are linear functions of the lathe speed. Although the lathe speeds could have been preset at selected values, they were not in this experiment. Thus, lathe speed should be treated as a covariate.

Note that lathe speed and tool type are not affected by one another, so lathe speed can be considered a covariate and not another response. Note too that the trends in Figure 5.2 suggest that the linear relationships between tool life and lathe speed have a common slope for the two tool types. This is not a necessary requirement for ANCOVA models, but a common slope is a reasonable assumption for many applications. Fitted lines with a common slope have been superimposed on Figure 5.2.

A general ANCOVA model having one experimental factor of interest and one covariate can be written as

$$y_{ij} = \mu_i + \beta_i x_{ij} + e_{ij}, \qquad i = 1, 2, \ldots, a, \quad j = 1, 2, \ldots, r.$$

In this model, y_{ij} and x_{ij} are the values of the response variable and the covariate, respectively, for the jth observation associated with the ith level of the experimental factor. The mean μ_i denotes the effect of the ith level of the factor on the response, β_i denotes the linear effect (slope) of the covariate for those observations associated with the ith level of the factor, and e_{ij} denotes the random error. As in Chapter 6, μ_i is often expressed as $\mu + \alpha_i$ with the constraint $\sum \alpha_i = 0$ imposed.

Assuming that there is a common slope for the covariate effect with all levels of the experimental factor, the ANCOVA model can be written as

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + e_{ij}. \tag{16.4}$$

A labeled scatterplot (Section 5.2) is an important guide to whether this assumption of a common slope is reasonable. If a common slope is not a reasonable assumption, general regression techniques (e.g., Chapter 15) can be used to model the effects of the factor and the covariate on the response.

The model (16.4) allows one to estimate the factor effects through estimation of the μ_i. One can also test for significant differences in the effects due to the factor levels by testing

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_\alpha \quad \text{vs} \quad H_\alpha: \text{at least two } \mu_i \text{ differ}$$

or, equivalently,

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0 \quad \text{vs} \quad H_a: \text{at least one } \alpha_i \neq 0.$$

Likewise, one can estimate the effect of the covariate on the response through the estimation of β. One can also test whether the covariate has a statistically significant effect on the response by testing

$$H_0: \beta = 0 \quad \text{vs} \quad H_a: \beta \neq 0.$$

ANCOVA models need not have a single covariate, nor need they be used only when there is a single experimental factor of interest. While this chapter explicitly covers only the single-covariate, single-experimental-factor models in either completely randomized or randomized block designs, ANCOVA models can be generalized to several factors and several covariates. Analysis of these more general models would utilize the general regression models and methodology contained in Chapter 15.

To analyze the data arising from model (16.4), the assumptions listed in Exhibit 16.3 must be valid.

EXHIBIT 16.3 ANALYSIS-OF-COVARIANCE ASSUMPTIONS FOR COMPLETELY RANDOMIZED DESIGNS

1. The experimental data are obtained using a completely randomized statistical design (Section 5.1).
2. The response can be represented as an additive function of the factor effects, a linear covariate effect, and a random error component as in Equation (16.4).
3. The covariate is unaffected by the factor levels.
4. The errors can be considered independently normally distributed with mean zero and constant standard deviation σ.

The statistical significance of the covariate can be assessed by comparing two models, one with and one without the covariate. The full model (M_1) is Equation (16.4). The reduced model (M_2) is

$$y_{ij} = \mu + \alpha_i + e_{ij}. \tag{16.5}$$

The statistical significance of the covariate is indicated in different ways in the output of different computer programs. Some computer programs report a t-statistic, others an F-statistic. Either way, if the program is coded correctly, the results will be the same as those using the principle of reduction in error sums of squares described in Section 8.1.

Denote the error sum of squares for the full model (16.4) by SSE_1. Denote the error sum of squares for the reduced model (16.5) by SSE_2. Then

$$R(M_1|M_2) = \text{SSE}_2 - \text{SSE}_1.$$

There is one more parameter (β) in the full model than in the reduced one. Therefore, using $\text{MS}_E = \text{SSE}_1/(n - a - 1)$ from the full model where $n = ar$,

$$F = R(M_1|M_2)/\text{MS}_E$$

has an F-distribution with $\nu_1 = 1$ and $\nu_2 = n - a - 1$ degrees of freedom under the hypothesis $H_0 : \beta = 0$. This procedure is summarized in Exhibit 16.4.

EXHIBIT 16.4 TESTING THE SIGNIFICANCE OF THE COVARIATE

1. Calculate the reduction in sum of squares for the full and reduced models, $R(M_1|M_2) = \text{SSE}_2 - \text{SSE}_1$.
2. Form the F-statistic $F = R(M_1|M_2)/\text{MS}_E$.
3. If $F > F_\alpha(1, n - a - 1)$, then conclude that $\beta \neq 0$, that is, that the covariate explains a statistically significant amount of the variation of the response variable

---

Table 16.3 exhibits two analysis of variance tables for the tool-life data of Table 5.4. The first ANOVA table is for the model containing the tool-type factor and the lathe-speed covariate. The second ANOVA table eliminates the covariate. Note that the degrees of freedom and the sum of squares for the error term in the reduced model have been increased by the respective quantities for the covariate in the ANOVA table for the full model. This illustrates that failure to include important model terms increases the variability attributable to the error component of the model.

Table 16.3 symbolically illustrates how one ANOVA table can be used to summarize the information from the two fits in Table 16.2. The lines in the table labeled "Model," "Error," and "Total" are the same as those in Table 16.2(a); i.e., they correspond to the full fit to the model, including the covariate. The line labeled "Factor A" includes the degrees of freedom, sum of squares, and mean square for the experimental factor from the second fit,

TABLE 16.2 Analysis of Variance Tables for Tool Life Data

Source	df	SS	MS	F	p-Value
(a) Model 1: Full Model, Including Covariate					
Tool type & speed	2	1418.00	709.00	76.74	0.000
Error	17	157.05	9.24		
Total	19	1575.05			
(b) Model 2: Reduced Model, Excluding Covariate					
Tool type	1	1097.72	1097.72	41.39	0.000
Error	18	477.33	26.52		
Total	19	1575.05			

TABLE 16.3 Symbolic Analysis-of-Variance Table

Source	df	SS*	MS	F	p-Value
Model	a	MSS	MSM	F_M	p_M
Factor A	$a-1$	SS_A	MS_A	F_A	p_A
Covariate	1	SS_C	MS_C	F_C	p_C
Error	$n-a-1$	SSE_1	MS_E		
Total	$n-1$	TSS			

*$SS_C = R(M_1|M_2)$.

excluding the covariate. The sum of squares in this line measures the effect of only factor A and not the covariate.

The line labeled "Covariate" in Table 16.3 measures the effect of the covariate using the principle of reduction in sum of squares. The number of degrees of freedom for the covariate and the sum of squares, SS_C, can be obtained by subtracting the respective quantities for the error term of the full model from those of the reduced model:

$$\text{df(covariate)} = \text{df}(SSE_2) - \text{df}(SSE_1) = (n-a) - (n-a-1) = 1,$$

$$SS_C = SSE_2 - SSE_1.$$

TABLE 16.4 Analysis of Covariance for Tool Life Data

Source	df	SS	MS	F	p-Value
Model	2	1418.00	709.00	76.74	0.000
Tool type	1	1097.72	1097.72	118.80	0.000
Speed	1	320.28	320.28	34.67	0.000
Error	17	157.05	9.24		
Total	19	1575.05			

This type of ANOVA table is frequently printed out by computer programs in lieu of two separate ANOVA tables. The indenting of the factor and covariate lines in Table 16.3 is used to stress that these sums of squares and degrees of freedom total to those for the "Model" line, the line under which the factor and covariate lines are indented.

Table 16.4 is such an ANOVA table for the tool life example. Note that the covariate is highly significant ($F = 34.67$, $p < 0.001$). This indicates that the linear trends seen in Figure 5.2 are statistically significant; that is, the lifetimes of the tools are affected by the lathe speed in addition to the tool type.

Had the covariate been judged nonsignificant, one could proceed to test the statistical significance of the main effect for tool type using the F-statistic in Table 16.4. However, because the speed covariate has been judged statistically significant, that cannot be done. Likewise, if an experimenter knows from theoretical considerations or past experimentation that a covariate is important to the satisfactory modeling of a response, the usual F-statistic for testing the significance of factor effects should not be used. A second example will illustrate the reason for this statement.

Table 16.5 lists data collected from an experiment that was conducted to compare the residence times of two chemical reactors. One of the reactors, the *production* model, is currently used to make an intermediate product for a synthetic fiber. The other reactor, the *experimental* model, incorporates new design features and is being compared with the production model. Short residence times are desirable, since they result in increased throughput and consequently increased production.

A straightforward comparison of the average residence times using a two-sample t-statistic ($t = 5.68$, $p < 0.001$) indicates that the average residence times for the two reactors are significantly different. On the basis of this test one would conclude that the experimental reactor ($\overline{y}_E = 30.71$ min) has a significantly lower average residence time than the production reactor ($\overline{y}_P = 34.45$ min).

It is known from past experience with the production reactor that the residence times are affected by the amount of impurities in the reactor. Data on the

TABLE 16.5 Residence Times and Impurity Measurements for Two Chemical Reactors

Reactor	Residence Time (min)	Impurity Measurement*
Production	33.9	−4.6011
	32.9	−4.5791
	32.7	−3.9769
	33.9	−4.7702
	35.4	−6.1052
	32.7	−4.4565
	33.1	−5.0066
	35.1	−6.3283
	37.3	−7.3091
	37.3	−6.2105
Experimental	30.4	−2.8473
	30.0	−2.6173
	31.0	−2.7969
	29.8	−2.5383
	31.7	−3.5066
	31.9	−3.5066
	29.1	−2.3645
	32.5	−4.0174
	30.5	−3.4122
	30.2	−3.3242

*Logarithm of percent impurity concentration.

amount of impurities were collected for each of the test runs and are recorded in Table 16.5 along with the residence times. Figure 16.1 is a labeled scatterplot of the residence times versus the impurity measurements. This scatterplot suggests that the observed differences in the average residence times for the two reactors might be due to differences in the amount of impurities in the reactors and not necessarily due to the reactor designs.

It is important to note that the experimenters did not believe that the reactors would affect impurity formation. If the reactors were thought to affect the amount of impurities in the test runs, both the residence times and the impurity levels would be response variables and an analysis of covariance would not be appropriate. In the following analysis we assume that a covariance analysis is appropriate, that is, that the level of impurities is not materially affected by the reactor design.

The statistical significance of the effect of the amount of impurities on the residence times is confirmed by the analysis shown in Table 16.6(a). A question that now arises is whether the reactors produce significantly different

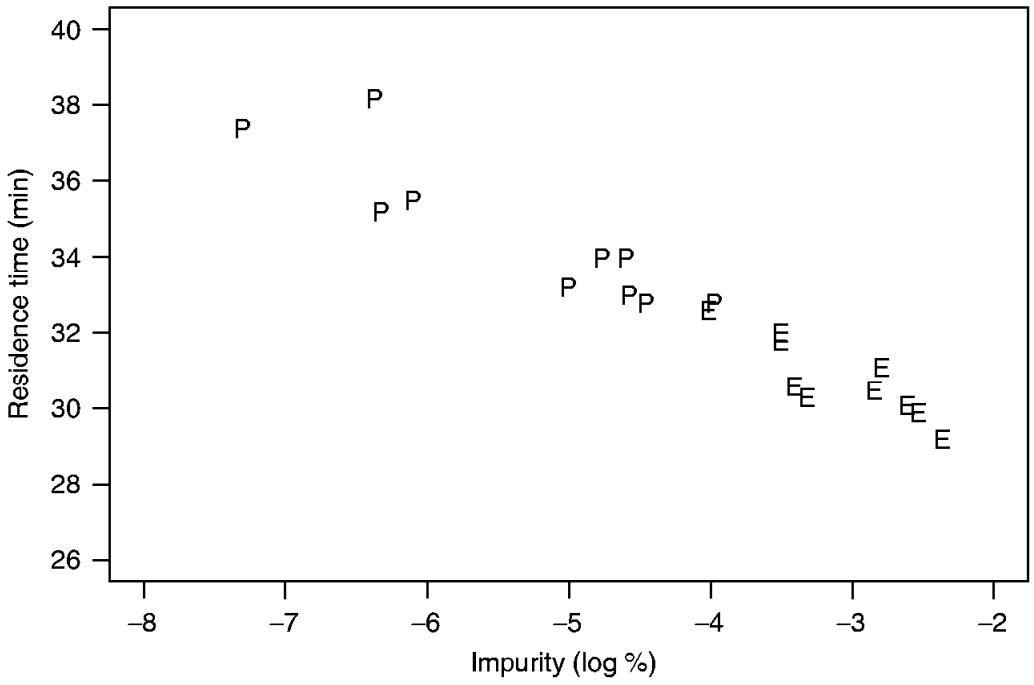


Figure 16.1 Chemical reactor residence-time data. P = Production; E = Experimental.

TABLE 16.6 Analysis-of-Variance Tables for Chemical-Reactor Residence Times

Source	df	SS	MS	F	p-Value
		(a) Impurity Effect Adjusted for Reactor Effect			
Model	2	100.424	50.212	116.96	0.000
Reactors	1	69.192	69.192	161.17	0.000
Impurity	1	31.232	31.232	72.75	0.000
Error	17	7.298	0.429		
Total	19	107.722			
		(b) Reactor Effect Adjusted for Impurity Effect			
Model	2	100.424	50.212	116.96	0.000
Impurity	1	100.325	100.325	233.69	0.000
Reactors	1	0.099	0.099	0.23	0.638
Error	17	7.298	0.429		
Total	19	107.722			

average residence times when the amount of impurities is the same for both. To answer this question the roles of the factor and covariate are reversed in the analysis of variance.

The principle of reduction in sums of squares can provide an appropriate test statistic to examine the effects of the factor on the response after adjusting for the effect of the covariate. Instead of the model (16.5), the reduced model M_3 is now

$$y_{ij} = \mu + \beta x_{ij} + e_{ij}. \tag{16.6}$$

The reduction in sums of squares, $R(M_1|M_3)$, now measures the incremental effect on the response of the factor above that which is provided by the covariate alone. Thus, if it is the covariate that is providing the major effect on the response, $R(M_1|M_3)$ will be small compared to the error mean square MSE_1 from the full model.

Table 16.6(b) reveals that the reactors do not have a statistically significant effect on the average residence times once the covariate has been accounted for. The difference observed in the average residence times is due to the difference in the amount of impurities in the two reactors. The production reactor has a lower average impurity measurement, -5.33, than does the experimental reactor, -3.09. Once this difference is accounted for by fitting the covariate, the remaining differences in residence times are minor.

Figure 16.2 shows the fitted model (16.4) for the two reactors. The common slope is estimated to be $b = -1.55$. One can obtain the two unadjusted

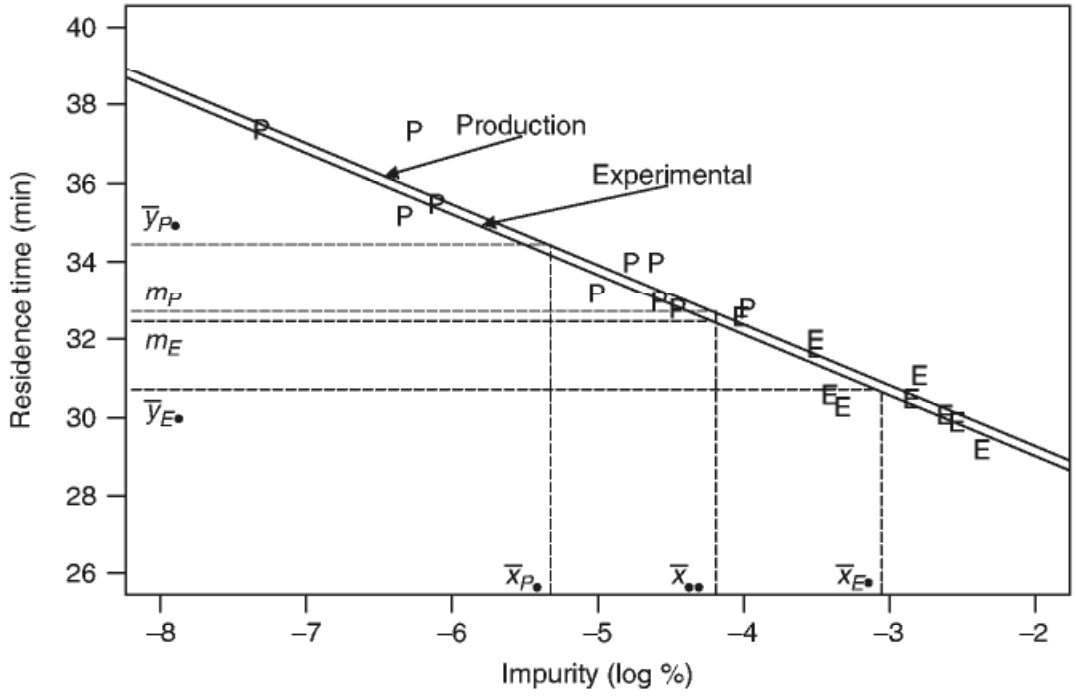


Figure 16.2 Reactor factor effects adjusted for average impurity.

residence-time averages by evaluating the fitted model for each reactor at its average impurity level. This is graphically shown in Figure 16.2 by proceeding vertically from each average impurity ($\bar{x}_{P\bullet}$ and $\bar{x}_{E\bullet}$) value to the line for the corresponding reactor and then reading the residence time ($\bar{y}_{P\bullet}$ and $\bar{y}_{E\bullet}$) from the vertical scale.

Fitting the ANCOVA model allows one to obtain adjusted averages, namely, the residence times corresponding to a common level of impurity. Any level of impurity could be chosen to adjust the residence-time averages. Although the adjusted averages will depend on the level of impurity chosen, the *difference* of any two adjusted averages does not depend on the level of impurity. It is common, when no preferred choice such as a trade or industry standard is available, to adjust the averages at the overall average covariate value, $\bar{x}_{\bullet\bullet}$, as indicated in Figure 16.2.

Adjusting the response averages to the common covariate value $\bar{x}_{\bullet\bullet}$ can be accomplished with the following formula (see the appendix to this chapter):

$$m_i = \bar{y}_{i\bullet} - b(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}). \tag{16.7}$$

Note that if the covariate averages for the different factor levels are all equal (that is, $\bar{x}_{i\bullet} = \bar{x}_{\bullet\bullet}$), the adjusted factor averages equal the ordinary averages $\bar{y}_{i\bullet}$. The cutting-tool lifetime example (see Figure 5.2) illustrates a situation where this is approximately true. When the covariate averages are not equal, as with the reactor residence times, the adjusted averages can be much different from the ordinary averages. For the reactor residence times,

$$m_E = 30.71 + 1.55(-3.09 + 4.21) = 32.45,$$
$$m_P = 34.43 + 1.55(-5.33 + 4.21) = 32.69.$$

The difference in these adjusted factor averages, 0.24 min, more accurately indicates the differences in the average residence times of the two reactors than does the difference in the ordinary averages, 3.72 min. Figure 16.2 graphically portrays the adjustment of the reactor averages for the effect of the difference in average impurities.

When there are more than two levels of the factor of interest, merely calculating the difference in adjusted factor averages will not suffice to determine which of the pairs of factor averages are significantly different. The overall F-test using the principle of reduction in error sums of squares as in Table 16.6(b) can be used to determine that one or more pairs of means are different from one another, but it cannot identify which pairs are different when there are three or more factor levels.

Multiple-comparison procedures discussed in Section 6.4 can be used for this purpose once the estimated standard errors of the differences have been calculated. The estimated standard error of the difference between two adjusted

factor averages, $m_i - m_k$, is given by

$$\widehat{\mathrm{SE}}(m_i - m_k) = \left[\mathrm{MS}_E \left(n_i^{-1} + n_k^{-1} + \frac{(\bar{x}_{i\bullet} - \bar{x}_{k\bullet})^2}{\mathrm{SS}_{xx}}\right)\right]^{1/2}, \tag{16.8}$$

where MS_E is the mean squared error from the fit to the full model (16.4), n_i and n_k are the numbers of observations for the two factor levels, $\bar{x}_{i\bullet}$ and $\bar{x}_{k\bullet}$ are the covariate averages for the two factor levels, and SS_{xx} is the error sum of squares from an ANOVA fit to the covariate:

$$\mathrm{SS}_{xx} = \sum_i \sum_j (x_{ij} - \bar{x}_{i\bullet})^2.$$

The estimated standard errors (16.8) can also be used to construct confidence intervals. The ANCOVA procedure for completely randomized designs is summarized in Exhibit 16.5.

EXHIBIT 16.5 ANALYSIS-OF-COVARIANCE PROCEDURES FOR COMPLETELY RANDOMIZED DESIGNS

1. Using appropriate computer software or calculating formulas, compute the error sums of squares for the models (16.4), (16.5), and (16.6).
2. Calculate the ANOVA table as in Table 16.3 and test the statistical significance of the covariate. If the covariate is not statistically significant, analyze the data using appropriate ANOVA procedures, ignoring the covariate.
3. If the covariate is statistically significant or if the experimenter knows from theoretical considerations that the covariate is required to satisfactorily model the response,
   (a) calculate the ANOVA table adjusting the factor effects for the covariate [e.g., Table 16.6(b)];
   (b) test the statistical significance of the adjusted factor averages, using the estimated standard errors of the differences, as appropriate, to assess which pairs of factor levels are statistically different.

16.3 ANALYSIS OF COVARIANCE FOR RANDOMIZED COMPLETE BLOCK DESIGNS

ANCOVA procedures can be extended to most of the experimental designs discussed in this book. In this section we illustrate the extension to randomized

TABLE 16.7 Randomized Block Design for Reactor Residence Times

	Production Reactor		Experimental Reactor	
Block	Residence Times	Impurity Measurement	Residence Times	Impurity Measurement
1	33.9	−4.6011	30.4	−2.8473
2	32.9	−4.5791	30.0	−2.6173
3	32.7	−3.9769	31.0	−2.7969
4	33.9	−4.7702	29.8	−2.5383
5	35.4	−6.1052	31.7	−3.5066
6	32.7	−4.4565	31.9	−3.5066
7	33.1	−5.0066	29.1	−2.3645
8	35.1	−6.3283	32.5	−4.0174
9	37.3	−7.3091	30.5	−3.4122
10	37.3	−6.2105	30.2	−3.3242

block designs. For simplicity and clarity of discussion we again consider only a single factor and one covariate.

Suppose that the raw materials used in the reactor example were delivered in batches and that only two test runs could be made from the raw material in each batch. One might decide to split the batches and make one run from each of the reactors from the raw material in a batch. The randomized block designs of Chapter 9 would then be appropriate for this experiment.

Table 16.7 is an illustration of how the experimental results might appear in such a randomized block design. The intent of the covariance analysis in this design would be the same as for the completely randomized design: to determine whether the two reactor types (experimental and production) have a significantly different mean residence time at a specified value of the impurity measurement. The analysis procedure must now take into account the fact that the experiment was blocked.

The ANCOVA model for data from a randomized complete block design in which no repeat tests have been made can be expressed as

$$y_{ij} = \mu + \alpha_i + c_j + \beta x_{ij} + e_{ij}, \quad i = 1, 2, \ldots, a, \quad j = 1, 2, \ldots, b, \quad (16.9)$$

where c_j now represents the (random) effect of the jth block (batch of raw material) on the response. The assumptions for ANCOVA models used with randomized complete block designs are given in Exhibit 16.6. The last assumption is critical for block designs. "Blocks" that interact with factors are actually additional factors, not truly blocks of homogeneous experimental units or test runs.

EXHIBIT 16.6 ANCOVA ASSUMPTIONS FOR RANDOMIZED COMPLETE BLOCK DESIGNS

1. The experimental data are obtained using a randomized complete block design (Section 9.2).
2. Block effects can be either fixed or random. If random, they are assumed to be independent normal variates having mean zero and constant standard deviation σ_c.
3. The response can be represented as an additive function of the factor effects, the block effects, a linear covariate effect, and a random error component as in Equation (16.9).
4. The errors can be considered independently normally distributed with mean zero and constant standard deviation σ. The error components are independent of the block components (if the blocks are random).
5. The covariate is unaffected by either the blocks or the factor levels.
6. The block and factor–level components of the model do not interact.

A variety of interesting questions can be asked about the ANCOVA model (16.9). Does the covariate influence the response? If so, are the factor levels significantly different from one another after adjusting for the covariate? Are the block effects statistically significant?

The complete model and three reduced models will be used to answer these questions. We denote the four models as follows:

$$
\begin{aligned}
M_1&: \text{ The complete model, (16.9);}\\
M_2&: \ y_{ij} = \mu + \alpha_i + \beta x_{ij} + e_{ij};\\
M_3&: \ y_{ij} = \mu + \alpha_i + c_j + e_{ij};\\
M_4&: \ y_{ij} = \mu + c_j + \beta x_{ij} + e_{ij}.
\end{aligned}
$$

Note that M_1 is the full ANCOVA model for a randomized complete block design. Models M_2 through M_4 are reduced models: M_2 has no block components, M_3 has no covariate, and M_4 has no factor–level components.

To investigate the questions raised above, three reductions in sums of squares are calculated: $R(M_1|M_2)$, $R(M_1|M_3)$, and $R(M_1|M_4)$. These reductions can be used, respectively, to test the following hypotheses:

$$
\begin{aligned}
R(M_1|M_2)&: \ H_0: \sigma_c = 0 \text{ (blocks random) or}\\
&\qquad H_0: \gamma_1 = \gamma_2 = \cdots = \gamma_b = 0 \text{ (blocks fixed);}\\
R(M_1|M_3)&: \ H_0: \beta = 0;\\
R(M_1|M_4)&: \ H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0.
\end{aligned}
$$

TABLE 16.8 ANCOVA for Residence-Time Data in a Randomized Block Design

Source	df	SS	MS	F	p-Value
Model	11	103.865	9.442	19.59	0.000
Batches	9	3.442	0.382	0.79	0.634
Reactors	1	0.478	0.478	0.99	0.348
Impurity	1	14.001	14.001	29.04	0.000
Error	8	3.857	0.482		
Total	19	107.722			

To test any one of these hypotheses, one could form ANOVA tables similar to that of Table 16.3. The mean squares for these hypotheses would be formed by dividing the reductions in sums of squares by their numbers of degrees of freedom, which are respectively $b - 1$, 1 and $a - 1$.

Table 16.8 is an ANOVA table for the reactor residence times under the assumption that the data in Table 16.7 were obtained from a randomized complete block design. This ANOVA table presents the analysis of the effects using the reductions in sums of squares. The results indicate that neither batches of raw materials nor the reactor effects are statistically significant once the adjustment for impurities is made. At this point the adjusted factor averages should be calculated as described in the appendix to this chapter. The formula for the adjusted factor averages is the same as equation (16.7), because the blocks are assumed to be random; however, the slope coefficient b is calculated from (16.A.5) in the appendix rather than from (16.A.2). The adjusted averages are

$$m_E = 30.71 + 2.07(-3.09 + 4.21) = 33.03,$$
$$m_P = 34.43 + 2.07(-5.33 + 4.21) = 32.11.$$

These averages are not significantly different from one another, as indicated by the nonsignificant F-value in Table 16.8. Any two adjusted averages can also be compared using t-statistics or confidence intervals. The standard error of the difference between adjusted averages is given by Equation (16.8) with

$$\mathrm{SS}_{xx} = \sum\sum(x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x}_{\bullet\bullet})^2.$$

To illustrate the dangers of using a sequential method of testing the model components, Table 16.9 was constructed from a sequential fitting of models by adding block, factor, and covariate components as described above. Note that the sequential F-tests now indicate that all the model components are

TABLE 16.9 Sequential ANCOVA for Residence Times in a Randomized Block Design

Source	df	SS	MS	F	p-Value
Model	11	103.865	9.442	19.59	0.000
Batches	9	20.672	2.297	4.76	0.019
Reactors	1	69.192	69.192	143.53	0.000
Impurity	1	14.001	14.001	29.04	0.000
Error	8	3.857	0.482		
Total	19	107.722			

statistically significant, contrary to the inferences drawn using the principle of reduction in error sums of squares, which is the correct analysis.

APPENDIX: CALCULATION OF ADJUSTED FACTOR AVERAGES

1. Completely Randomized Design, One Factor

Denote estimates of the model parameters in Equation (16.4) by m, a_i and b. The estimates m and a_i are not unique (see Section 6.2). Using the constraint $\sum \alpha_i = 0$, the estimates are

$$m = \bar{y}_{\bullet\bullet}, \qquad a_i = \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet} - b\bar{x}_i, \tag{16.A.1}$$

where b is the estimate of the slope parameter for the covariate in the full model. The estimate of β is unique, regardless of the solutions used for the other model parameters:

$$b = \frac{\sum_i \sum_j (x_{ij} - \bar{x}_{i\bullet})(y_{ij} - \bar{y}_{i\bullet})}{\sum_i \sum_j (x_{ij} - \bar{x}_{i\bullet})^2} \tag{16.A.2}$$

Using these estimates of the model parameters, the response averages for the factor levels are adjusted using the following expressions:

$$\begin{aligned} m_i &= m + a_i + b\bar{x}_{\bullet\bullet} \\ &= \bar{y}_{i\bullet} - b(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}). \end{aligned} \tag{16.A.3}$$

2. Randomized Block Design, One Factor

Adjusted factor averages for randomized block designs with fixed block effects are computed similarly to those for completely randomized designs. The estimates m, a_i, and c_j again are not unique. One set of solutions is

$$m = \bar{y}_{\bullet\bullet}, \qquad a_i = \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet} - b\bar{x}_{i\bullet},$$
$$c_j = \bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet} - b\bar{x}_{\bullet j}. \tag{16.A.4}$$

The estimate b is again unique, regardless of the solutions used for the other model parameters:

$$b = \frac{\sum_i \sum_j (x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j})(y_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j})}{\sum_i \sum_j (x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j})^2}. \tag{16.A.5}$$

Adjusted factor–level averages are given by

$$m_i = m + a_i + b\bar{x}_{\bullet\bullet},$$

which reduces to (16.A.3) with the estimated slope (16.A.5) inserted.

REFERENCES

Text References

Many of the texts listed at the end of Chapter 14 discuss ANCOVA techniques. The following texts provide fairly elementary coverage of ANCOVA techniques, including basic calculation formulas:

Anderson, R. L. and Bancroft, T. A. (1952). *Statistical Theory in Research*, New York: McGraw-Hill, Inc.

Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*, New York: John Wiley & Sons, Inc.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., and Nizam, A. (1998). *Applied Regression Analysis and Other Multivariate Methods*, Pacific Grove, CA: Duxbury Press.

Ostle, B. and Malone, L. C. (1988). *Statistics in Research*, Fourth Edition, Ames, IA: Iowa State University Press.

Winer, B. J. (1991). *Statistical Principles in Experimental Design*, Third Edition, New York: McGraw-Hill, Inc.

More theoretical coverage is provided in

Hinkelmann, K. and Kempthorne, O. (1994). *Design and Analysis of Experiments, Vol. 1: Introduction to Experimental Design*, New York: John Wiley & Sons, Inc.

Myers, R. H. and Milton, J. S. (1991). *A First Course in the Theory of Linear Statistical Models*, Boston: PWS-Kent Publishing Company.

Data References

The data for Exercises 16 (lib.stat.cmu.edu/datasets/strikes, prepared by Bruce Western) and 17 (lib.stat.cmu.edu/datasets/cars.data, prepared by Lawrence H. Cox.) were taken from the Statlib web site, July, 2001.

EXERCISES

1 An experiment was conducted to study the friction properties of lubricants. A key constituent of lubricants that is of interest to the researchers is the additive that is mixed with the base lubricant. In order to ascertain whether two competing additives produce a different effect on the friction properties of lubricants, several mixtures of a base lubricant and each of the additives were made. The mixtures of base lubricant cannot be made sufficiently uniform to ensure that all batches have identical physical properties. Consequently, the plastic viscosity, an important characteristic of the base lubricant that is related to its friction-reducing capability, was measured for each mixture prior to the addition of the additives. Analyze the data below to determine whether the additives differ in the mean friction measurements associated with each.

	Additive 1		Additive 2	
Batch	Plastic Viscosity	Friction Measurement	Plastic Viscosity	Friction Measurement
1	12	27.1	15	28.6
2	13	26.6	13	37.1
3	15	28.9	14	37.9
4	14	27.1	14	30.6
5	10	23.6	13	33.6
6	10	26.4	13	34.9
7	13	28.1	13	33.1
8	14	26.1	14	34.4
9	12	24.4	14	32.6
10	14	29.1	17	35.6

2 An experiment was conducted to assess the effects of three new high-nutrient feeds for producing accelerated weight gain in guinea pigs. Pigs from three strains were administered the feeds, and their weight gains y in grams were recorded. As is common in studies of this type, the initial weights x of the pigs were also recorded for use as a covariate. Analyze these data and determine which, if any, of the feeds produce higher mean weight gains than the others.

	Feed A		Feed B		Feed C	
Strain	x	y	x	y	x	y
1	152	301	160	298	162	336
2	177	337	168	310	181	379
3	114	241	125	234	122	258

3 An automobile manufacturer is interested in determining factors that are important to the stopping of vehicles on wet roads. Three factors were chosen for this study: car type, driver, and road surface material. Two cars, a light-weight and a medium-weight, were chosen. Two drivers were selected at random from a pool of experienced test car drivers. Two road surfaces were selected: asphalt and concrete. The response values measured are the stopping distances (feet) for vehicles traveling at a constant speed of 60 mph. All three factors are crossed in this study. Assume that no three-factor interaction exists. The data collected from this study are given below:

	Driver 1		Driver 2	
Car	**Asphalt**	**Concrete**	**Asphalt**	**Concrete**
1	194.1 (49)	197.3 (48)	185.0 (45)	175.7 (45)
	184.1 (47)	188.4 (46)	183.2 (45)	176.2 (51)
	189.0 (45)	187.6 (43)	184.0 (47)	180.9 (49)
2	188.7 (58)	186.3 (62)	191.8 (60)	182.9 (61)
	190.2 (60)	197.0 (63)	194.2 (60)	186.0 (63)
	189.4 (61)	197.5 (61)	190.5 (59)	184.3 (60)

The road surface temperature (°F) was measured in this study, and the values are given in the parentheses in the table above. Analyze this data set by using the road surface temperature as a covariate. Calculate adjusted factor–level means for the factors.

4 A rating procedure is being developed in an engine laboratory that will better quantify the subjective measurement of carbon deposits on engine parts. Two different methods of teaching this complex procedure to newly hired technicians are under investigation. Three senior technicians were chosen to instruct new technicians. Eighteen newly hired technicians were chosen for this study. Each technician rated an engine part before and after completing the training program with instruction from one of the two methods. The results are as follows:

	Method A		Method B	
Senior Technician	Before-Class Rating	After-Class Rating	Before-Class Rating	After-Class Rating
1	42	94	55	87
	59	81	69	94
	80	104	45	77
2	62	97	35	95
	47	99	50	92
	79	98	69	104
3	53	95	68	90
	30	77	37	85
	52	73	48	87

Perform an analysis on this data set to determine if there are significant differences in the teaching methods, the senior technicians, or the interaction between the methods and the senior technicians.

5 A study was conducted to determine the cold-startability of three different makes of diesel engines with a methanol test fuel. Fuel was run through each of three test engines, and the start times (seconds) were recorded. Five tests were run on each engine. It was hypothesized that the amount of time (minutes) the engine was left idle between the tests affected the resulting start times. Suppose that the methanol fuel was received in five 20-gallon drums and that samples were taken from each drum for testing in the three engines. Analyze these data to determine whether the starting times differ for the three engines.

Engine	Block (Drum)	Start Time	Idle Time
1	1	24	40
1	2	19	36
1	3	10	41
1	4	17	36
1	5	29	46
2	1	38	31
2	2	31	30
2	3	24	31
2	4	31	34
2	5	15	21
3	1	14	32
3	2	34	36
3	3	25	37
3	4	21	32
3	5	34	36

6 Compare an ANCOVA model in which the design factor has two levels with a two-factor regression model in which one of the predictors is an indicator variable. Algebraically relate the regression coefficient for the indicator variable to the effects parameters for the ANCOVA model. Algebraically relate the least-squares estimators of the regression parameters to the estimators of the ANCOVA model parameters.

7 Model the weight-gain data in Exercise 2 as a regression model using indicator variables. Compute the total of the sums of squares for the categorical variables using the principle of reduction in sums of squares. Compare this total with the corresponding total for the main effects sum of squares from the analysis of covariance.

8 Model the engine start-time data in Exercise 5 as a regression model using (a) values 1, 2, and 3 in a single predictor to designate the engines, and (b) two indicator variables to identify the engines. Calculate the least-squares estimates for the two models. How do the regression-coefficient estimates affect predictions for the two fits?

9 An engineer wants to evaluate whether three welding methods yield different weld strengths. The welding rods used in the tests differed by the amount of an additive that is known to enhance weld strength. The engineer decided to use the amount of additive as a covariate to get a more accurate look at weld method differences. The data collected are shown below. Analyze these data (use both graphical and numerical procedures) to determine if the weld strengths differ for the three welding methods.

Weld Method	Additive Amount	Weld Strength
Current	7	70.0
Current	8	71.6
Current	7	67.6
Current	9	78.8
Current	10	82.0
Current	13	90.0
Current	7	79.6
Current	9	81.2
Current	7	69.2
Current	5	68.4
XT	9	73.2
XT	6	74.0
XT	7	74.8
XT	10	72.4
XT	12	71.6
XT	13	73.2
XT	14	75.6

Weld Method	Additive Amount	Weld Strength
XT	11	74.0
XT	10	72.4
XT	9	72.4
CT	10	74.0
CT	11	74.8
CT	7	72.4
CT	8	69.2
CT	9	68.4
CT	6	73.2
CT	13	72.4
CT	14	73.2
CT	13	70.8
CT	11	73.2

10 Analyze the data in Exercise 5 of Chapter 6 using a regression analysis approach. Compare the results with the ANOVA approach. Annotate the ANOVA tables for the two analyses by mapping the entries from one to the other. Do you come to the same conclusions about the performance of the printer ribbons?

11 For the data in Exercise 12 of Chapter 6, fit a regression model that includes a week-by-transfer interaction and assess the significance of the terms. Interpret the coefficients for the eight interaction terms in the model. Display your results graphically and make recommendations on how to increase machine efficiency.

12 Exercise 6 of Chapter 6 describes an experiment and the resulting data. Is ANCOVA an appropriate statistical method for this data? Why or why not?

13 Fit a regression model to the data in Exercise 20 of Chapter 6 that has linear and two- and three-factor interaction terms. Are categorical variables needed for all three factors? Why or why not? Interpret (in words and graphically) the gap-by-material-by-diameter interaction. Specify the levels of the three factors to achieve high fragmentation.

14 For the scenario describe in Exercise 13 of Chapter 14 the following information was obtained about the gender of the bird. Does the maximum heart rate depend on the bird's gender?

Max Heart Rate	Distance from Bird	Gender
287	1816	Female
283	1588	Male
301	1410	Female
293	1192	Female

Max Heart Rate	Distance from Bird	Gender
298	1012	Female
295	915	Male
313	809	Male
311	603	Male
301	469	Male
314	398	Female
327	201	Female
369	0	Female

15 Suppose that in Exercise 9 of Chapter 14 the first 19 vehicles (going down the columns) are from one manufacturer, F, and the remaining 19 are from manufacturer G. Analyze these data to determine whether the NO_x emissions differ for the two automobile manufacturers.

16 The following data consist of annual observations on strike volume (defined as the days lost due to industrial strikes per 1000 hourly workers) and inflation for three countries. Determine if there are differences among the countries relative to the number of lost days. Which country has the least fallout from labor disputes?

Country	Year	Strike Volume	Inflation
1	1951	296	19.8
1	1952	397	17.2
1	1953	360	4.3
1	1954	300	0.7
1	1955	326	2.0
1	1956	352	6.3
1	1957	195	2.5
1	1958	133	1.3
1	1959	109	1.8
1	1960	208	3.8
1	1961	173	2.5
1	1962	142	−0.3
1	1963	158	0.5
1	1964	243	2.4
1	1965	211	4.0
1	1966	183	3.0
1	1967	168	3.2
1	1968	249	2.7
1	1969	439	2.9
1	1970	515	3.9

Country	Year	Strike Volume	Inflation
1	1971	645	6.1
1	1972	416	5.9
1	1973	527	9.5
1	1974	1252	15.1
1	1975	702	15.1
1	1976	760	13.5
1	1977	328	12.3
1	1978	421	7.9
1	1979	778	9.1
1	1980	633	10.1
1	1981	779	9.7
1	1982	370	11.2
1	1983	313	10.1
1	1984	241	4.0
1	1985	226	6.7
3	1951	242	9.6
3	1952	356	0.9
3	1953	170	−0.3
3	1954	182	1.3
3	1955	403	−0.5
3	1956	372	2.8
3	1957	1453	3.2
3	1958	114	1.3
3	1959	383	1.2
3	1960	129	0.3
3	1961	35	1.0
3	1962	100	1.4
3	1963	90	2.2
3	1964	157	4.2
3	1965	25	4.1
3	1966	189	4.2
3	1967	64	2.9
3	1968	129	2.7
3	1969	56	3.8
3	1970	483	3.9
3	1971	410	4.3
3	1972	117	5.5
3	1973	282	7.0
3	1974	184	12.7
3	1975	196	12.8

Country	Year	Strike Volume	Inflation
3	1976	291	9.2
3	1977	216	7.1
3	1978	325	4.5
3	1979	197	4.5
3	1980	69	6.7
5	1951	3	10.5
5	1952	3	3.8
5	1953	2	0.9
5	1954	17	0.0
5	1955	7	5.5
5	1956	789	6.0
5	1957	5	2.7
5	1958	6	0.7
5	1959	12	1.7
5	1960	41	1.3
5	1961	1486	3.5
5	1962	9	7.4
5	1963	15	6.1
5	1964	11	3.1
5	1965	143	5.5
5	1966	9	7.1
5	1967	6	8.2
5	1968	19	8.0
5	1969	31	3.5
5	1970	56	6.5
5	1971	11	5.9
5	1972	11	6.6
5	1973	2001	9.3
5	1974	94	15.3
5	1975	52	9.6
5	1976	107	9.0
5	1977	115	11.1
5	1978	64	10.0
5	1979	84	9.6
5	1980	90	12.3
5	1981	319	11.7
5	1982	45	10.1
5	1983	38	6.9
5	1984	62	6.3
5	1985	1056	4.7

17 In the early 1970s, data were collected on miles per gallon (MPG) and vehicle weight for cars from three different countries (the coded levels are 1 = American, 2 = European, and 3 = Japanese). Are there differences among the vehicles that come from the three producing nations? If so, which country produces the most fuel-efficient vehicle? If you had the opportunity to augment this data with additional observations, what characteristics would you want the additional data to have?

MPG	Weight	Origin
18	3504	1
15	3693	1
18	3436	1
16	3433	1
17	3449	1
15	4341	1
14	4354	1
14	4312	1
14	4425	1
15	3850	1
15	3563	1
14	3609	1
15	3761	1
14	3086	1
24	2372	3
22	2833	1
18	2774	1
21	2587	1
27	2130	3
26	1835	2
25	2672	2
24	2430	2
25	2375	2
26	2234	2
21	2648	1
10	4615	1
10	4376	1
11	4382	1
9	4732	1
27	2130	3
28	2264	1

MPG	Weight	Origin
25	2228	3
25	2046	1
19	2634	1
16	3439	1
17	3329	1
19	3302	1
18	3288	1
14	4209	1
14	4464	1
14	4154	1
14	4096	1
12	4955	1
13	4746	1
13	5140	1
18	2962	1
22	2408	1
19	3282	1
18	3139	1
23	2220	1

CHAPTER 17

Designs and Analyses for Fitting Response Surfaces

This chapter contains a discussion of designs and analyses that are useful when an experimenter wants to explore an unknown response surface. Topics to be addressed include:

- *uses of response-surface methodology,*
- *methods for locating an appropriate experimental region,*
- *description of designs for fitting response surfaces,*
- *regression methods for characterizing response surfaces,*
- *using fitted response surfaces to identify factor levels that optimize a response, and*
- *analysis of quality improvement data from designs with crossed or combined arrays*

A common goal in many types of experimentation is to characterize the relationship between a response and a set of quantitative factors of interest to the researcher. This is accomplished by constructing a model that describes the response over the applicable ranges of the factors of interest. In many industrial applications the fitted model is referred to as a *response surface* because the response can then be graphed as a curve in one dimension (one factor of interest) or a surface in two dimensions (two factors of interest). The response surface can be explored to determine important characteristics such as optimum operating conditions (that is, factor levels that produce the maximum or minimum estimated response), or relevant tradeoffs when there are multiple responses. The objective of this chapter is to present the design strategies for and the analysis of these types of response functions.

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

17.1 USES OF RESPONSE-SURFACE METHODOLOGY

Certain types of scientific problems involve expressing a response variable, such as the viscosity of a fluid, as an empirical function of one or more quantitative factors, such as reaction time and reaction temperature. This is accomplished using a *response function* to model the relationship:

$$\text{Viscosity} = f(\text{reaction time, reaction temperature}).$$

A general form of this type of response function is

$$y = f(x_1, x_2, \ldots, x_k), \tag{17.1}$$

where y is the response and $x_1, x_2, \ldots, x_k$ are quantitative levels of the factors of interest. Knowledge of the form of the function, f, often found by fitting models to data obtained from designed experiments, allows one to both summarize the results of the experiment and predict the response for values of the quantitative factors. The function f defines the response surface (see Exhibit 17.1).

EXHIBIT 17.1

Response Surface. A response surface is the geometric representation obtained when a response variable is plotted as a function of one or more quantitative factors.

Designing experiments to study or fit response surfaces is important for several reasons, including the following:

- the response function is characterized in a region of interest to the experimenter,
- statistical inferences can be made on the sensitivity of the response to the factors of interest,
- factor levels can be determined for which the response variable is optimum (e.g., maximum or minimum), and
- factor levels can be determined that simultaneously optimize several responses; if simultaneous optimization is not possible, tradeoffs are readily apparent.

Each of these uses will now be discussed.

A response surface can have various shapes, depending on the form of the response function in equation (17.1). For example, if it is a second-order (that

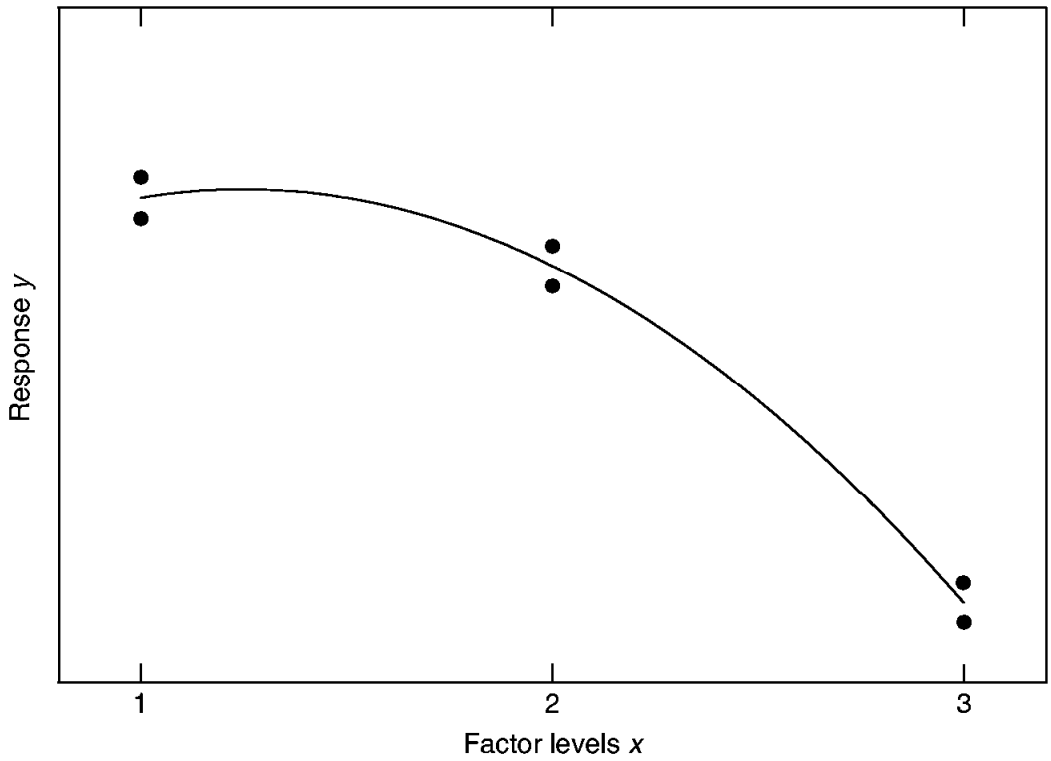


Figure 17.1 Quadratic response surface in one factor.

is, quadratic) function of only one factor, the surface might look similar to the curve shown in Figure 17.1. The plotted points represent pairs of observed responses (y) for each of three quantitative values of the factor (x). The fitted model, represented by the smooth curve, characterizes the response surface and identifies where the maximum response is obtained.

One experimental strategy to use when a response is believed to be a quadratic function of a single factor is to select a three-level, completely randomized design with repeat tests. Note that using a two-level design would only allow a straight line to be fitted. Repeat tests are included to provide an estimate of the uncontrolled experimental-error variation. The plotted points in Figure 17.1 depict the location of the design points and the corresponding responses for a typical experiment.

When the response is a function of two factors, the experimental situation can still be depicted graphically. As before, assume the response function is a second-order polynomial. In this situation the response surface might be graphed similarly to the one shown in Figure 17.2. An experimental strategy to use in characterizing this surface might consist of a completely randomized design in two factors, each at three levels. Typical responses at these design points are illustrated on the graph (without repeats).

An alternative approach to plotting the two-factor response surface is to plot contours (see Exhibit 17.2) of constant response as a function of the two

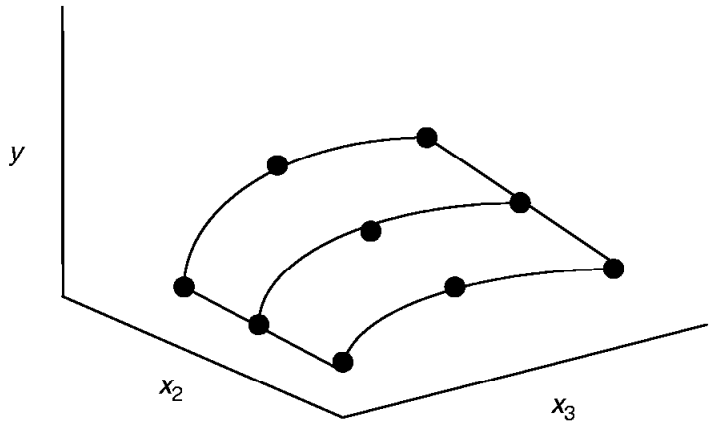


Figure 17.2 Quadratic response surface in two factors.

factors. These are similar to the contours of equal elevation on a topographical map.

EXHIBIT 17.2

Contour Plot. A contour plot is a series of lines or curves that identify values of the factors for which the response is constant. Curves for several (usually equally spaced) values of the response are plotted.

Contour plots can be constructed in several ways. If the response function is a simple enough function of the two factors, it can be solved for one of the factors as a function of the response and the other factor. Fixing the response at a value then enables one to plot the contour having that response value by plotting one of the factors in terms of the other one.

When the response function is complicated, a more direct solution is to calculate the values of the response variable for a grid of values of the two factors. Instead of plotting points, one can then plot the numerical values of the response on a graph as a function of the two factors; that is, the two axes represent the factor values and the numerical value of a calculated response is placed on the grid at the intersection of the two factor values used to calculate it. Contours can then be approximated by interpolating between values of the response variable.

Figure 17.3 illustrates several common forms of contour plots from two-factor response surfaces. Each of the plots shows curves for values of the response in increments of 10 units. Figure 17.3*a* illustrates the shape of contours when a conical or symmetrical "mound-shaped" surface is located in the center of the experimental region. Figure 17.3*b* depicts the identification of a *stationary ridge*, a sequence of flat contours with a maximum along one of the contours. With response surfaces such as this, there is a wide range

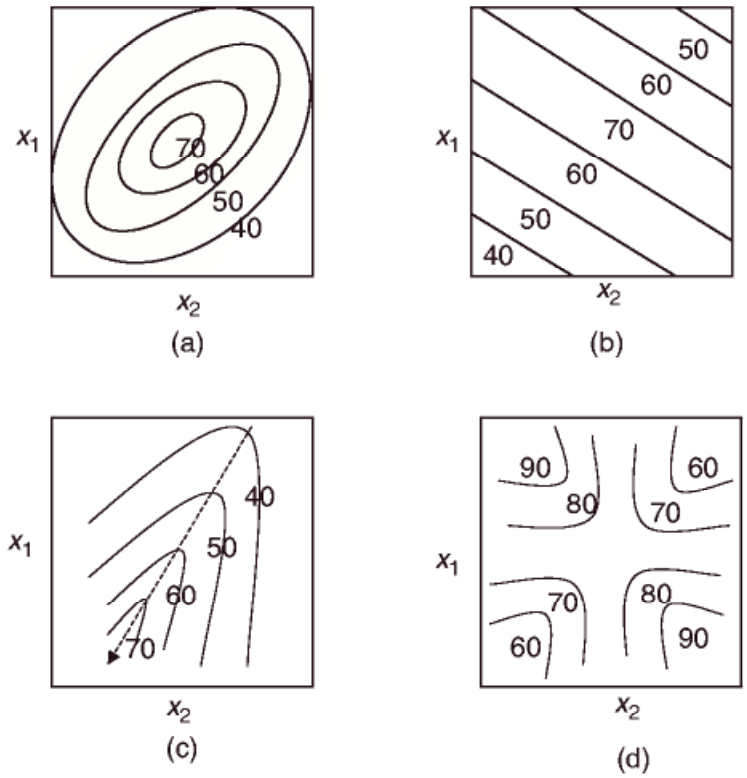


Figure 17.3 Response-surface contours in two factors: (a) "mound-shaped" maximum, (b) stationary ridge, (c) rising ridge, and (d) saddle.

of values of the two factors that produce an optimal response. Sometimes fitted response surfaces look like Figure 17.3*b*, but the response values are all increasing in one direction; that is, there is no contour with a maximum response. This indicates that the fitted surface is a plane and that the region of optimum response is distant from the experimental region.

Figure 17.3*c* shows a *rising ridge*, such that the location of a minimum response is just outside the experimental region. A path up the ridge line (dashed in Figure 17.3*c*) leads to the optimum response. Figure 17.3*d* depicts a *saddle*, on which one can either decrease or increase the response by selecting factor levels along a 45° or a 135° line, respectively, from the center of the region.

The variety of response-surface contours shown in Figure 17.3 demonstrate why response-surface characterization is one primary reason for many studies of responses and the factors that influence them. An equally important reason for fitting and studying response surfaces is the determination of the sensitivity of the response to the various factors. If the contours in Figure 17.3*b* were more horizontal it would indicate graphically that the response was very sensitive to the first factor (x_1) and relatively insensitive to the second one (x_2). This is because contours that are almost horizontal indicate that

the horizontal factor could change greatly and the response would remain fairly constant.

Polynomials models of first and second order (that is, linear and quadratic equations with interactions) are routinely used to model the response surface in Equation (17.1). Experimental designs are selected that ensure that information about the response is gathered in an efficient manner in the regions of interest (Sections 17.2 and 17.3). Fitted response surfaces or their contours are graphed to illustrate the changes in the response as a function of the various factors in the study. Statistical inference techniques (Section 17.4) can be used to assess the importance of the individual factors, the appropriateness of their functional form, and the sensitivity of the response to each factor.

A third use of response-surface experimentation is to find the factor levels that provide the optimum response. This could be either a maximum or a minimum response. For example, in a study on Sulphlex binders used in developing synthetic asphalt, a response-surface approach was used to determine the reaction temperature and reaction time that would yield a binder with the highest possible viscosity. Using a 3^2 factorial experiment with several repeat tests, a second-order response surface was fitted to the observed viscosities, and the contour plot given in Figure 17.4 was obtained. It can be seen that the highest viscosity values occurred at the higher reaction times and at the lower temperatures. The contours suggest that even higher viscosities could

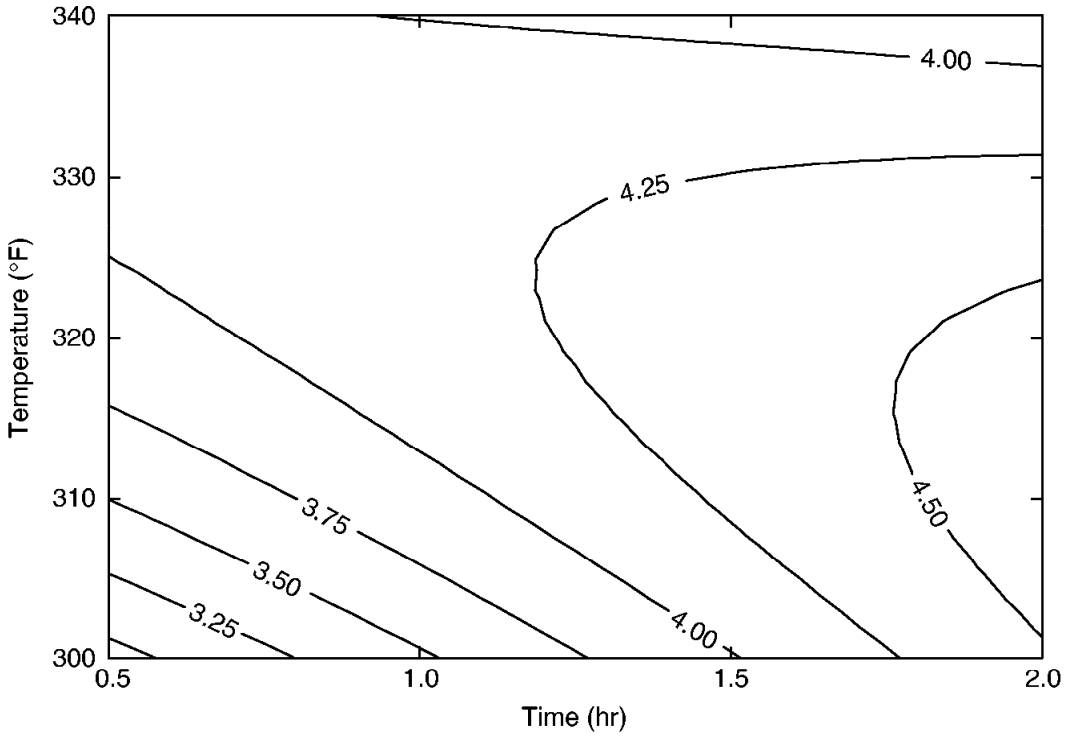


Figure 17.4 Response-surface contours for Sulphlex study: contours of constant viscosity.

be obtained for moderate temperatures with longer reaction times. This was the direction taken in further experimentation.

A fourth use of response-surface design is to find factor regions that produce the best combinations of several different responses. To illustrate this use, consider an experiment that was run to study the sensitivity of an analytical method used to assay blood serum for the presence of the therapeutic antibiotic drug gentamicin. The purpose was to determine the volumes of two reagents, a particle reagent and an antibody, used in an analytical assay test pack that were necessary to maintain uniform performance.

Four responses were measured as a function of the volumes of these two reagents: (a) separation, (b) ratio, (c) blank, and (d) final blank absorbance. The experimental region of interest included volume combinations of the reagents for which separation was greater than 200 milliabsorbance units per minute (ma/min), ratio was between 0.4 and 0.5, blank rate exceeded 300 ma/min, and final blank absorbance was less than 0.8 absorbance units (800 ma).

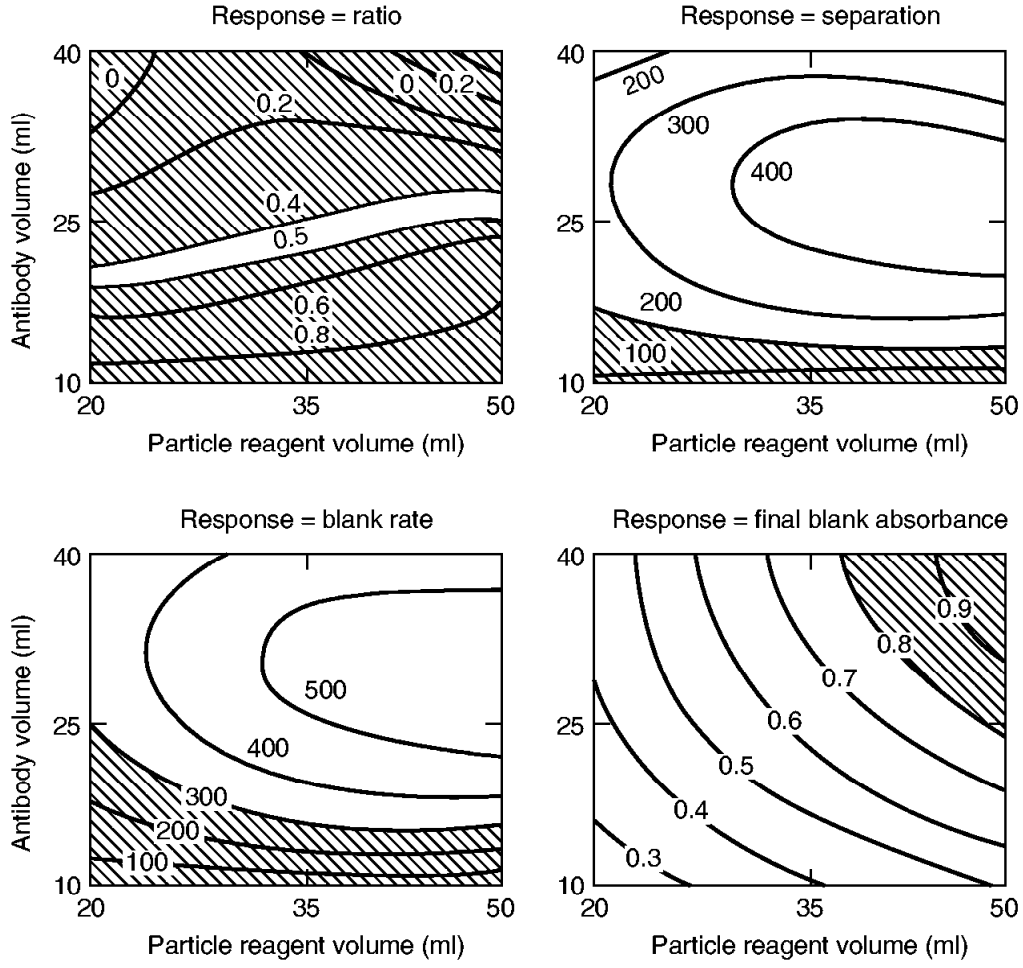


Figure 17.5 Four response contours for gentamicin study (shaded regions are unacceptable.).

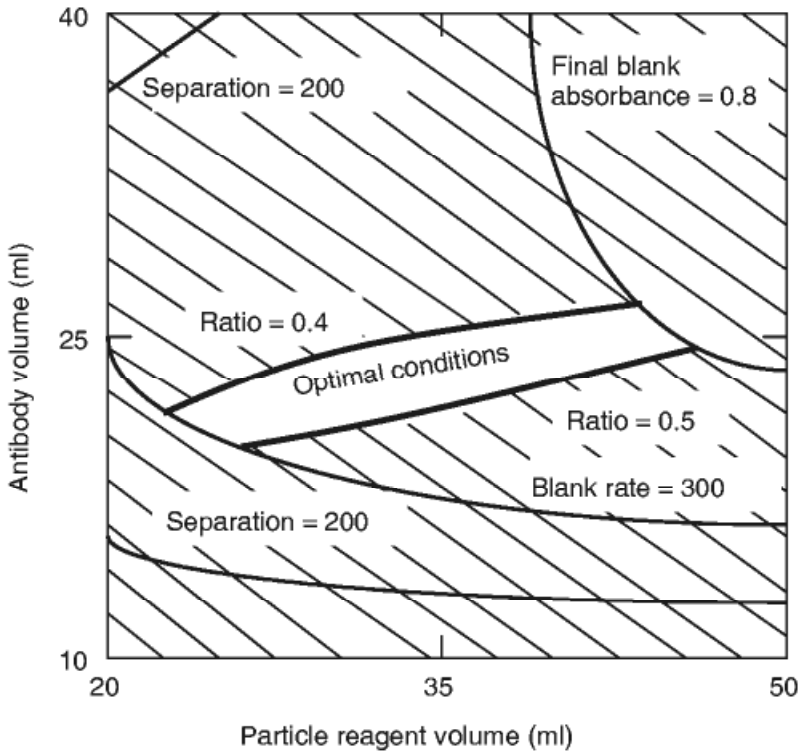


Figure 17.6 Simultaneous optimization of four responses: gentamicin study.

The purpose of the response-surface investigation was to find the particle reagent and antibody volumes that maximized the separation and blank rates subject to the above restrictions. A 3^2 factorial experiment with three repeat tests was used to develop the response-surface contours shown in Figure 17.5. These contours are overlaid in Figure 17.6. The region of optimal conditions for the two factors is indicated in Figure 17.6 by the unshaded region.

In this example there exists a portion of the experimental region for which all four responses are optimized. There is no guarantee that this will always occur. Often there will be regions in which some responses are optimized, but others are not. Constructing contour plots similar to Figure 17.5 will allow reasonable compromises to be made.

17.2 LOCATING AN APPROPRIATE EXPERIMENTAL REGION

As noted in Section 17.1, one of the uses of a response-surface design is to identify values of the factors that produce optimum or near optimum values of the response. This often is accomplished through a series of experiments whereby one searches for the region of the optimum response. One method of experimentation that is popular in this process is the one-factor-at-a-time strategy. Each factor is individually increased or decreased in an effort to find

the maximum response. The combination of these optimum factor levels is then chosen as the conditions for obtaining the overall maximum. Unfortunately, as discussed in Section 4.2, this method frequently fails to locate the region of the optimum response, since the procedure does not take account of any joint effects of the factors on the response.

A preferred alternative procedure involves up to three stages, the second of which is a sequence of test runs conducted along a path of *steepest ascent.* In the initial stage of this procedure a simple experiment is conducted over some small area of the region of interest. Usually this initial experiment does not include the extremes of the experimental region, but it may include the extremes of a desirable subregion. A first-order model is then fit to the data, and the resultant equation is used to determine the direction to move toward the surface optimum. Observations are sequentially taken along this path of steepest ascent until the vicinity of the optimum of the surface is located. A more comprehensive experiment is then run in this new region, and usually a higher-order model is fit in order to characterize the surface. The procedure is outlined in Exhibit 17.3.

EXHIBIT 17.3 LOCATING AN OPTIMUM RESPONSE

1. Conduct a small-scale experiment (often a two-level fractional factorial) in a region of the factor space that is believed to include the optimum response.
2. Fit a model (usually first-order) to the data collected in step 1.
3. Use the fitted model (equation) from step 2 to find the direction of the steepest ascent or descent in the response. The path is perpendicular to the contour lines. If the model is nonlinear and is a function of three or more factors, a canonical analysis (see the appendix to this chapter) may be needed to determine the direction of steepest ascent.
4. Conduct a series of test runs along the path determined in step 3 until the increase or decrease in the response becomes small or reverses.
5. Repeat steps 1–4 in the new region of the factors, if needed, until the region of the optimum response is located.
6. When the factor region containing the optimum response is located, conduct a more extensive experiment that will permit the fitting of higher-order models so the curvature of the response surface can be adequately approximated.

Complete or fractional factorial experiments are the most common initial experiments in the study of response surfaces. We shall use data from the Sulphlex study to illustrate the method of steepest ascent. Figure 17.7 shows four design points (a 2^2 factorial) taken from the Sulphlex study introduced

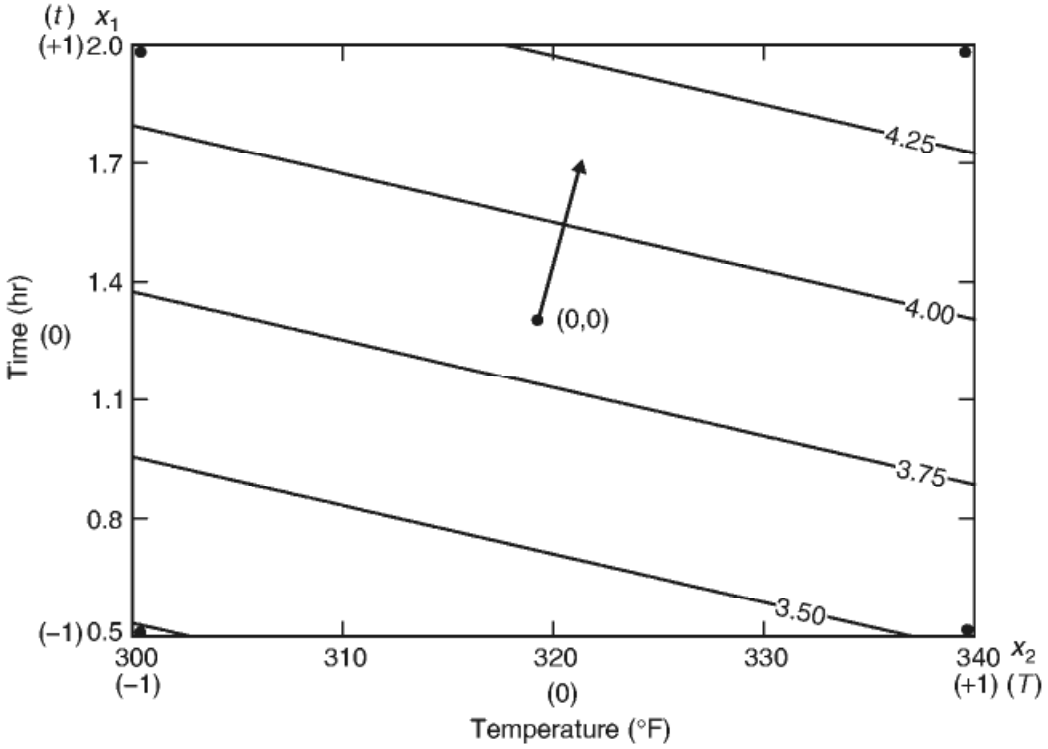


Figure 17.7 Response-surface contours from a 2^2 factorial experiment for the Sulphlex study.

in the previous section. Linear contours for response values from 3.25 to 4.25 in increments of 0.25 are shown.

The path of steepest ascent can be determined graphically as indicated in Figure 17.7. The direction indicated by the arrow, perpendicular to the contour lines, is the direction in which additional test runs will be taken if the goal is to maximize viscosity. Ordinarily, test runs are equally spaced along the path of steepest ascent.

The path of steepest ascent is usually determined from the center of the experimental region. The determination can be facilitated by coding the region so that the lower and upper levels of each factor are -1 and $+1$, respectively. To do so for each factor, use the coding technique given in Exhibit 17.4. Using this technique, the coded levels of the two factors, x_1 = time and x_2 = temperature, in the Sulphlex study are

$$t = \frac{x_1 - 1.25}{0.75}, \qquad T = \frac{x_2 - 320}{20}.$$

As indicated on the axes in Figure 17.7, the contours can be plotted using the coded or the raw factor levels.

EXHIBIT 17.4 TWO-LEVEL FACTOR CODING

For each factor:

1. Determine the average of the two factor–level values. Denote it by AVG.
2. Determine the midrange (MID) of the factor levels:

$$\text{MID} = \frac{\text{upper level} - \text{lower level}}{2}.$$

3. Code the factor levels:

$$\text{coded level} = \frac{\text{level} - \text{AVG}}{\text{MID}}.$$

Because the contours in Figure 17.7 are straight lines, the path of steepest ascent is a line from the center of the region that is perpendicular to the contours (a general technique for finding the path of steepest ascent from the fitted model is given in the appendix to this chapter). Perpendicular lines to these contours have slopes opposite in sign and inverse in magnitude to those in the figure; that is, the lines

$$t = bT \quad \text{and} \quad t = cT$$

are perpendicular when $c = -1/b$.

All the contour lines in Figure 17.7 have slopes equal to -0.33; consequently, the perpendiculars to these lines have slopes equal to $+3.00$. Thus the path of steepest ascent goes from the center of the region along the line $t = 3.00T$, as indicated in Figure 17.7. Table 17.1 shows the path of steepest ascent in both the coded and the original scales of the factors. These values were determined (arbitrarily) by incrementing the coded temperature factor by 0.5. Other increments can be used if they are deemed more reasonable by the experimenter. Note that one need not start the test runs at the center of the current design; rather, one could begin at the first combination (e.g., $t = 1.5, T = 0.5$) that is outside the region shown in Figure 17.7.

The objective now is to advance along the path $t = 3.00T$ until the maximum of the response is obtained. As experimentation continues along the path of steepest ascent, the increases in the response will become smaller; eventually the responses will decrease. One should then conduct a new set of experimental test runs to ensure that the optimum experimental region has been located. The new test runs should be selected so that if the optimum region has been identified, additional runs can be added to these to allow the fitting of higher-order models. The inclusion of repeat tests with these new test runs will allow an assessment of the adequacy of the fit.

A good design for this latter purpose is again a two-level factorial with the levels chosen so that a third level for each factor can be added if it is

TABLE 17.1 Path of Steepest Ascent for Sulphlex Data*

Coded Factors		Original Factors	
Time t	Temperature T	Time x_1(hr)	Temperature x_2(°C)
0.0	0.0	1.25	320
1.5	0.5	2.38	330
3.0	1.0	3.50	340
4.5	1.5	4.62	350
6.0	2.0	5.75	360
7.5	2.5	6.88	370
.	.	.	.
.	.	.	.
.	.	.	.

*Increments of 0.5 Units in Coded Temperature.

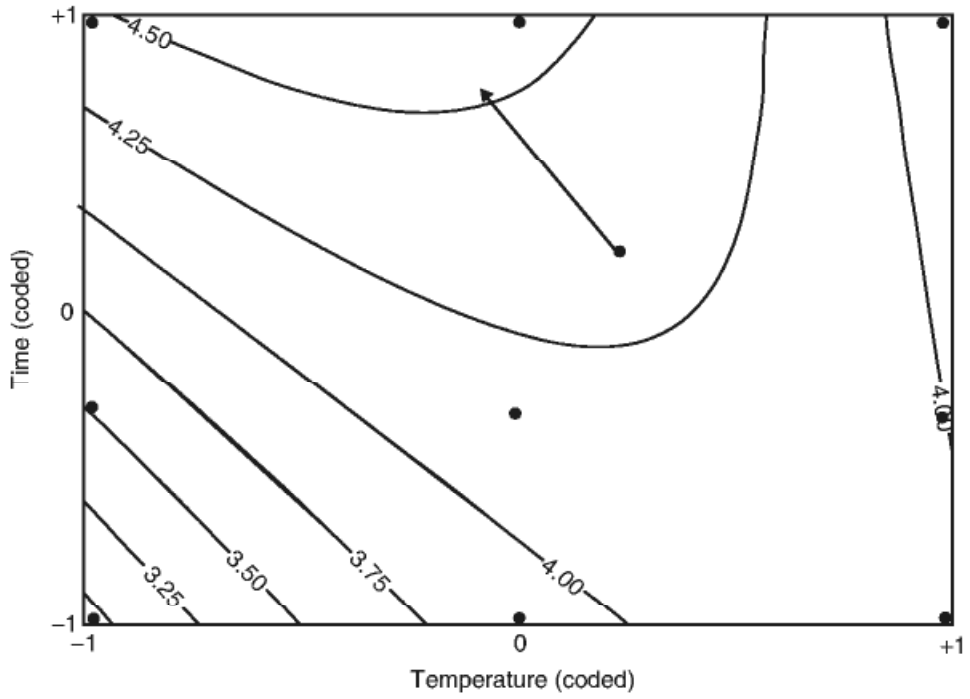


Figure 17.8 Response-surface contours from a 3^2 factorial experiment: Sulphlex study.

determined that the optimum experimental region has been located. Repeat tests can be made at the center of this new experimental region.

For a large number of factors, the initial experiment in the steepest-ascent procedure may be a two-level fractional factorial. When using two-level complete or fractional factorial experiments, caution should be exercised in moving along the path of steepest ascent. A biasing effect may lead one along an

incorrect path if higher-order effects are substantial. This is especially true when nearing the region of the optimum response, where curvature and joint factor effects are likely to be most pronounced.

Figure 17.8 shows response contours for the coded Sulphlex factors after fitting interaction and quadratic terms from a complete 3^2 design. Because short reaction times were of interest in this experiment, the middle level of time was chosen closer to the lower level than to the upper one. If a two-level factorial experiment had been run in this region, the response contours would have been those in Figure 17.7. By using a three-level factorial experiment and fitting a model including terms for curvature, the path of steepest ascent is as indicated in the figure, quite different from that shown in Figure 17.7.

17.3 DESIGNS FOR FITTING RESPONSE SURFACES

Several different types of designs are useful when one is attempting to fit a response surface. Complete and fractional factorial experiments in completely randomized designs are extremely useful when one is exploring the factor space in order to identify the region where the optimum response is located. As stressed at the end of the last section, two-level factorials are highly efficient but must be used with some caution. They allow the fitting of only first-order models with or without interaction terms and cannot detect curvature.

When one has located the region of the optimum response, curvature can be pronounced. Three-level factorial experiments are often conducted in order to fit such response surfaces. The nine equally spaced factor–level combinations for a 3^2 experiment are shown in Table 17.2. The coded factor levels are designated as low (-1), middle (0), and high $(+1)$.

The gentamicin study described in Section 17.1 uses a 3^2 factorial. The antibody volume (factor A) has three levels (10, 25, and 40 ml), and the particle-reagent volume (factor B) has three levels (20, 35, and 50 ml). Three repeat tests were taken at each of the nine design points.

As the number of factors increases, the 3^k factorials become inefficient and impractical. These experiments need large numbers of observations; for example, $3^5 = 243$ and $3^{10} = 59,049$. Further, these designs do not give equal precision for fitted responses at points (factor–level combinations) that are at equal distances from the center of the factor space. A design that has this property is termed a *rotatable* design (see Exhibit 17.5). Rotatability is a desirable property for response-surface models because prior to the collection of data and the fitting of the response surface, the orientation of the design with respect to the surface is unknown. Thus, the exploration of the response surface is dependent on the orientation of the design. In particular, procedures such as the method of steepest ascent, which utilize the fitted response surface, can be jeopardized if some estimated responses are less precise than others.

TABLE 17.2 Typical Factor Levels for a 3^2 Factorial

Coded Factors		Original Factors	
Factor *A*	Factor *B*	Factor *A*	Factor *B*
−1	−1	Low	Low
−1	0	Low	Middle
−1	+1	Low	High
0	−1	Middle	Low
0	0	Middle	Middle
0	+1	Middle	High
+1	−1	High	Low
+1	0	High	Middle
+1	+1	High	High

EXHIBIT 17.5

Rotatable Design. When fitting specified response-surface models, a design is rotatable if fitted models estimate the response with equal precision at all points in the factor space that are equidistant from the center of the design.

In general, rotatable designs can be constructed from equally spaced points on circles or spheres. If one is interested in constructing rotatable designs for use with models in which only linear terms, with or without interactions, are to be included, points in regular polygons about the center of the experimental region can be used. For example, in a two-factor experiment the vertices of a square (two-level factorial), pentagon, hexagon, or octagon could be used. These are two-dimensional polygons centered at the origin with vertices on a circle. Coordinates of such designs are multiples of those shown in Table 17.3. Ordinarily, in such designs, repeat observations are taken at the origin. If there are to be repeat tests conducted at the vertices of the design, there must be an equal number of repeat tests at all the design points for the design to be rotatable.

For more than two factors, design points should lie on a sphere, or a hypersphere in four or more dimensions. The design points must also form a regular geometric figure such as a cube for the design to be rotatable. All 2^k complete factorials are rotatable, but 3^k factorials are not. Fortunately, there are classes of designs for two or more factors that can be used in place of 3^k factorials for fitting second-order polynomials to response surfaces.

Two designs that make more efficient use of the experimental units or test runs than the 3^k factorial experiments are the central composite design and the Box–Behnken design. Both of these designs are fractions of the 3^k factorials,

TABLE 17.3 Some Rotatable Designs for Two Factors*

Factor–Level Combination	Square		Pentagon		Hexagon		Octagon	
	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
1	−1	−1	1	0	1	0	1	0
2	−1	1	0.309	0.951	0.500	0.866	0.707	0.707
3	1	−1	−0.809	0.588	−0.500	0.866	0	1
4	1	1	−0.809	−0.588	−1	0	−0.707	0.707
5			0.309	−0.951	−0.500	−0.866	−1	0
6					0.500	−0.866	0.707	−0.707
7							0	−1
8							−0.707	−0.707

*Coordinates are cosines and sines of multiples of $360/k$ degrees, where k is the number of points in the design.

but they only require enough observations to estimate the second-order effects of the response surface. Both of these designs can be made rotatable, or approximately so.

17.3.1 Central Composite Design

The central composite design is constructed following the steps given in Exhibit 17.6

EXHIBIT 17.6 CENTRAL COMPOSITE DESIGN

1. Construct a complete or fractional 2^k factorial layout, depending on the need for efficiency and the ability to ignore interaction effects.
2. Add $2k$ *axial*, or *star*, points along the coordinate axes. Each pair of star points is denoted, using coded levels, as follows:

$$
\begin{aligned}
&(\pm a, \quad 0, \quad 0, \cdots, \quad 0),\\
&(0, \quad \pm a, \quad 0, \cdots, \quad 0),\\
&\qquad\qquad \cdots\\
&(0, \quad 0, \quad 0, \cdots, \quad \pm a),
\end{aligned}
$$

   where a is a constant, which can be chosen to make the design rotatable or to satisfy some other desirable property.
3. Add m repeat observations at the design center:

$$(0, \quad 0, \quad 0, \ldots, \quad 0).$$

4. Randomize the assignment of factor–level combinations to the experimental units or to the run sequence, whichever is appropriate.

The total number of test runs in a central composite design based on a complete 2^k factorial is $n = 2^k + 2k + m$. This count usually is less than 3^k, so that fewer observations are required than in a 3^k factorial. The central composite design can be made to be rotatable by choosing $a = F^{1/4}$, where F is the number of factorial points (e.g., $F = 2^k$ when a complete factorial is used). An illustration of the three-factor central composite design based on a complete 2^k factorial is given in Figure 17.9.

We now present an example of a central composite design that was used to evaluate the performance of a method for determining the level of the hormone thyroxine in blood serum. After initial experimentation, it was found that the key factors in this study were two reagent concentrations and the serum sample volume. The two reagents included an enzyme and a thyroxine conjugate, which acts as an enzyme inhibitor.

A rotatable central composite design could be used for this experiment. The design would consist of the eight corner points of the 2^3 cube, the six star points, and m center points. The star points would have $a = 8^{1/4} = 1.68$. If three center points were selected, the design (prior to randomization) would be as given in Table 17.4.

The actual design used is illustrated in Figure 17.10. It is a special form of the central composite design in that the star points are on the face of the cube formed from the 2^3 factorial. Thus, $a = 1$ and the design is called a *face-centered cube design.* The face-centered cubic design was chosen over

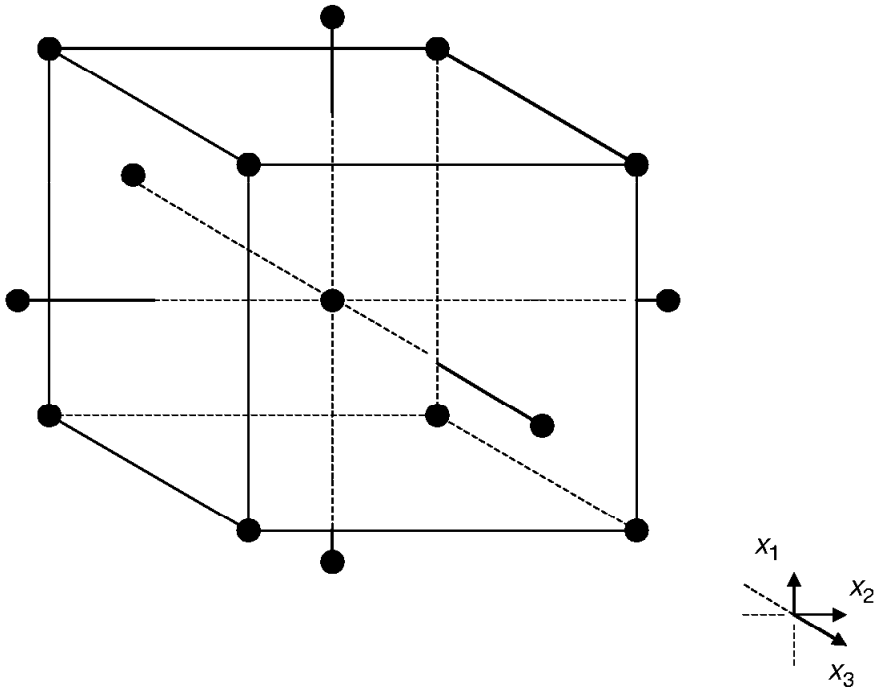


Figure 17.9 Central composite design in three factors.

TABLE 17.4 Rotatable Central Composite Design for Thyroxine Study*

	Coded Factor Levels		
Run	Enzyme	Conjugate	Sample Volume
1	−1	−1	−1
2	−1	−1	+1
3	−1	+1	−1
4	−1	+1	+1
5	+1	−1	−1
6	+1	−1	+1
7	+1	+1	−1
8	+1	+1	+1
9	−1.68	0	0
10	+1.68	0	0
11	0	−1.68	0
12	0	+1.68	0
13	0	0	−1.68
14	0	0	+1.68
15	0	0	0
16	0	0	0
17	0	0	0

*Nonrandomized.

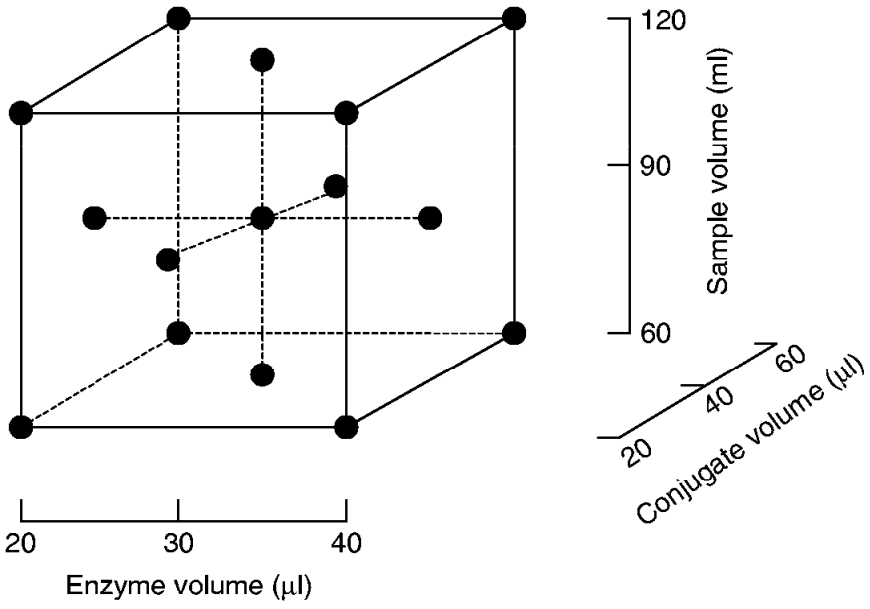


Figure 17.10 Face-centered central composite design for thyroxine study.

alternatives such as a rotatable design because the face-centered design only uses three levels of each factor, whereas other central composite designs would require five levels of each factor $(0, \pm 1, \pm a)$. Having three levels instead of five was cited as desirable because it reduced the preparation time and lessened the potential for mistakes in preparing the test serum. Three replicates were taken at the design center, so that the total number of observations was $n = 8 + 6 + 3 = 17$. This is slightly over half the number of observations that would be required for a three-level factorial without repeats ($3^3 = 27$).

17.3.2 Box–Behnken Design

The Box–Behnken design is another alternative to the 3^k factorial. The designs are formed by combining 2^k factorials with incomplete block designs. The result is a design that makes efficient use of the experimental units and is also rotatable or nearly so. Useful Box–Behnken designs are listed in Table 17A.1 in the appendix to this chapter. Sequences of ± 1 signs in the rows of Table 17A.1 mean that each level of the factors is to be run with each level of all the other factors in that row of the design that have ± 1.

The three-factor Box–Behnken design is shown in Figure 17.11. The equivalent design for the thyroxine study is given in Table 17.5 (without randomization). The total number of test runs needed for this design is 15, fewer than the 17 required for a central composite design with the same number of repeats and the 27 required for a 3^3 factorial without repeats. Three repeat tests are included at the center of the design, as was done with the central composite design. The Box–Behnken design in Table 17.5 only requires three levels of

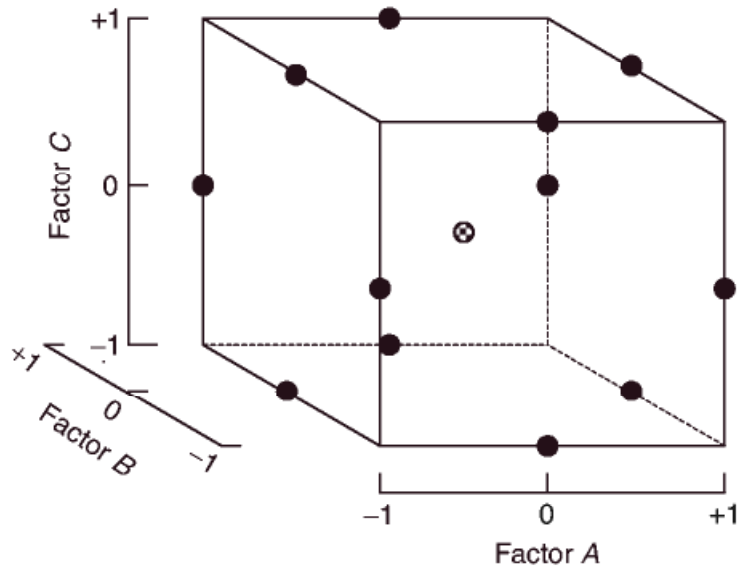


Figure 17.11 Three factor Box–Behnken design (coded factor levels).

TABLE 17.5 Box–Behnken Design for Thyroxine Study*

	Coded Factor Levels		
Run	Enzyme	Conjugate	Sample Volume
1	−1	−1	0
2	−1	+1	0
3	+1	−1	0
4	+1	+1	0
5	−1	0	−1
6	−1	0	+1
7	+1	0	−1
8	+1	0	+1
9	0	−1	−1
10	0	−1	+1
11	0	+1	−1
12	0	+1	+1
13	0	0	0
14	0	0	0
15	0	0	0

*Nonrandomized.

each factor. It is preferable to the face-centered central composite design not only because it requires fewer test runs but also because it is rotatable.

As indicated in Figure 17.11, Box–Behnken designs do not contain any points at the extremes of the cubic region created by the two-level factorial. All of the design points are either on a sphere or at the center of a sphere. This design is advantageous when the points on one or more corners of the cube represent factor–level combinations that are prohibitively expensive or impossible to test due to physical constraints on the experimentation.

A careful examination of the three-level face-centered central composite design (Figure 17.10) and the three-level Box–Behnken design (Figure 17.11) reveals an interesting geometric pattern. When the two designs are overlaid, one obtains a complete 3^3 factorial as shown in Figure 17.12. Thus, these alternative designs are simply fractional parts of the three-level factorial. As such, both provide good results for a wide range of practical problems.

17.3.3 Some Additional Designs

A variety of additional experimental designs are useful when fitting response surfaces. Myers and Montgomery (1995) provide a helpful set of properties

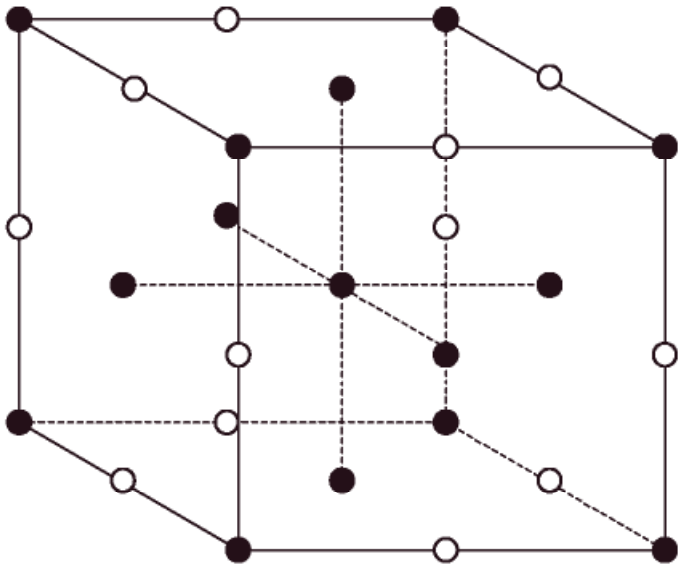


Figure 17.12 3^3 factorial layout: combination of Box–Behnken (○) and face-centered central composite (●) designs.

that should be considered when choosing such designs. These are listed in Exhibit 17.7. The designs discussed in Sections 17.3.1 and 17.3.2 meet most of these criteria. Some other alternative designs are briefly discussed below. More details on these designs can be found in the many references contained at the end of the chapter.

EXHIBIT 17.7 PREFERRED PROPERTIES OF RESPONSE SURFACE DESIGNS

Result in a good fit of the model to the data.
Give sufficient information to allow a test for lack of fit.
Allow models of increasing order to be constructed sequentially.
Provide an estimate of "pure" experimental error.
Be robust to the presence of outliers in the data.
Be robust to errors in control of design levels.
Be cost effective.
Allow experiments to be done in blocks.
Provide a check on the homogeneous variance assumption.
Provide a good distribution of $v(x) = \text{Var}[\hat{y}(x)]/\sigma^2$.

Designs for process improvement, such as *robust parameter* designs or *integrated* designs, are types of design that are closely connected to the fitting of a response surface. These were discussed in Section 12.2, and several examples were included. These designs are based on fractional factorial experiments or

the response surface designs described previously, but they focus on quality improvement issues.

Mixture designs are used when the response of interest is influenced not by the actual amounts of the factors (that is, components of the mixture) but only by their relative amounts. In a mixture experiment the factor space of allowable component values is constrained, and the particular combination of the components in a formulation is often referred to as a *recipe*. For example, if $x_1, x_2, \ldots, x_p$ denote the proportions of p components that are to be used in a mixture design, the factor space of permissible component values is $0 \leqslant x_i \leqslant 1$, and $x_1 + x_2 + \ldots + x_p = 1$ (that is, 100%). A class of designs that satisfies these constraints is the class of so-called *simplex* designs. Simplex designs and other mixture designs are described in the references to this chapter.

Computer-generated designs that satisfy desired design properties are very popular in industry. The designs are constructed using computer algorithms that produce designs that are deemed to be efficient in terms of a selected design criterion. These programs require input such as the form of the model, the ranges of the variables, the required sample size, and any constraints. Although computer-aided designs can be useful, they have several pitfalls, such as the requirement to specify a model, and can be narrow in application.

Many varieties of classical designs also have become popular. One example is a *small composite* design, which is a special Resolution-III design that has fewer runs than a central composite design. Another is a *hybrid design*, which is a saturated or near-saturated second-order design. A recent innovation is a *noncentral composite* design, which is composed of two designs with different centers, and is of value when the location of interest shifts during the experiment. More information on these and many other alternative designs for studying response surfaces can be found in the references to this chapter.

17.4 FITTING RESPONSE-SURFACE MODELS

A response-surface model represents the functional form of a response surface. Response surface models can be based on either theoretical or empirical considerations. When a theoretical model cannot be specified in an experimental investigation (the usual case), polynomial models often are used to approximate the response surface. A quadratic polynomial [equation (15.25)] can provide a useful approximation for a broad range of applications.

A quadratic response-surface model in two predictor variables consists of the following terms:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + e. \qquad (17.2)$$

As in equation (15.25), the regression coefficients in this model for the second-order terms have double subscripts. The number of times a 1 occurs

in the subscripts indicates the power of x_1, and similarly for x_2. In addition to a sufficient number of observations, at least three distinct values of a predictor variable are needed to fit linear and quadratic powers of the predictor.

Response-surface models, and regression models in general, can be fitted to basically two types of data: observational data and data collected in a prescribed manner according to a designed experiment. Observational data are known by several names, such as "historical data," "old process data," and "happenstance data." Whatever the name, fitting observational data with response-surface models has several potential pitfalls, as shown in Table 17.6.

The list of potential problems with observational data is presented in part to point out the strengths of statistical designs that meet the criteria presented in Table 4.9. Data from factorial and fractional factorial experiments (Chapters 5 and 7) and from central composite and Box–Behnken designs can be efficiently used to fit a response-surface model. Three-level factorial experiments, central composite designs, and Box–Behnken designs will provide data that can be used to fit a full quadratic response-surface model. Two-level factorial experiments and their fractions will yield data to fit a more limited model, linear in all the factors with some product terms.

Table 17.7 displays data from an experiment that used a central composite design to study the affects of annealing time and annealing temperature on the density of a polymer. Run 5 was an additional time and temperature condition included in the experiment by the investigator. Figure 17.13 shows the central–composite design for this experiment.

A quadratic response-surface model will be fit to these data using the standardized form of the predictor variables shown in Equation (15.27). Table 17.8(a) lists the two standardized predictor variables, the products of these two variables, and the two squared variables for the complete quadratic model, Equation (17.2). The correlations in Table 17.8(b) among the linear, product, and squared terms are modest and in many cases near zero. This indicates that each model term is essentially contributing independent information about the

TABLE 17.6 Frequent Defects in Observational Data

- Collinearities among predictor variables are common. This partial confounding among the predictors is especially a problem among polynomial functions of a predictor variable.
- Important predictor effects may go undiscovered because the predictors vary over too narrow a range.
- When significant effects are identified, causation cannot be confirmed. The variable(s) that are significant may be surrogates for other unobserved or uncontrolled variables.
- Excessive effort may be spent dealing with gross data errors, missing values, and inconsistent data-collection periods.

TABLE 17.7 Responses from a Central Composite Design for a Polymer Density Study

Number	Location	Annealing Time (min)	Annealing Temperature (°C)	Polymer Density (g/ml)
1	Star point	60	190	101
2	Corner point	50	155	72
3	Corner point	50	225	101
4	Star point	30	140	70
5	"Extra" point	30	170	91
6	Center point	30	190	98
7	Star point	30	240	Missing
8	Corner point	10	155	70
9	Corner point	10	225	83
10	Star point	2	190	70

Data from Snee, R. D. (1985). "Computer-Aided Design of Experiments — Some Practical Examples." *Journal of Quality Technology*, **17**, 222–236. Copyright American Society for Quality Control, Inc., Milwaukee, WI. Used with permission.

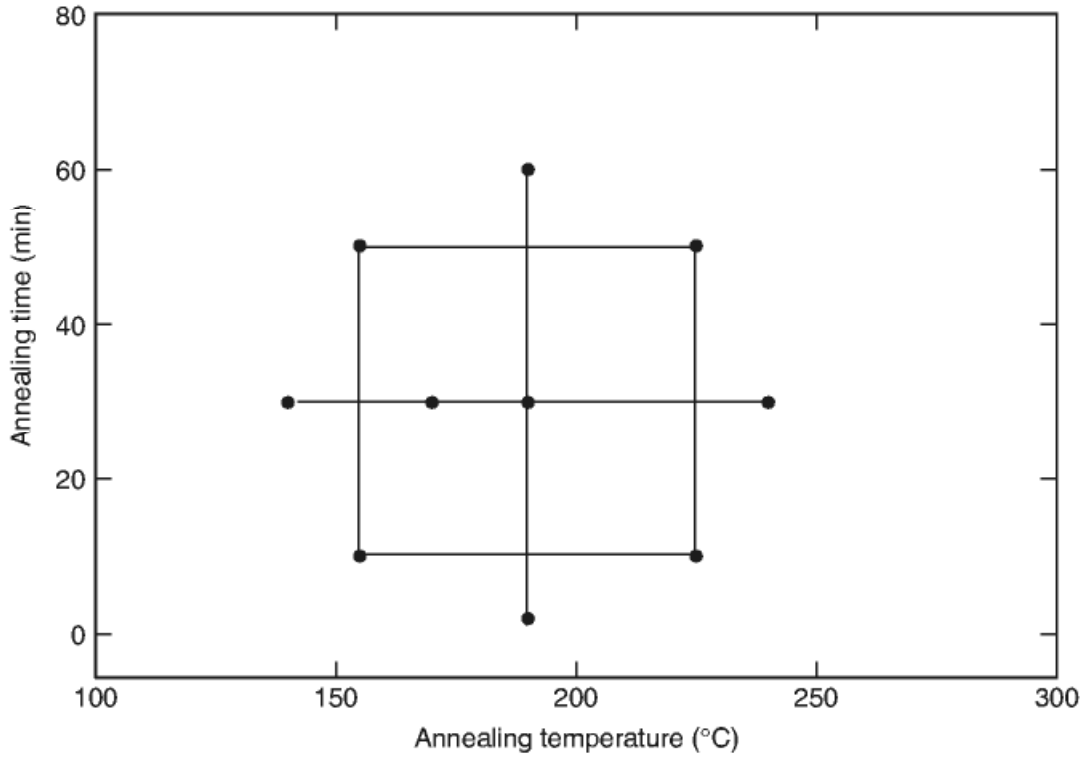


Figure 17.13 Polymer-density study: central composite design (with extra design point added).

TABLE 17.8 Standardized Predictor Values for a Quadratic Fit to the Polymer Density Data

(a) *Standardized Predictors*

No.	Time t	Temp. T	tT	t^2	T^2
1	1.5601	0.0592	0.0924	2.4340	0.0035
2	1.0366	−0.9774	−1.0132	1.0745	0.9553
3	1.0366	1.0958	1.1360	1.0745	1.2009
4	−0.0105	−1.4216	0.0149	0.0001	2.0210
5	−0.0105	−0.5331	0.0056	0.0001	0.2842
6	−0.0105	0.0592	−0.0006	0.0001	0.0035
7	−0.0105	1.5401	−0.0161	0.0001	2.3719
8	−1.0575	−0.9774	1.0336	1.1184	0.9553
9	−1.0575	1.0959	−1.1589	1.1184	1.2009
10	−1.4764	0.0592	−0.0875	2.1797	0.0035

(b) *Correlations*

	t	T	tT	t^2	T^2
T	0.001	1			
tT	0.082	−0.014	1		
t^2	0.054	0.065	0.003	1	
T^2	−0.012	0.184	−0.004	−0.488	1

response. The small correlations among these model terms is a property of most of the statistical experimental designs presented in this book.

A summary of the quadratic fit is given in Table 17.9. Even though one of the experimental conditions resulted in no density measurement (the polymer melted at the highest temperature), the remaining data allowed the fit without any appreciable confounding of the effects of the terms in the model. This is another advantage of well-designed experiments: loss of one or two observations usually does not impair fitting of response surfaces or meaningful interpretation of the fitted models.

The estimated response function for the quadratic fit to the polymer density data is shown graphically in the contour plot displayed in Figure 17.14. The quadratic nature of the response surface is evident from this fit, a fit that required only nine well-chosen observations.

17.4.1 Optimization

When a polynomial representation of a response surface is obtained, the experimenter often is interested in determining the factor levels that provide an

TABLE 17.9 Summary of the Quadratic Fit (Standardized Predictors) to the Polymer Density Data

(a) *Coefficient Estimates*

Model Term	Estimate	t-Value
Intercept	97.59	21.32
Time (z_1)	7.55	3.79
Temperature (z_2)	10.05	4.00
$z_1 z_2$	3.69	1.34
z_1^2	−6.09	−2.33
z_2^2	−7.88	−2.44

(b) ANOVA

Source	df	SS	MS	F	p-Value
Regression	5	1449.78	289.96	8.19	0.057
Error	3	106.22	35.41		
Total	8	1556.00			

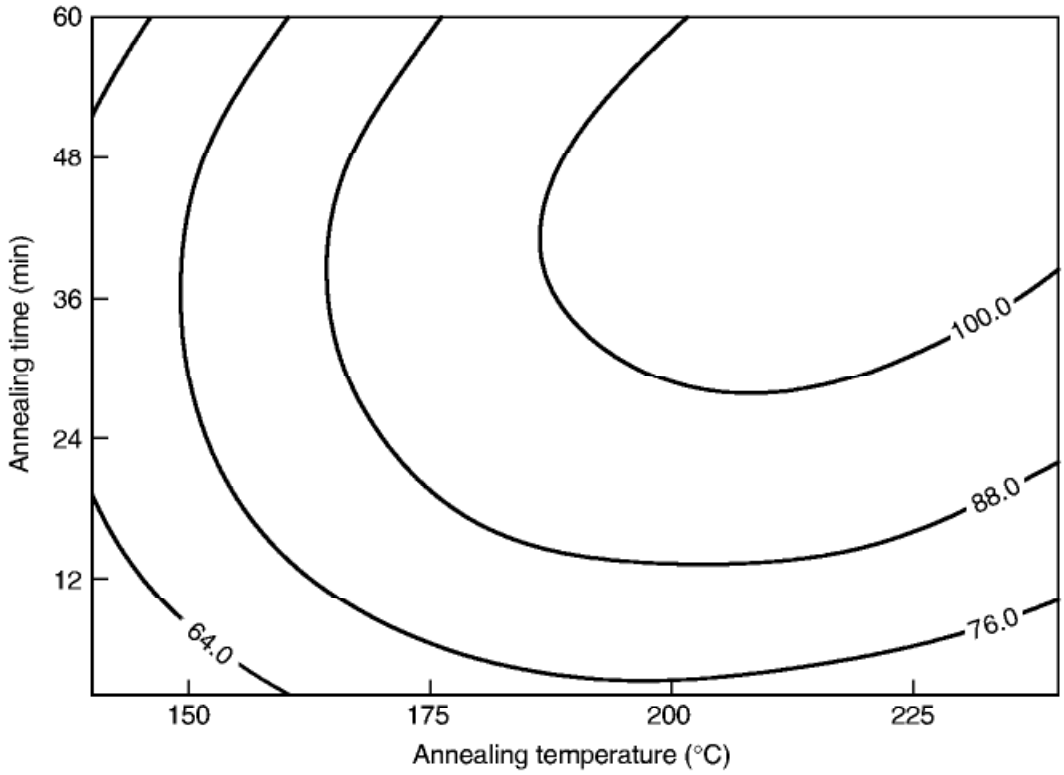


Figure 17.14 Contour plot of fitted polymer-density response surface.

optimum response. The preferred approach for accomplishing this task is the contour-plotting method described in Section 17.1. This use of contour plots is preferred for a variety of reasons. The foremost of these is that in most practical situations there is more than one response of interest in an experimental program. Thus, identification of an optimum in any one response is not as important as is knowing the tradeoffs available among several responses. Whether there is a set of predictor-variable conditions that will result in the optimum of all responses simultaneously becomes apparent when contour plots are overlaid. This approach is discussed in terms of the gentamicin study example in Section 17.1.

A second reason for employing this graphical method is that in addition to the optimum condition being apparent in a contour plot (it is often on the edge of an experimental region), there may also be portions of the experimental region where the response is insensitive to changes in the predictor variables. Where a response is robust to changes in predictor-variable values is important, especially in designing for product quality.

A drawback of the contour-plot approach is that as the number of responses and particularly the number of predictor variables increase, the number of plots to review may appear overwhelming. This can be minimized by eliminating nonsignificant variables from the model and putting predictor variables involved in interactions and curvature on the two axes available while holding other predictor variables at constant values. A contour plot is generated for each response at each combination of the levels of the predictor variables that are being held fixed. A rule of thumb that keeps the number of plots within bounds while making it large enough to allow the salient features of the system under study to be seen is to have each predictor variable being held constant take on three equally spaced values.

For example, if there are four responses of interest and five predictor variables, three of the predictor variables will have to be held constant. If they are each held constant at three different levels, the number of contour plots that result for study is $3 \times 3 \times 3 = 27$ per response, for a grand total of 108. After studying the initial set of plots, the investigator may decide to interchange the fixed predictor variables with one or both of those on the axes; in addition, one or more of the fixed variables may be held at different levels. A new set of contour plots can then be generated.

The question of an optimum response can be mathematically investigated. If an maximum or minimum response value exists, it corresponds to the predictor variable values that are a solution of the set of equations obtained by setting each of the partial derivatives of the estimated response function with respect to the predictor variables equal to zero. The predictor-variable values that satisfy this set of equations are called a *stationary point* (in a p-dimensional space). Canonical analysis is a mathematical approach that can be used to determine the stationary point and whether it represents a maximum, minimum,

or saddle point. Due to the mathematical background required to understand and use canonical analysis (a facility with matrix algebra and calculus), a comprehensive treatment of this topic is beyond the scope of this book. The appendix to this chapter outlines algebraic methods for locating predictor-variable values that optimize a response function.

17.4.2 Optimization for Robust Parameter Product-Array Designs

The concepts associated with Taguchi's *robust parameter design* were introduced and discussed in Chapter 12. Two key design features that emerged from this description include the use of *crossed arrays*, which are often labeled *product arrays*, and the use of *combined arrays*. These arrays are useful when studying the effects of a set of control and noise factors on a response variable. In a crossed-array design (e.g., see Figure 12.4), the inner array consists of a complete or fractional factorial experiment for the control factors, and the outer array consists of an entirely separate complete or fractional factorial experiment for the noise factors.

One method for analyzing the data from product-array designs is the ANOVA procedures introduced in Chapter 8. Even though noise factors are ordinarily random effects in practice, they are analyzed as though they are fixed effects because their levels are specifically chosen in product-array designs. Thus, the natural variation in responses that is due to the noise factors in practice is translated to mean differences in the analysis of product-array data. A critical component of the analysis is the conclusions that are drawn from interaction plots between the control and noise factors. These interactions provide critical information about the robustness (insensitivity) of the response to changes in noise factors as a function of levels of the control factors.

As an example, consider the crossed-array experiment described in Section 12.2.1. An objective of the experiment is to determine the torque associated with an intermediate shaft steering column used to connect the steering wheel of an automobile to the power steering motor. A torque value of 30 is considered ideal for separating the tube from the pockets in the yoke. Large deviations from this torque target value can lead to problems because too little torque causes play in the steering wheel, while too much torque makes it difficult to disassemble the column for repairs.

Table 17.10 (also see Figure 12.4) contains the experimental design layout for the torque data. The first four factors listed in the table are the control factors, while the last two factors are the noise factors. The torque values are contained in the last column. A 2^{4-1} fractional factorial was used with the control factors, and a separate 2^2 complete factorial was used for the noise factors. Because no repeat data were taken in this experiment, unreplicated analysis procedures, such as a normal quantile plot, would be needed to analyze

TABLE 17.10 Experimental Design for Steering Column Experiment

Run	Pocket Depth	Yoke Concentricity	Tube Length	Power Setting	Clearance	Line Voltage	Torque
1	−1	−1	−1	−1	−1	−1	23.4
2	−1	−1	1	1	−1	−1	34.7
3	−1	1	−1	1	−1	−1	33.8
4	−1	1	1	−1	−1	−1	22.6
5	1	−1	−1	1	−1	−1	26.9
6	1	−1	1	−1	−1	−1	16.5
7	1	1	−1	−1	−1	−1	14.9
8	1	1	1	1	−1	−1	28.1
9	−1	−1	−1	−1	−1	1	33.4
10	−1	−1	1	1	−1	1	21.7
11	−1	1	−1	1	−1	1	22.3
12	−1	1	1	−1	−1	1	34.1
13	1	−1	−1	1	−1	1	17.0
14	1	−1	1	−1	−1	1	28.6
15	1	1	−1	−1	−1	1	28.5
16	1	1	1	1	−1	1	14.6
17	−1	−1	−1	−1	1	−1	6.7
18	−1	−1	1	1	1	−1	17.9
19	−1	1	−1	1	1	−1	18.1
20	−1	1	1	−1	1	−1	7.1
21	1	−1	−1	1	1	−1	42.6
22	1	−1	1	−1	1	−1	33.5
23	1	1	−1	−1	1	−1	33.0
24	1	1	1	1	1	−1	43.6
25	−1	−1	−1	−1	1	1	17.8
26	−1	−1	1	1	1	1	6.3
27	−1	1	−1	1	1	1	7.0
28	−1	1	1	−1	1	1	16.9
29	1	−1	−1	1	1	1	30.7
30	1	−1	1	−1	1	1	43.1
31	1	1	−1	−1	1	1	42.6
32	1	1	1	1	1	1	32.4

the effects of the factors. A comprehensive analysis of this data set is left as an exercise at the end of the chapter.

The presence of at least one strong interaction between the noise and control factors in this experiment would lead to the ability to select levels of the control factors that would stabilize, hopefully also minimize, the effects of noise factors. This is Taguchi's concept of design robustness to noise effects.

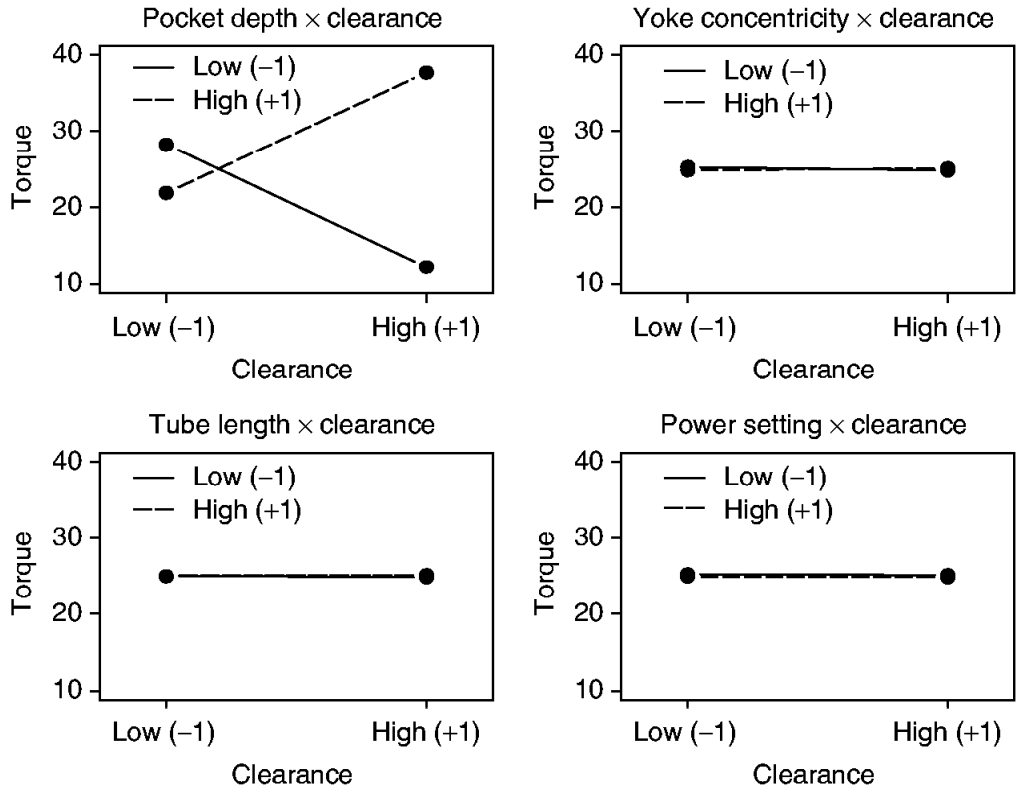


Figure 17.15 Control by noise interaction plot for clearance.

Thus, it becomes of interest to find the levels of the control factors that are least sensitive to the noise factors. This goal is connected to the simultaneous need of keeping the torque near the target value of 30.

Figures 17.15 and 17.16 contain the interaction plots for the eight two-factor interactions between the control and noise factors for the steering column torque data. Scanning the plots, it is clear that strong interactions exist between clearance and pocket depth and between power setting and line voltage. This occurs because the average torque for the different pocket depths and power settings varies greatly for both line voltage and clearance. A deep pocket depth (that is, level = 1) appears to keep the torque near its target value of 30 for both levels of line voltage, and it lessens (relative to the other level of pocket depth) the torque spread from the target value across the levels of clearance. Although power setting does not interact with clearance, it does with line voltage. Neither level of power setting is unaffected by line voltage; either power setting appears to be equally affected by changes in line voltage. Finally, neither of the levels of yoke concentricity or of tube length substantively affect the average torque and both are insensitive to the noise factors.

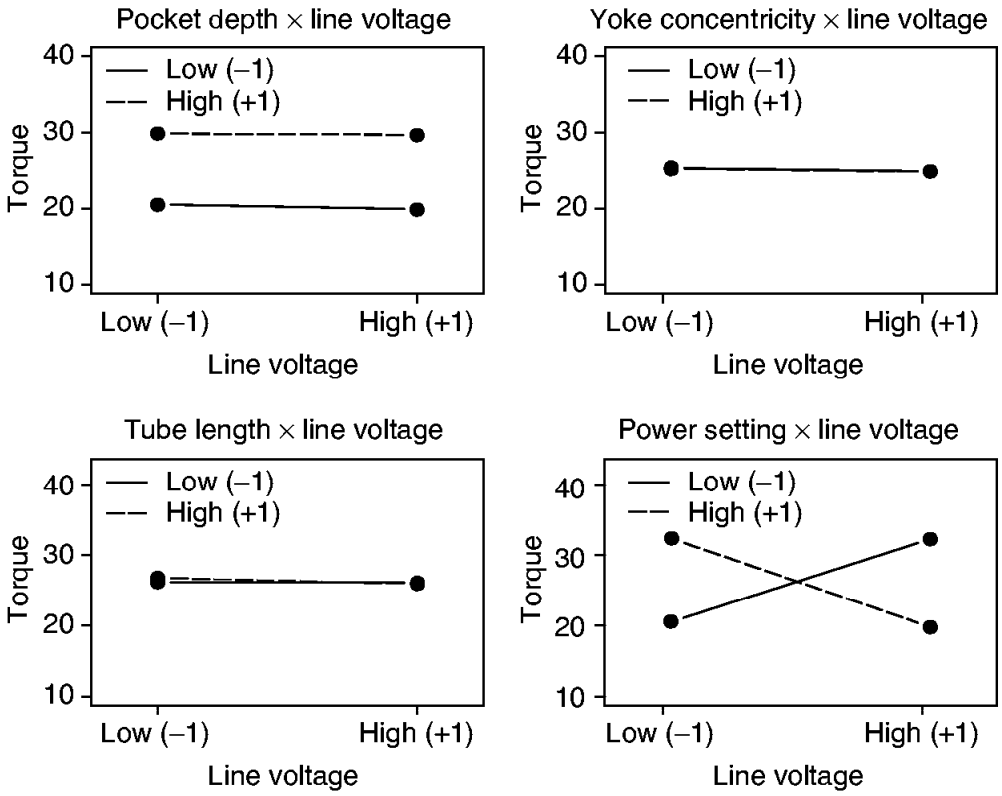


Figure 17.16 Control by noise interaction plot for line voltage.

A more detailed statistical analysis of the data can be found in Lorenzen and Andersen (1993).

17.4.3 Dual Response Analysis for Quality Improvement Designs

In the integrated design approach introduced in Chapter 12, a single complete or fractional factorial experiment is used for the combined set of control and noise factors. This type of combined-array design often offers more flexibility than a crossed-array design because it allows for the experimenter to exploit the control-noise interactions when forming the design. Crossed-array designs do not permit the efficiency that can be obtained from fractionating the control-noise interactions.

While the analysis outlined in the previous section can be applied to combined-array designs, an extremely important alternative is also available. The *dual response surface* approach enables investigators to fit two response surfaces: one for the process mean and one for the process variability. Typically this procedure is used with a combined-array design, so that only a single

experimental design is necessary, but it can also be used with a crossed-array design as in the example below.

A basic dual response model with two control factors and two noise factors has the following form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_{12} z_1 z_2 + \delta_{11} x_1 z_1 + \delta_{12} x_1 z_2 + \delta_{21} x_2 z_1 + \delta_{22} x_2 z_2 + e \tag{17.3}$$

where

y = response variable,
x_i = fixed control factor with unknown coefficient β_i,
$x_i x_j$ = control-by-control interaction with unknown coefficient β_{ij},
z_i = random noise factor with unknown coefficient γ_i,
$z_i z_j$ = noise-by-noise interaction with unknown coefficient γ_{ij},
$x_i z_j$ = control-by-noise interaction with unknown coefficient δ_{ij},
e = random error term.

Note that the control factors in model (17.3) can interact with each other, as can the noise factors. Also, there are control-by-noise interactions for each control factor. As was discussed in the last section, if at least one of the coefficients of the control-noise interactions is nonzero, some form of robustness to noise variation may be achievable. The model in Equation (17.3) can be generalized by adding higher-order terms for the control and noise factors. More details on these types of models can be found in the reference list at the end of the chapter.

In the analysis of this model, it is assumed that the errors are independently and normally distributed with zero means and a constant variance of σ^2. It also is assumed that the noise factors are random variables (even though they are controllable in the experiment) and have zero means, a constant variance of σ_z^2, zero covariances with one another, and are independent of the error terms. Other assumptions are that the control and noise factors are expressed as coded variables as in Exhibit 17.4. By coding the noise variables, $\sigma_z = 1$; this result becomes useful when attempting to evaluate the process variance model.

With the above assumptions, the response model for the process mean is given by

$$E_z(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2, \tag{17.4}$$

and the response model for the process variance is given by

$$V_z(y) = (\gamma_1^2 + \gamma_2^2 + \delta_{11}^2 x_1^2 + +\delta_{12}^2 x_1^2 + \delta_{21}^2 x_2^2 + \delta_{22}^2 x_2^2)\sigma_z^2 + \sigma^2. \tag{17.5}$$

The application of regression modeling to the fitting of dual response models is very straightforward. Typical steps in such an analysis are listed in Exhibit 17.8.

EXHIBIT 17.8 STEPS FOR FITTING A DUAL RESPONSE MODEL

1. Fit a response surface model [such as the one given in (17.3)] to the data, including control-by-noise interactions.
2. Delete any nonsignificant terms from the fitted model and refit the data using the reduced set of terms.
3. From the fitted model, select the appropriate terms for the process mean and process variance models in equations (17.4) and (17.5).
4. Substitute the estimated coefficients obtained in Step 2 into the models specified in Step 3.
5. Use the error mean square from the fit in Step 2 to estimate σ^2, and set σ_z equal to an appropriate value (such as 1) based on the coding scheme used for the noise factors.
6. Optimize the resulting estimated mean and variance models by treating them as a multiple-response problem (i.e., see Sections 17.1 and subsection 17.4.1).

To illustrate the above method, the torque data given in Table 17.10 will be reexamined. This experiment used a crossed-array design in which the control factors form a half-fraction of a 2^4 complete factorial (R_{IV}, so two-factor interactions are aliased with one another) and the noise factors form a complete 2^2 factorial for each combination of levels of the control factors. The four control factors are pocket depth (x_1), yoke concentricity (x_2), tube length (x_3), and power setting (x_4), and the two noise factors, z_1, and z_2, are clearance and line voltage. An initial model fit included all 6 linear (main) effects, 3 (aliased) two-factor interactions between the control factors, the two-factor interaction between the noise factors, the 8 two-factor interactions between the control and noise factors, and 13 additional higher-order interactions between the control and noise factors. Since this is a saturated model (no degrees of freedom remain to estimate the error variance because there are no repeat tests; see Exhibit 7.7), a normal quantile plot of the estimated effects was made to determine which of the high-order interactions could be assumed to be negligible. From this analysis, all 13 of the interaction effects higher than second-order were assumed to be negligible and were pooled together to form the error variance estimate.

A reduced response model was then fit to the data. It included linear terms for the 6 factors as well as 12 two-factor interactions between these factors. Finally, all nonsignificant terms were deleted and the model was refit. The resulting fitted response surface model is:

$$\hat{y} = 25.013 + 4.775x_1 + 7.956x_1z_1 - 5.656x_4z_2,$$

and the mean square error from the fit is 0.651. Based on these results, the estimated mean model is given by

$$\hat{E}_z(y) = 25.013 + 4.775x_1.$$

Assuming that $\sigma_z^2 = 1$, and using the above mean square error to estimate σ^2, the estimated variance model is given by

$$\hat{V}_z(y) = 0.651 + 63.301x_1^2 + 31.993x_4^2.$$

From the estimated mean model, the torque is nearest its target value of 30 when $x_1 = 1$; (that is, when pocket depth is at its deep setting). Differentiating the estimated variance model, the estimated process variance is minimized for values of x_1 and x_4 satisfying $126.602x_1 + 63.986x_4 = 0$, or $x_4 = -1.979$. If the selection of control levels is to be restricted to the (coded) experimental region, the optimum value would appear to be $x_4 = -1$, which occurs when the power setting is 80. While this power setting minimizes the estimated process variance for the deep pocket depth setting, note that the predicted torque $\hat{y}$ still depends on line voltage. Thus, the optimal power setting does not keep the predicted torque near 30 for both line voltages. This is consistent with what was noted in the last section from the interaction plots. Also, because no terms involving x_2 and x_3 are significant, neither of these control factors affects the above conclusions. This, too, is consistent with the analysis in the last section.

APPENDIX: BOX–BEHNKEN DESIGN PLANS; LOCATING OPTIMUM RESPONSES

1. Box–Behnken Design Plans

TABLE 17.A.1 Box–Behnken Designs

Number of Factors	Coded Factor Levels									Number of Points
	1	2	3	4	5	6	7	8	9	
3	±1	±1	0							4
	±1	0	±1							4
	0	±1	±1							4
	0	0	0							3
										15

TABLE 17.A.1 (***continued***)

Number of Factors	Coded Factor Levels 1	2	3	4	5	6	7	8	9	Number of Points
4	±1	±1	0	0						4
	±1	0	±1	0						4
	±1	0	0	±1						4
	0	±1	±1	0						4
	0	±1	0	±1						4
	0	0	±1	±1						4
	0	0	0	0						3
										27
5	±1	±1	0	0	0					4
	±1	0	±1	0	0					4
	±1	0	0	±1	0					4
	±1	0	0	0	±1					4
	0	±1	±1	0	0					4
	0	±1	0	±1	0					4
	0	±1	0	0	±1					4
	0	0	±1	±1	0					4
	0	0	±1	0	±1					4
	0	0	0	±1	±1					4
	0	0	0	0	0					6
										46
6	±1	±1	0	±1	0	0				8
	0	±1	±1	0	±1	0				8
	0	0	±1	±1	0	±1				8
	±1	0	0	±1	±1	0				8
	0	±1	0	0	±1	±1				8
	±1	0	±1	0	0	±1				8
	0	0	0	0	0	0				6
										54
7	0	0	0	±1	±1	±1	0			8
	±1	0	0	0	0	±1	±1			8
	0	±1	0	0	±1	0	±1			8
	±1	±1	0	±1	0	0	0			8
	0	0	±1	±1	0	0	±1			8
	±1	0	±1	0	±1	0	0			8
	0	±1	±1	0	0	±1	0			8
	0	0	0	0	0	0	0			6
										62

(*continued overleaf*)

TABLE 17.A.1 (*continued*)

Number of Factors	Coded Factor Levels									Number of Points
	1	2	3	4	5	6	7	8	9	
9	±1	0	0	±1	0	0	±1	0	0	8
	0	±1	0	0	±1	0	0	±1	0	8
	0	0	±1	0	0	±1	0	0	±1	8
	±1	±1	±1	0	0	0	0	0	0	8
	0	0	0	±1	±1	±1	0	0	0	8
	0	0	0	0	0	0	±1	±1	±1	8
	±1	0	0	0	±1	0	0	0	±1	8
	0	0	±1	±1	0	0	0	±1	0	8
	0	±1	0	0	0	±1	±1	0	0	8
	±1	0	0	0	0	±1	0	±1	0	8
	0	±1	0	±1	0	0	0	0	±1	8
	0	0	±1	0	±1	0	±1	0	0	8
	±1	0	0	±1	0	0	±1	0	0	8
	0	±1	0	0	±1	0	0	±1	0	8
	0	0	±1	0	0	±1	0	0	±1	8
	0	0	0	0	0	0	0	0	0	10
										130

2. Locating Optimum Responses

2.1 Path of Steepest Ascent

Graphical or algebraic techniques for identifying predictor-variable values that correspond to a maximum or a minimum of a response surface are most effective when the predictor variables are equivalently centered or scaled. This can be accomplished by using coded predictor or factor levels as in Exhibit 17.4 or by standardization as in equations (15.26) or (15.27).

The path of steepest ascent through the experimental region follows the direction of maximal increase or decrease in the predicted response variable. It is most frequently used with prediction equations that are linear in the predictors. Canonical analysis is usually used with second-order prediction equations.

The path of steepest ascent is most satisfactorily implemented when the predictor variables or factor levels are coded so that the center of the experimental region corresponds to the value 0 for each predictor and the limits on each factor are ±1 (see Exhibit 17.4). Denoting the scaled predictors by x_j, $j = 1, 2, \ldots, k$, the fitted prediction equation is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k. \qquad (17.A.1)$$

The path of steepest ascent is determined by finding the values of x_j that either maximize or minimize Equation (17.A.1) subject to the constraint

$$\sum x_j^2 = c, \tag{17.A.2}$$

where c is any constant. The constraint (17.A.2) simply requires that the solution to the maximization or minimization of (17.A.1) lie on a sphere of radius $c^{1/2}$. The path of steepest ascent is then a line from the origin through the point on the sphere (17.A.2) for which the response (17.A.1) is the largest or smallest, as appropriate.

One can show that the solution to this problem is

$$x_j = qb_j, \tag{17.A.3}$$

where q is a constant that depends on the radius of the sphere in (17.A.2), and b_j is the coefficient of x_j in (17.A.1). Since choices of the radius of the sphere are arbitrary, so are values of q. Because of this, the procedure in Exhibit 17.A.1 can be used to locate an optimum response using the path of steepest ascent.

EXHIBIT 17.A.1 PATH OF STEEPEST ASCENT

1. Code the factor or response variable so that the center of the experimental region is at the origin $x_1 = \cdots = x_k = 0$ and the extremes in each variable are -1 and $+1$.
2. Fit a linear response surface of the form (17.A.1) to the response.
3. Choose a value for one of the predictors, say x_1, at which a new test run or observation is to be made. This new value should be some fractional change in the largest or smallest coded value, depending on the sign of b_1 and on whether the optimum response is a maximum or a minimum of the response surface. For example, if b_1 is positive and a maximum response is sought, choose x_1 to be 1.1, 1.2, or some other convenient value.
4. With a selected value of x_1, solve (17.A.3) for $q : q = b_1/x_1$. Insert this value of q into (17.A.3), and solve for the other x_j. The values of $x_1, x_2, \ldots, x_k$ determine a set of predictor-variable values or factor–level combinations at which the next observation is to be taken.
5. Determine the value of the response for the x_j-values from step 4. If a maximum (minimum) is sought and the new response is an increase (a decrease) over the previous one, return to step 3. Otherwise, terminate experimentation along the path of the steepest ascent.

Once the path of steepest ascent no longer leads to increases or decreases, as appropriate, or the changes become small, this procedure should be discontinued in favor of a more elaborate experiment. The response-surface designs

discussed in this chapter should then be used to collect data from which a satisfactory response-surface model can be fitted.

2.2 Canonical Analysis

When second-order regression models such as Equation (15.26) are used to characterize a response surface, the path of steepest ascent cannot be used to identify the location of the optimum response. A *canonical* analysis of the fitted prediction equation is used.

For a canonical analysis, it is convenient to use standardized predictor variables, because polynomial models are being fitted. A complete second-order model in matrix notation (see the appendix to Chapter 15) can be written as

$$\hat{y} = b_0 + \mathbf{x}'\mathbf{b} + \mathbf{x}'B\mathbf{x}. \tag{17.A.4}$$

Here $\mathbf{b}' = (b_1, b_2, \ldots, b_k)$ is a vector that contains the least-squares estimates of the linear terms of the model (15.25), $\mathbf{x}$ is a vector of the values of the k predictor variables, and B is a symmetric matrix containing the estimates of the second-order terms of the model (15.25) in the following form:

$$B = \begin{bmatrix} b_{11} & b_{12}/2 & \cdots & b_{1k}/2 \\ b_{12}/2 & b_{22} & \ldots & b_{2k}/2 \\ \vdots & \vdots & & \vdots \\ b_{1k}/2 & b_{2k}/2 & \cdots & b_{kk} \end{bmatrix}.$$

A stationary point of the fixed response surface (17.A.4) is obtained by setting the partial derivatives of the fitted surface with respect to each of the predictors equal to zero. The result is

$$\mathbf{x}_s = -B^{-1}\mathbf{b}/2. \tag{17.A.5}$$

The nature of the stationary point — whether it locates a maximum, a minimum, or a saddle point — is determined by expressing the prediction equation (17.A.4), evaluated at the stationary point (17.A.5), in canonical form.

Denote the eigenvalues of B by $l_1, l_2, \ldots, l_k$ and the corresponding eigenvectors by $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$. The eigenvalues and eigenvectors satisfy the matrix equation $B\mathbf{v}_j = l_j\mathbf{v}_j$, $j = 1, 2, \ldots, k$. The canonical form of the prediction equation can then be expressed as follows:

$$\hat{y} = \hat{y}_s + l_1u_1^2 + l_2u_2^2 + \cdots + l_ku_k^2, \tag{17.A.6}$$

where $\hat{y}_s$ is the value of the prediction equation (17.A.4) at the stationary point $\mathbf{x}_s$, and $u_j = \mathbf{v}_j'(\mathbf{x} - \mathbf{x}_s)$ is a transformation of the original k predictor variables to k new canonical variables u_j.

If all the l_j are positive, (17.A.6) shows that any departure from the stationary point $\mathbf{x}_s$ results in an increase in the predicted response. Thus, the stationary point locates a minimum of the response surface. If all the l_j are

negative, the stationary point locates a maximum of the fitted response surface, because any departure from the stationary value decreases the predicted response. If some of the l_j are positive and some are negative, the stationary point is a saddle point: the fitted response increases in directions corresponding to positive l_j, and it decreases in directions corresponding to negative l_j. These directions emanate from the stationary point and are defined by the transformations $\mathbf{v}_j'\mathbf{x}$ of the original coordinate axes.

The magnitudes of the l_j indicate the relative sensitivity of the predicted response variable to changes in the direction defined by u_j. Large values of l_j indicate [see Equation (17.A.6)] that unit changes in the corresponding u_j-direction result in large increases or decreases in the predicted response relative to those directions corresponding to small l_j.

REFERENCES

Text References

Response surface designs and analyses, including computer-generated designs, small composite designs, and hybrid designs, are discussed in the following texts. Also included are discussions of the role of empirical models, including polynomial models, in describing response surfaces, and illustrations and discussions of the method of canonical analysis for describing the fitted surface.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. New York: John Wiley & Sons, Inc.

Khuri, A. and Cornell, J. A. (1996). *Response Surface Designs and Analyses*. New York: John Wiley & Sons, Inc.

Myers, R. H. and Montgomery, R. H. (1995). *Response Surface Methodology*. New York: John Wiley & Sons, Inc.

The following journal articles contain additional information on a variety of designs and models that are useful for fitting response surfaces. Some of these articles are highly technical.

Box, G. E. P. (1954). "The Exploration and Exploitation of Response Surfaces: Some General Considerations and Examples," *Biometrics*, **10**, 16–60.

Box, G. E. P. (1959). "A Basis for the Selection of a Response Surface Design," *Journal of the American Statistical Association*, **54**, 622–654.

Box, G. E. P. (1975). "Robust Designs," *Biometrika*, **67**, 347–352.

Box, G. E. P. and Draper, N. R. (1963). "The Choice of a Second Order Rotatable Design," *Biometrika*, **50**, 335–352.

Box, G. E. P. and Youle, P. V. (1955). "The Exploration and Exploitation of Response Surfaces: An Example of the Link between the Fitted Surface and the Basic Mechanism of the System," *Biometrics*, **11**, 287–323.

Hunter, J. S. (1958–1959). "Determination of Optimum Operating Conditions by Experimental Methods," *Industrial Quality Control*, **15**, Part II-1 (Dec. 1958), Part II-2 (Jan. 1959), **16**, Part II-3 (Feb. 1959).

Myers, R. H. (1999). "Response Surface Methodology—Current Status and Future Directions," with discussion, *Journal of Quality Technology*, **31**, 30–74.

Myers, R. H., Khuri, A. I., & Vining, G. (1992). "Response Surface Alternatives to the Taguchi Robust Parameter Design Approach," *The American Statistician*, **46**, 131–139.

An extensive list of Box–Behnken designs for 3–12 factors is contained in:

Box, G. E. P. and Behnken, D. W. (1960). "Some New Three Level Designs for the Study of Quantitative Variables," *Technometrics*, **2**, 455–476.

Two important references on mixture designs and analyses are:

Cornell, J. A. (1981). *Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data*, New York: John Wiley & Sons, Inc.

Snee, R. D. and Marquardt, D. W. (1976). "Screening Concepts and Designs for Experiments with Mixtures," *Technometrics*, **18**, 19–30.

A useful reference for noncentral composite designs is:

Mee, R. (2001). "Noncentral Composite Designs," *Technometrics*, **43**, 34–43.

Data References

Designs for the gentamicin method study and the thyroxine experiment were taken from the following article:

Myers, G. C., Jr. (1985). "Use of Response Surface Methodology in Clinical Chemistry," in Snee, R. D., Hare, L. B., and Trout, J. R. (Eds.), *Experiments in Industry: Design, Analysis and Interpretation of Results*, Milwaukee, WI: American Society for Quality Control. Copyright American Society for Quality Control, Inc., Milwaukee, WI. Reprinted by permission.

The design for the Sulphlex experiment was taken from:

Dale, J. M. (1984). "Design for Sulphlex® Binders," Federal Highway Administration, Contract No. DTFH-61-82-C-00049. San Antonio, TX: Southwest Research Institute.

The design and analyses for the steering column torque experiment was taken from:

Lorenzen, T. J. and Villalobos, M. A. (1990). "Understanding Robust Design, Loss Functions, and Signal To Noise Ratios," Research Publication GMR-7118, General Motors Research Laboratories, Warren, MI.

Lorenzen, T. J. and Anderson, V. L. (1993). *Design of Experiments: A No-Name Approach*. New York: Marcel Dekker, Inc.

EXERCISES

1 Sketch three-dimensional surfaces for the contours graphed in Figure 17.3.

2 A fitted second-order model for a response surface has the following equation:

$$y = 300.0 + 70.0x_1 + 70.5x_2 + 3.0x_1x_2 - 10.5x_1^2 - 10.0x_2^2.$$

Sketch a contour plot of this three-dimensional surface by calculating the responses for a grid of values of the two factors and interpolating between adjacent responses. The ranges of interest on each of the two factors are 0 to 10. Plot contours for response values of 100 to 500 in increments of 100.

3 Use the Sulphlex-study contour plot in Figure 17.4 to plan a sequence of test runs along the path of steepest ascent to find the maximum viscosity (see also Figure 17.8). Specify ten factor–level combinations.

4 A ceramic diesel engine is being tested in buses to measure the amount of combustion soot being deposited in the engines. It is known that engine temperature plays an important role in the depositing of soot in engines. A study is being conducted to determine the maximum temperature (°F) in the engine during a combustion test. Two factors are of interest in one phase of the study: the time (sec) during which combustion takes place, and the engine operating speed (rpm). Initially, a 2^2 factorial experiment was conducted with the results shown below. Design a sequence of tests using the path of steepest ascent. List at least five test runs.

Time (sec)	**Speed (rpm)**	**Temperature (°F)**
10	1000	350
100	1000	625
10	1600	200
100	1600	400

5 A disk-type test rig was designed and fabricated to measure the wear of graphite material under specified test conditions. Two central composite designs were constructed using the following ranges on the two factors

temperature: 500–1000°F,
contact pressure: 15–21 psi.

The designs below are in coded factor levels. Replace the coded levels with the actual factor levels. Is either design rotatable? Why (not)? If not, construct a rotatable central composite design for this project.

	Design 1		**Design 2**	
Run	**Temperature**	**Pressure**	**Temperature**	**Pressure**
1	−1	−1	−1	−1
2	−1	1	−1	1
3	1	−1	1	−1

	Design 1		Design 2	
Run	**Temperature**	**Pressure**	**Temperature**	**Pressure**
4	1	1	1	1
5	0	−1	0	−1.5
6	0	1	0	1.5
7	−1	0	−1.5	0
8	1	0	1.5	0
9	0	0	0	0
10	0	0	0	0
11	0	0	0	0

6 Construct a Box–Behnken design for the previous exercise. Under which of the following conditions would the various central composite designs or the Box–Behnken design be preferable? Why?

(a) The limits in Exercise 5 are not constraints, only suggested starting ranges.

(b) No design points can exceed the ranges specified in Exercise 5, but test runs are desired as close to the limits as possible.

(c) No design points can exceed the ranges specified in Exercise 5, and test runs are not desired in the extreme corners of the region.

7 Suppose that in addition to temperature and contact pressure, sliding speed and disk hardness are important factors in the experiment described in Exercise 5. These latter two factors have the following ranges:

Sliding speed	54–60 ft/sec
Disk hardness	58–60 R_c

Construct a rotatable central composite design for this experiment. Include six repeat tests. List the factor–level combinations in both coded and noncoded form.

8 Construct a Box–Behnken design for the experiment in Exercise 7.

9 Refer to Exercise 4. Using the methodology described in the appendix to this chapter, design a sequence of test runs along the path of steepest ascent to locate the maximum temperature. List at least five test runs.

10 The data below are coded wear measurements for the second design in Exercise 5. Fit a second-order model to these data. Do the t-statistics from the model indicate that a second-order model is necessary for an adequate fit to the wear measurements?

	Coded Factor Levels	
	−1	+1
Temperature (°F)	500	1000
Contact Pressure (psi)	15	21

Run	Temperature	Pressure	Wear
1	−1	−1	0.1014
2	−1	1	0.5009
3	1	−1	0.8152
4	1	1	0.4026
5	0	−1.5	0.7001
6	0	1.5	0.1995
7	−1.5	0	0.5753
8	1.5	0	0.8747
9	0	0	0.4893
10	0	0	0.5031
11	0	0	0.5118

11 Use the coefficient estimates in Table 17.9 to locate the stationary point for the polymer-density fitted response surface. Transform the coded annealing-time and annealing-temperature values at the stationary point to values in the original scales of the two predictors. Using a canonical analysis, determine whether the stationary point locates a maximum, a minimum, or a saddle point for the response.

12 Use a canonical analysis to investigate the nature of a second-order response surface fitted to the data in Exercise 10. Is there a maximum or a minimum within the experimental region? If so, identify the factor levels that correspond to the optimum response.

13 Re-analyze the torque data given in Table 17.10. Ignore the two noise factors and separately perform an ANOVA on the four control factors using as responses the mean torque and the natural logarithm of the torque. Also, plot the main effects and interactions plots for each factor. Based on these results, determine an optimal combination for the four control factors if the goal is to minimize process variance while attaining a target torque value of 30.

14 The coefficient estimates for a fitted response surface model are given below. Use these coefficient estimates to locate the stationary point for the fitted surface. Using canonical analysis, determine whether the stationary point locates a maximum, a minimum, or a saddle point for the response.

Assume that the data are in the original scales for the four predictor variables, X_1, X_2, X_3, and X_4.

$$\hat{y} = 4.9253 - 0.1560^*X_1 - 10.4734^*X_2 - 0.006887^*X_3 - 0.05506^*X_4 - 0.2902^*X_1{}^*X_2 - 2.2992E - 005^*X_1{}^*X_3 - 0.0005808^*X_1{}^*X_4 - 0.00060591^*X_2{}^*X_3 + 0.27750^*X_2{}^*X_4 + 0.0001687^*X_3{}^*X_4 + 0.01056^*X_1{}^*X_1 + 24.9211^*X_2{}^*X_2 + 7.1517E - 006^*X_3{}^*X_3 + 0.005955^*X_4{}^*X_4$$

15 The performance evaluation of a fuel cell (FC) assembly for an engine involves operating the fuel cell under certain conditions of applied voltage, gas pressure, gas flow, cell temperature, and humidification temperature of the anode and cathode. It is desired to run an experiment to measure the current density (mA/cm^2) supplied by the FC assembly when a fixed amount of voltage is applied at atmospheric pressure. In addition, gas flow and gas pressure are to be held constant. The factors of interest are as follows:

cell temperature	65–75°C
anode humidification temperature	75–90°C
cathode humidification temperature	75–90°C

Construct a face-centered central composite design for this experiment, and list the test runs. Assume three repeats will be run at the cube center.

16 To determine the performance of an engine, its throttle was measured at various combinations of speed and torque. The resultant data are given below. Fit a second-order model to these 134 observations, and present your results.

Speed (rpm)	Torque (ft-lbs)	Coded Throttle (volts)	Speed (rpm)	Torque (ft-lbs)	Coded Throttle (volts)
900	0	0.360	1500	0	1.396
900	100	0.698	1500	100	1.426
900	200	0.725	1500	200	1.457
900	300	0.748	1500	300	1.466
900	400	0.768	1500	400	1.514
900	500	0.803	1500	500	1.544
900	600	0.837	1500	600	1.557
900	700	0.996	1500	700	1.604
900	715	1.008	1500	800	1.617
1000	0	0.752	1500	900	1.659
1000	100	0.781	1500	1000	1.665
1000	200	0.827	1500	1030	1.744

Speed (rpm)	Torque (ft-lbs)	Coded Throttle (volts)	Speed (rpm)	Torque (ft-lbs)	Coded Throttle (volts)
1000	300	0.861	1600	0	1.546
1000	400	0.884	1600	100	1.566
1000	500	0.899	1600	200	1.600
1000	600	0.923	1600	300	1.618
1000	700	0.973	1600	400	1.642
1000	800	0.996	1600	500	1.672
1000	900	1.018	1600	600	1.700
1000	998	1.197	1600	700	1.727
1100	0	0.902	1600	800	1.762
1100	100	0.922	1600	900	1.780
1100	200	0.955	1600	1000	1.817
1100	300	0.975	1600	1010	1.875
1100	400	1.008	1700	0	1.661
1100	500	1.028	1700	100	1.672
1100	600	1.061	1700	200	1.718
1100	700	1.095	1700	300	1.739
1100	800	1.123	1700	400	1.754
1100	900	1.167	1700	500	1.788
1100	1000	1.284	1700	600	1.813
1100	1022	1.293	1700	700	1.848
1200	0	1.023	1700	800	1.879
1200	100	1.044	1700	900	1.913
1200	200	1.058	1700	950	1.970
1200	300	1.090	1800	0	1.785
1200	400	1.120	1800	100	1.823
1200	500	1.152	1800	200	1.847
1200	600	1.192	1800	300	1.892
1200	700	1.209	1800	400	1.921
1200	800	1.245	1800	500	1.940
1200	900	1.259	1800	600	1.981
1200	1000	1.300	1800	700	2.010
1200	1039	1.381	1800	800	2.044
1300	0	1.139	1800	900	2.100
1300	100	1.163	1800	900	2.110
1300	200	1.199	1900	0	1.925
1300	300	1.234	1900	100	1.964
1300	400	1.268	1900	200	1.984
1300	500	1.292	1900	300	2.030
1300	600	1.316	1900	400	2.045

Speed (rpm)	Torque (ft-lbs)	Coded Throttle (volts)	Speed (rpm)	Torque (ft-lbs)	Coded Throttle (volts)
1300	700	1.329	1900	500	2.070
1300	800	1.354	1900	600	2.120
1300	900	1.372	1900	700	2.162
1300	1000	1.414	1900	800	2.180
1300	1064	1.510	1900	810	2.254
1400	0	1.249	2000	0	2.090
1400	100	1.280	2000	100	2.107
1400	200	1.324	2000	200	2.141
1400	300	1.350	2000	300	2.167
1400	400	1.365	2000	400	2.190
1400	500	1.382	2000	500	2.240
1400	600	1.439	2000	600	2.266
1400	700	1.460	2000	700	2.313
1400	800	1.494	2000	750	2.790
1400	900	1.514	2050	0	3.650
1400	1000	1.547			
1400	1060	1.585			

17 Using the data of Exercise 16, determine the combination of engine speed and torque that produces the lowest throttle value.

18 Using the data given in Exercise 16 of Chapter 15, fit a second-order model to the given two predictor variables and present your results.

19 Sketch a contour plot of the three-dimensional surface obtained in Exercise 18, and determine an approximate location of the optimum value of the response.

20 An experiment was run relating a response variable, y, to five predictor variable, X_1, X_2, X_3, X_4, and X_5. The predictor X_3 is categorical and indicates a location; the other predictor variables are continuous. The resulting data is given below. Fit a second-order model to the fifty observations, and determine the optimum combination of the predictor variables if the desire is to maximize the y value.

X_1	X_2	X_3	X_4	X_5	Y
1	140	1	50	352	213
0.6	140	0	50	1379	800
0.6	60	1	50	352	225
0.6	140	1	50	965	866
0.6	60	0	50	352	345
1	60	1	50	965	244
0.8	140	1	50	1379	346

X_1	X_2	X_3	X_4	X_5	Y
0.8	60	1	50	965	303
0.6	100	1	50	965	938
0.6	60	0	50	1379	980
1	140	1	50	1379	255
1	140	0	50	352	175
0.6	60	1	50	1379	1044
1	60	0	50	1379	175
1	60	0	50	352	147
0.6	140	1	50	352	540
0.8	60	0	50	965	185
1	60	1	100	1379	250
0.6	60	1	100	965	748
0.6	60	1	100	352	480
1	60	0	100	1379	164
1	60	0	100	1379	165
1	140	0	100	1379	170
1	140	0	100	1379	172
0.8	60	0	100	352	157
0.8	60	0	100	352	155
1	140	1	100	1379	232
0.8	60	1	100	1379	319
0.6	140	0	100	352	185
0.6	140	0	100	352	202
0.6	140	1	100	352	460
0.6	60	0	100	1379	660
0.6	60	0	100	1379	645
0.6	100	1	100	1379	912
0.6	60	0	100	352	202
0.6	60	0	100	352	182
1	140	1	100	965	236
0.8	140	1	100	352	244
0.8	60	1	100	1379	309
1	140	0	100	965	156
1	140	0	100	965	156
0.8	100	0	100	1379	183
0.8	100	0	100	1379	182
1	60	1	100	352	211
1	140	0	100	352	133
1	140	0	100	352	132
0.6	140	0	100	1379	850
0.6	140	1	100	1379	912
0.6	100	0	100	352	216
0.6	100	0	100	352	218

CHAPTER 18

Model Assessment

Assessing the validity of model assumptions is often critical to the successful application of statistical inference procedures. In this chapter techniques for investigating the soundness of model assumptions are presented for both single predictor variables and multiple regression models. Also introduced are some procedures that are useful when the model is found to be inadequate. The topics discussed in this chapter include:

- *outlier detection techniques,*
- *procedures for evaluating the adequacy of model assumptions, and*
- *reexpression of response and predictor variables to better satisfy model assumptions.*

The methods presented in this chapter are techniques for assessing common statistical model assumptions, model assumptions that are used in a variety of statistical inference procedures. It is assumed that initial model specifications have been made, and, if appropriate, a cursory inspection of the reasonableness of those model assumptions has been made through an examination of simple plots such as point plots, histograms, boxplots, and scatterplots of the variables. We consider in this chapter the evaluation of a specific statistical model for which assumptions of normality, independence of errors, or correct model specification are of concern. Further, we describe several model respecification procedures that are useful when there is evidence that one or more of these assumptions are not appropriate.

18.1 OUTLIER DETECTION

Outliers are observations that have extreme values relative to other observations observed under the same conditions. Observations may be outliers because of a

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

single large or small value of one variable or because of an unusual combination of values of two or more variables.

Outliers are important for at least two reasons. First, their presence in a data set may obscure characteristics about the phenomena being studied that are present in the bulk of the other data values. Second, outliers may provide unique information about the phenomenon of interest that is not contained in the other observations. Outliers can have deleterious effects on traditional summary statistics; they can influence the results obtained in the analysis of designed experiments, and they can seriously alter the least squares fits of regression models.

Statistical techniques for dealing with the possible presence of outliers in a data set fall into two general categories: identification and accommodation. Identification refers to techniques used to determine whether any outliers exist in a data set and, if so, which observations are outliers. Accommodation of outliers refers to techniques used to mitigate their effects. These techniques include the deletion of outliers, trimming extreme observations to less extreme values, using outlier-resistant estimators, respecifying the assumed model, or collecting additional data. The focus of this section is on the identification of outliers.

18.1.1 Univariate Techniques

For univariate models of the form $y_i = \mu + e_i$, the graphical techniques introduced in previous chapters are useful for identifying potential outliers in one variable. These techniques include box plots and normal quantile–quantile plots. We recommend the use of normal quantile–quantile plots, in part because many of the statistical procedures recommended in this book are based on an assumption of normality.

Table 2.1 in Chapter 2 contains fuel-economy measurements for four vehicles, each using eight different fuels. Figure 18.1 displays normal quantile–quantile plots of the fuel-economy data for the Volkswagen and the Peugeot. Outliers in a quantile–quantile plot usually appear as one or more values at the extremes that depart from a line that passes through the bulk of the remaining observations. The argument that the extreme values that depart from a line through the remaining points are outliers is offered cautiously, because there are often a few points that depart modestly from a straight line at the extremes of quantile-quantile plots even when the data do conform to the reference distribution.

The plot of the Peugeot values in Figure 18.1a is not sufficient to label the largest observation an outlier, because its departure is not large enough to justify an unequivocal conclusion. Contrast the uncertainty about the largest observation in Figure 18.1a with the gross departure of the smallest fuel-economy value for the Volkswagen from the remaining seven values in

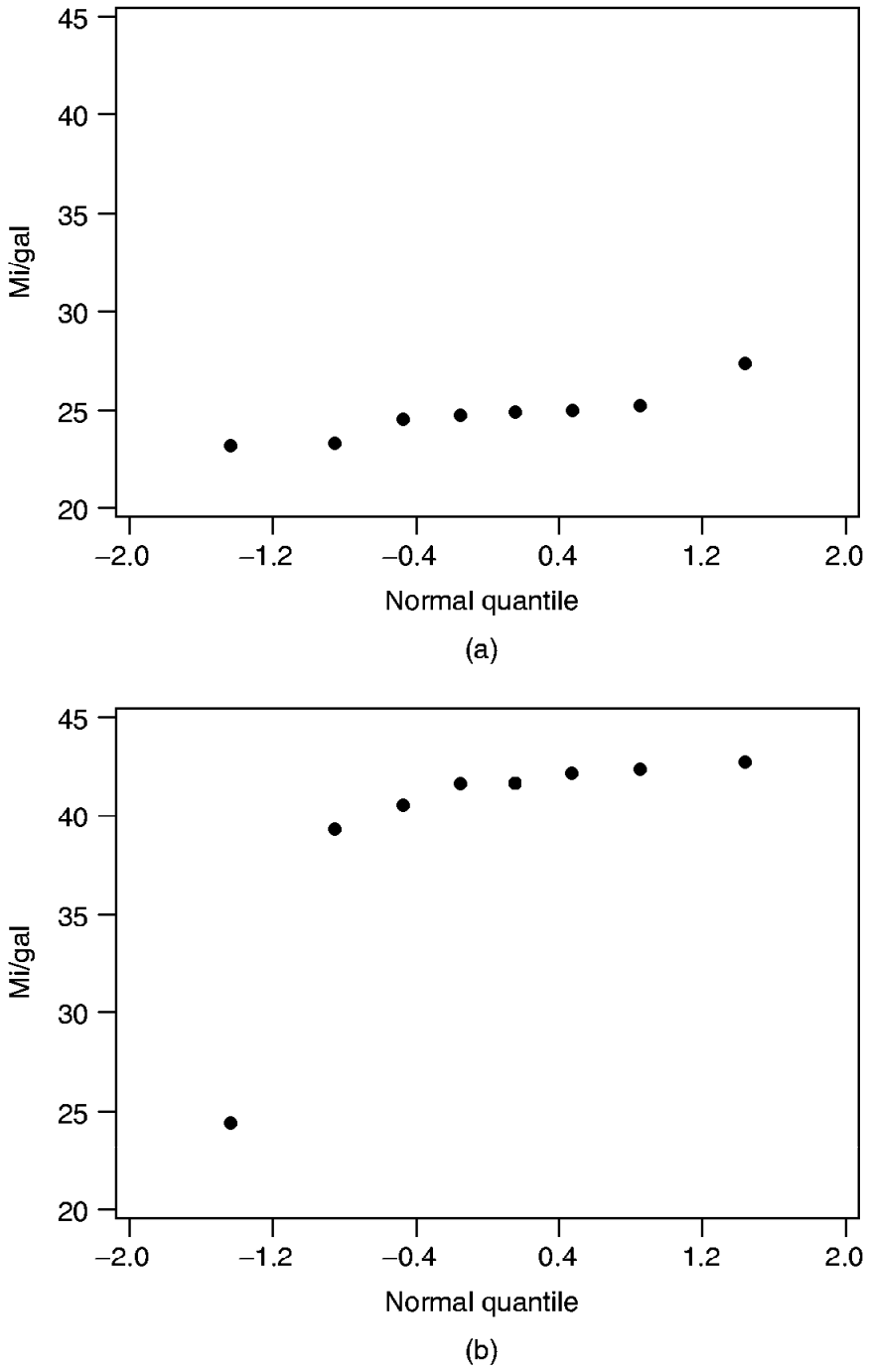


Figure 18.1 Fuel-economy normal quantile plots: (a) Peugeot, (b) Volkswagen.

Figure 18.1b. This large a deviation from a straight line through the other seven points is beyond the modest departures which might be expected of extreme values.

Statistical tests for outlying observations are available to help confirm patterns seen in plots of the data. These statistical tests are especially valuable

when extreme observations exhibit modest departures from the bulk of the other observations in the various plots mentioned above.

A commonly used statistic for detecting outliers among observations on a single variable is the *Grubbs* test. This test uses ratios of two sums of squares. The numerator sum of squares does not contain the suspect observations. The denominator sum of squares contains all the observations, including the suspect values. Grubbs tests are based on an assumption of normally distributed model errors (other than the outliers) in Equation (3.1). The Grubbs test, detailed in Exhibit 18.1, is now applied to the fuel economy data in Figure 18.1

EXHIBIT 18.1 GRUBBS TEST FOR OUTLIERS

1. Let $y_{(i)}$ denote the ith smallest observation; i.e., let the ordered responses be denoted by
$$y_{(1)} \leqslant y_{(2)} \leqslant \cdots \leqslant y_{(n)}.$$
2. If the k largest values in a data set are suspected as outliers, calculate
$$L_k = \frac{1}{S_{yy}} \sum_{i=1}^{n-k} (y_{(i)} - \bar{y}_L)^2, \tag{18.1}$$
where $\bar{y}_L = \sum_{i=1}^{n-k} y_{(i)}/(n-k)$ and $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$. If the k smallest values in a data set are suspected as outliers, calculate
$$S_k = \frac{1}{S_{yy}} \sum_{i=k+1}^{n} (y_{(i)} - \bar{y}_S)^2, \tag{18.2}$$
where S_{yy} is as defined above and $\bar{y}_S = \sum_{i=k+1}^{n} y_{(i)}/(n-k)$.
3. Conclude that the group of k observations are outliers if the calculated value of L_k or S_k is less than the critical value given in Table A10 of the appendix.
4. If the k most extreme observations are suspected as outliers but some are the largest values in the data set while others are the smallest ones, calculate
$$E_k = \frac{1}{S_{yy}} \sum_{i=1}^{n-k} (z_{(i)} - z_E)^2, \tag{18.3}$$
where $z_{(i)}$ is the y_i corresponding to the ith smallest value of the ordered absolute values of the deviations, i.e., the y_i corresponding to ith smallest $|y_i - \bar{y}|$. The value of $\bar{z}_E$ is the average of the y_i corresponding to the $n-k$ smallest deviations. Conclude that the group of k observations are outliers if the calculated value of E_k is less than the critical value given in Table A11 in the appendix.

The calculated value of $S_k(k = 1)$ for the Volkswagen fuel-economy data that corresponds to the lowest test result is

$$S_1 = 8.714/265.429 = 0.033.$$

The p-value associated with this statistic (see Table A10) is less than 0.01. Thus the lowest observed fuel-economy value is not consistent with the remainder of the data. As noted in Chapter 2, this low test result is not an error in the experimentation; it was confirmed in subsequent test runs. It is, nevertheless, significantly different from the average of the other Volkswagen fuel economy measurements.

The value of L_1 associated with the highest fuel-economy test result for the Peugeot is

$$L_1 = 4.000/11.875 = 0.337.$$

This value of the test statistic is not significant at the 0.05 level of significance ($0.05 < p < 0.10$).

The procedures just discussed for outlier testing should not be used without an accompanying plot of the data. If the statistics L_k, S_k and E_k are routinely used as outlier diagnostics without any visual inspection of data sets, the researcher may be misled due to problems such as *masking*.

Masking occurs when two or more outliers have similar values. If a data set contains, say, two large values that are similar in magnitude, an outlier test for one of the observations ordinarily will not be statistically significant. This is because both the numerator and the denominator of the statistics (18.1) to (18.3) will be large, due to the presence of one outlier in the numerator and both outliers in the denominator. Only a test for both the two largest observations will be statistically significant. Plotting the data provides a measure of protection against this difficulty by identifying sets of outlying observations.

One other caution is important to understand when considering the use of either graphical or computational outlier-detection techniques. If all model assumptions (e.g., fixed-effects models with independent normal errors) are correct, test statistics for single outliers will identify observations as being outliers $100\alpha\%$ of the time; that is the proportion of observations erroneously identified as outliers will equal the significance level of the test. Tests for groups of outliers also will erroneously lead to the conclusion that the group of observations significantly deviates from the bulk of the remaining data $100\alpha\%$ of the time. For these reasons, a small significance level should be used with these tests, for example, 0.01 or smaller.

With the above two cautions in mind, a general procedure for identifying outliers is outlined in Exhibit 18.2.

EXHIBIT 18.2 OUTLIER DETECTION

1. Plot responses using point plots, sequence plots, box plots, or normal quantile–quantile plots. Note any unusual trends or clustering of observations, as well as all extreme observations.
2. Based on the plots constructed in step 1, a scan of the data set, or any other analysis deemed appropriate, select the observation or group of observations that are suspected to be outliers.
3. Use a statistical test for outliers with a small significance level to determine whether the group of observations deviate significantly from the bulk of the observations.

18.1.2 Response-Variable Outliers

Outliers in response variables can occur in the analysis of data from designed experiments and in regression analyses from either designed experiments or observational studies. In designed experiments outliers usually do not occur in the factor variables. This is because the levels of the factor variables are generally selected by the experimenter according to a balanced arrangement in the factor space. In regression analyses using either observational data or predictor variables that cannot satisfactorily be controlled, outliers can occur in the predictor variables. This section presents outlier detection techniques for response variables from designed experiments or regression analyses. Outlier diagnostic techniques for predictor variables in regression models are discussed in section 18.1.3.

Outlier-detection techniques for response variables are based on residuals: differences between observed (y) and predicted ($\hat{y}$) responses, $r = y - \hat{y}$. Residuals were introduced in Section 14.4 in the context of fitting regression models. Residuals can be defined similarly for data from designed experiments that are analyzed using ANOVA techniques. In designed experiments the predicted values are obtained by inserting estimates of all fixed effects for the corresponding model parameters [e.g., estimates of the parameters in Equation (6.2)]. Any solution to the normal equations can be inserted for the model parameters to obtain the predicted response.

Before detailing the value of and techniques for performing a residual analysis, a comment on the use of residuals from fits to ANOVA models is in order. Residuals from ANOVA fits can have characteristics that prevent them from providing relevant information about the corresponding errors in the responses. For example, a saturated (see Exhibit 7.7) ANOVA model for which there are exactly two repeats for each factor–level combination will result in pairs of residuals that have exactly the same magnitudes and are

opposite in sign. This is an algebraic property of the fit and has nothing to do with the actual distribution of the model errors. Hence, residuals from ANOVA fits should be used cautiously for the assessment of error assumptions and for outlier detection purposes. This cautionary note is illustrated and discussed further in Section 18.2.1.

Residual plots are very effective for detecting outliers. Recommended techniques include plotting residuals in scatterplots versus the corresponding predicted responses, each factor or predictor variable, experimental run order, or any other meaningful variable. Outliers generally occur as points far above or below the bulk of the plotted residuals. Other graphical techniques include plotting the residuals in box plots and normal quantile–quantile plots as discussed in the previous section.

Table 18.1 contains 58 observations on yield and four operating conditions (catalyst type, percent conversion, flow of raw material, and ratio of reactants)

TABLE 18.1 Yield and Operating Conditions for a Chemical Process

Obs.	Yield	Catalyst	Conversion	Flow	Ratio
1	55.45	0	11.79	118.9	0.155
2	54.83	0	11.95	105.0	0.089
3	52.21	0	12.14	97.0	0.094
4	50.40	0	12.06	101.0	0.108
5	49.32	0	12.04	44.0	0.100
6	41.36	0	12.28	30.0	0.036
7	49.28	0	12.36	38.0	0.113
8	47.89	0	12.22	32.0	0.123
9	52.40	0	11.90	220.0	0.135
10	52.62	0	11.34	350.0	0.183
11	59.34	0	11.20	160.0	0.166
12	45.32	0	12.03	328.0	0.221
13	48.97	0	12.04	325.0	0.192
14	49.87	0	12.02	330.0	0.188
15	47.06	0	12.02	330.0	0.201
16	40.05	0	12.05	322.0	0.153
17	51.29	0	12.05	322.0	0.194
18	43.73	0	12.14	326.0	0.097
19	51.44	0	12.05	366.0	0.136
20	49.33	0	12.13	350.0	0.143
21	51.85	0	11.94	510.0	0.116
22	52.33	0	12.02	513.0	0.195
23	47.72	0	12.02	523.0	0.160
24	46.65	0	11.80	430.0	0.164
25	49.83	0	11.40	325.0	0.197
26	51.36	0	11.19	380.0	0.233
27	51.19	0	11.19	380.0	0.211

TABLE 18.1 (*continued*)

Obs.	Yield	Catalyst	Conversion	Flow	Ratio
28	48.10	0	11.18	383.0	0.222
29	48.28	0	11.21	375.0	0.223
30	51.28	0	11.21	375.0	0.229
31	47.81	0	11.88	410.0	0.170
32	46.78	0	12.09	485.0	0.163
33	44.24	0	11.97	445.0	0.153
34	53.07	0	11.30	232.0	0.180
35	54.08	0	10.90	220.0	0.126
36	54.19	0	10.90	223.0	0.152
37	54.26	0	10.90	222.0	0.184
38	55.94	0	10.80	240.0	0.225
39	55.85	0	10.90	238.0	0.169
40	56.57	0	10.80	258.0	0.161
41	53.85	0	10.75	272.0	0.197
42	54.81	0	10.72	280.0	0.201
43	54.14	0	11.10	426.0	0.221
44	53.74	0	11.12	404.0	0.215
45	58.40	1	10.90	466.0	0.343
46	51.02	1	11.00	490.0	0.278
47	49.53	1	10.90	483.0	0.280
48	48.50	1	10.90	427.0	0.296
49	51.97	1	10.96	465.0	0.339
50	52.35	1	11.42	527.0	0.306
51	51.83	1	11.60	480.0	0.274
52	44.51	1	11.65	465.0	0.243
53	49.46	1	11.60	150.0	0.159
54	49.58	1	11.50	160.0	0.165
55	51.42	1	11.50	166.0	0.185
56	48.85	1	11.30	166.0	0.165
57	47.42	1	11.50	167.0	0.160
58	48.33	1	11.50	167.0	0.161

for a chemical process. The experimenters believed that yield would be linearly related to conversion, flow, and the reciprocal of the reactant ratio ("invratio"). In addition, a conversion-by-flow interaction term has physical meaning, and catalyst type is suspected to cause a shift in the mean response. Because an interaction term is to be included in the model, we follow the recommendation of Section 15.4 and standardize the continuous predictor variables (conversion, flow, invratio) using the normal-deviate form, Equation (15.27), prior to forming the interaction term. The fit of the five predictor variables to the yield is shown in Table 18.2.

TABLE 18.2 Regression Analysis for Chemical-Process Yield

ANOVA

Source of Variation	df	Sum of Squares	Mean Square	F	p-Value
Regression	5	379.218	75.843	8.34	0.000
Error	52	472.728	9.091		
Total	57	851.946			

Variable*	Coefficient Estimate	t-Statistic	p-Value
Intercept	51.096	110.77	0.000
Catalyst	−2.202	−2.15	0.036
Conversion	−2.182	−4.73	0.000
Flow	−1.107	−2.24	0.030
Conversion × flow	−0.046	−0.10	0.919
Invratio	−1.088	−1.96	0.055

*Normal-deviate standardization for conversion, flow, and invratio.

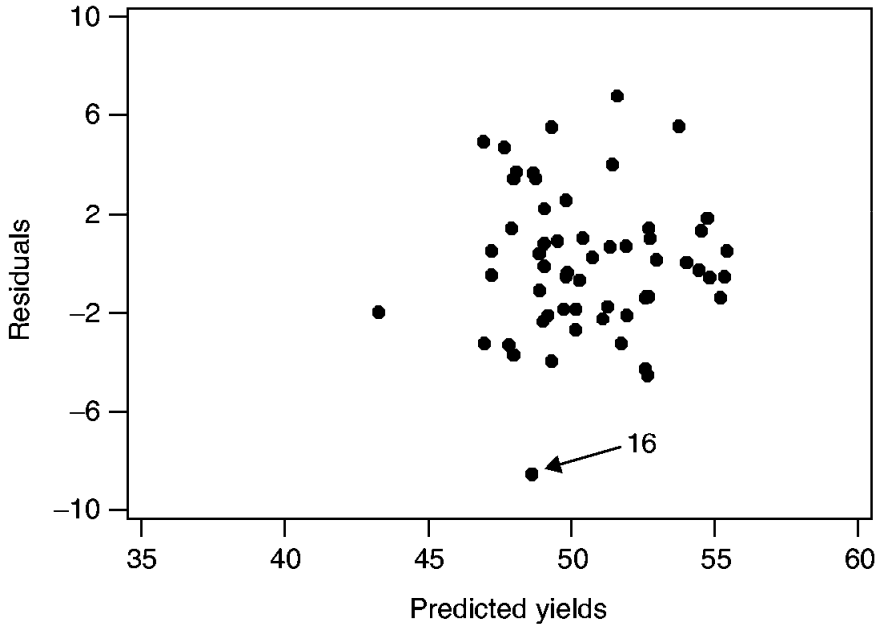


Figure 18.2 Scatterplot of residuals and predicted yields.

Figure 18.2 displays a plot of the residuals versus the fitted values, one of the residual plots suggested above. One large negative residual (observation 16) appears to be separated from the bulk of the remaining residuals. This residual corresponds to the observation having the smallest response in Table 18.1.

Studentized deleted residuals are popular outlier-detection statistics. The computation of studentized deleted residuals is based on deleting an observation from the data set, fitting the regression model with the remaining $n - 1$ observations, and using the fitted model to predict the observation that was deleted from the data set. If the deleted residual is large, an outlier is indicated. Studentized deleted residuals are deleted residuals that have been scaled by independent estimates of their standard errors so they individually follow Student t-distributions (see Exhibit 18.3).

EXHIBIT 18.3 STUDENTIZED DELETED RESIDUALS

Let $\hat{y}_{(-i)}$ denote the predicted response for the ith observation from a fit to a regression or ANOVA model when the ith observation has been deleted from the data set. The ith studentized deleted residual is

$$t_{(-i)} = r_{(-i)}/\widehat{\mathrm{SE}}(r_{(-i)}), \tag{18.4}$$

where

$$r_{(-i)} = y_i - \hat{y}_{(-i)}, \tag{18.5}$$

and $\widehat{\mathrm{SE}}(r_{(-i)})$ is a statistically independent estimate of the standard error of the ith deleted residual $r_{(-i)}$. The studentized deleted residual $t_{(-i)}$ follows Student's t-distribution with $v - 1$ degrees of freedom, where v equals the number of degrees of freedom for error from the fitted model using all the data.

Calculation of studentized deleted residuals is outlined in the appendix to this chapter. Studentized deleted residuals can always be found for regression models once the least-squares estimates are calculated. For some designed experiments, Studentized deleted residuals cannot be calculated. This occurs when there is a single response for a factor–level combination. Deletion of that factor–level combination for some designs does not allow the estimation of all model effects. If all the effects corresponding to the (deleted) combination of interest cannot be estimated, its response cannot be predicted and the studentized deleted residual cannot be calculated.

Table 18.3 lists the raw and the studentized deleted residuals for a subset of the 58 observations for the chemical-process yield example of Table 18.1. The three other statistics (h_i, DFFITS$_i$, DFBETAS) in this table are defined later in this chapter. The large studentized deleted residual ($t_{(-16)} = -3.14$, $p < 0.001$) for observation 16 confirms the visual impression left by the residual plot, Figure 18.2. Note too the large value for observation 45($t_{(-45)} =$

TABLE 18.3 Diagnostics for Influential Observations: Selected Observations, Chemical-Process Yield Data

Observation	h_i	r_i	$t_{(-i)}$	DFFITS$_i$
6	0.834	−1.933	−1.594	−3.566
7	0.250	0.407	0.154	0.089
8	0.207	−1.856	−0.688	−0.351
11	0.084	5.565	1.982	0.559
16	0.043	−8.562	−3.141	−0.668
21	0.134	4.925	1.792	0.705
45	0.144	6.799	2.564	1.050

DFBETAS

Observation	Intercept	Catalyst	Conversion	Flow	Conversion × Flow	Invratio
6	−0.164	−0.425	0.842	−0.999	0.261	−3.158
7	0.020	−0.021	0.037	−0.026	−0.062	−0.039
8	−0.091	0.084	−0.140	0.142	0.220	0.173
11	0.344	−0.099	−0.199	−0.397	0.311	−0.054
16	−0.447	0.132	−0.446	−0.020	−0.078	0.150
21	0.266	0.010	0.083	0.442	0.291	0.398
45	0.033	0.311	−0.293	0.448	−0.606	−0.046

2.56, $p = 0.012$). This observation has the second largest response value in Table 18.1 and is the first observation taken with the second catalyst type. There are now two outliers which the experimenter may wish to study further.

Residual plots are often made with studentized deleted residuals rather than raw residuals r_i. One reason for this is that studentized deleted residuals are conveniently scaled; since they behave like Student t-statistics, most of them should have values between −3 and 3. Another reason for plotting studentized deleted residuals is that they have equal standard errors, unlike raw residuals, which are not equally variable. Both types of plots generally display the same trends and highlight the same outliers.

One of the key concerns of experimenters when deciding how to accommodate the presence of outliers is whether the outliers would seriously affect statistical procedures. Outliers that do affect statistical estimation or inference procedures are termed *influential observations* (see Exhibit 18.4).

EXHIBIT 18.4

Influential Observations. An outlier (extreme observation) is termed an influential observation if its presence in a data set strongly affects parameter estimates or statistical inference procedures.

Studentized deleted residuals are useful diagnostics for influential observations because they are measures of when a data set cannot satisfactorily predict a particular response. A related statistic, referred to by the mnemonic DFFITS, measures the change in the predicted value of the ith response using two fits to the data, one with and one without the ith response:

$$\begin{aligned} \text{DFFITS}_i &= \frac{\hat{y}_i - \hat{y}_{(-i)}}{\widehat{\text{SE}}(\hat{y}_i)} \\ &= \sqrt{\frac{h_i}{1 - h_i}} t_{(-i)}, \end{aligned} \tag{18.6}$$

where $\hat{y}_i$ is the predicted response from the full (all n observations) fit to the model, $\hat{y}_{(-i)}$ is the predicted response from the fit in which the ith observation is deleted, and $\widehat{\text{SE}}(\hat{y}_i)$ is an independent estimate of the standard error of $\hat{y}_i$. As shown in the second expression in Equation (18.6), DFFITS_i is a multiple of the ith studentized deleted residual. The multiplier of $t_{(-i)}$ in (18.6) is a simple function of the ith "leverage value" h_i (see Section 18.1.3).

While similar in form to studentized deleted residuals, DFFITS values measure change in predicted responses. As such they are another measure of influence and are important in their own right. Criteria for assessing whether an observation is influential according to the DFFITS statistic depend on both the sample size and the number of predictor variables in the model. One useful cutoff for regression models with intercept terms is $|\text{DFFITS}_i| > 3[(p + 1)/n]^{1/2}$, roughly corresponding to a Student t-value of 3. This cutoff would use p rather than $p + 1$ for no-intercept models.

Another outlier diagnostic that is especially useful for assessing the influence of the ith observation on the estimation of the jth regression coefficient in a regression analysis is referred to by the mnemonic DFBETAS:

$$\text{DFBETAS}_{ij} = \frac{b_j - b_{j(i)}}{\widehat{\text{SE}}(b_j)}, \tag{18.7}$$

where the two estimates of β_j are from fits with and without the ith observation and $\widehat{\text{SE}}(b_j)$ is an independent estimate of the standard error of the least-squares estimator b_j. A cutoff for DFBETAS_{ij} is $3/n^{1/2}$.

Table 18.3 lists the outlier diagnostics DFFITS_i and DFBETAS_{ij} for selected observations. The cutoff values for DFFITS and DFBETAS are, respectively, $3[(p + 1)/n]^{1/2} = 3(6/58)^{1/2} = 0.96$ and $3/n^{1/2} = 3/(58)^{1/2} = 0.39$. Observation 16 does not have a large value on DFFITS but it does on DFBETAS (intercept, conversion). Observation 45 has a relatively large value for both DFFITS and DFBETAS (flow, conversion $\times$ flow). Note too that observation 6 has large values for DFFITS and DFBETAS (catalyst, flow, invratio)

and the following observations have large values for DFBETAS on at least one predictor variable: observation 11 (flow), observation 21 (flow, invratio).

Because the diagnostics presented in Table 18.3 are from a regression analysis, the influential observations that were identified could be the result of outliers in the response or in the predictor variables. In the next section we discuss techniques for identifying outlying observations in the predictors. We shall return to this example at that time.

Each of the outlier diagnostics, (18.4) to (18.7), appears to require $n + 1$ fits to a data set, one for the full data set and one for each reduced data set in which one observation is deleted. In fact, one can show that all these statistics can be calculated from quantities available after only one fit to the data, the fit to the complete data set. The interested reader is referred to the references and the appendix to this chapter for details.

As in the univariate case, the outlier diagnostics presented in this section can fail to detect extreme observations due to the clustering of two or more observations in the $(p + 1)$-dimensional space of response and predictor (or factor) variables. The techniques described in this section are most beneficial when outliers occur singly. The most effective outlier diagnostics for multiple outliers make use of group deletion of observations. To date there are few economically feasible procedures for identifying an arbitrary number of clustered outliers. The references cite several procedures that have been suggested for identifying multiple outliers.

18.1.3 Predictor-Variable Outliers

Response–predictor and predictor–predictor scatterplots are useful tools for discovering extreme values of the predictor variables. Figure 18.3 shows a scatterplot of yield versus one of the predictor variables, the inverse of the reactant ratio, for the chemical-process example. Observation 6 appears to be an outlier in this plot. This observation has an extremely large value for invratio, a value that removes it from the bulk of the other observations plotted in Figure 18.3. On occasion, a very large value for a predictor variable occurs, but the plotted point is consistent with a trend in the data. The extreme value for invratio does not appear to be consistent with any trend among the other points plotted.

Figure 18.4 is a plot of two of the predictor variables, conversion versus invratio. Again, observation 6 is an outlier on this plot. This observation appears to be an outlier on any plot involving the invratio. It is now clear why the extremely large DFBETAS value (the largest in the entire data set) for the invratio on observation 6 exists. This large value is distorting the estimated regression coefficient for invratio and is also the cause of the large DFFITS value for this observation.

At this juncture in a statistical analysis, the question of accommodation arises — in particular, the deletion of observations. While we urge caution in

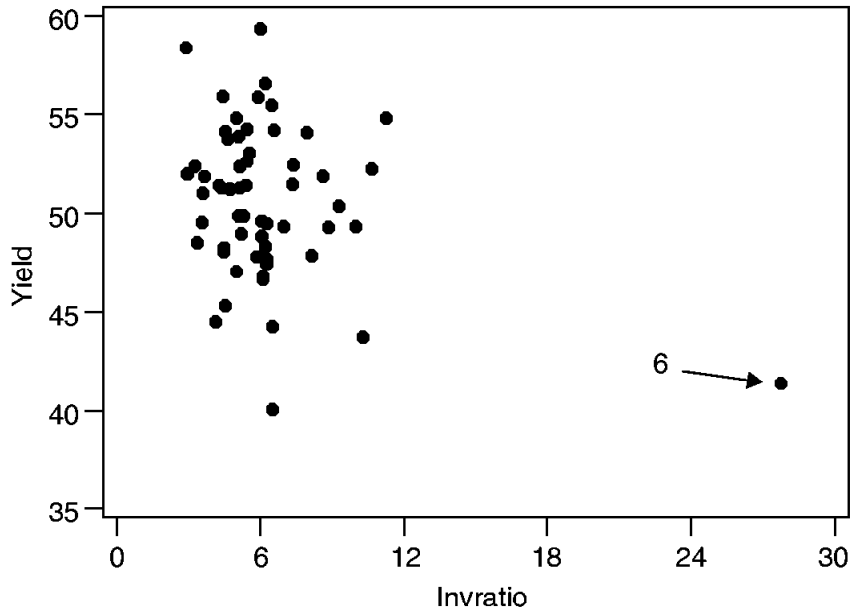


Figure 18.3 Scatterplot of yield and invratio.

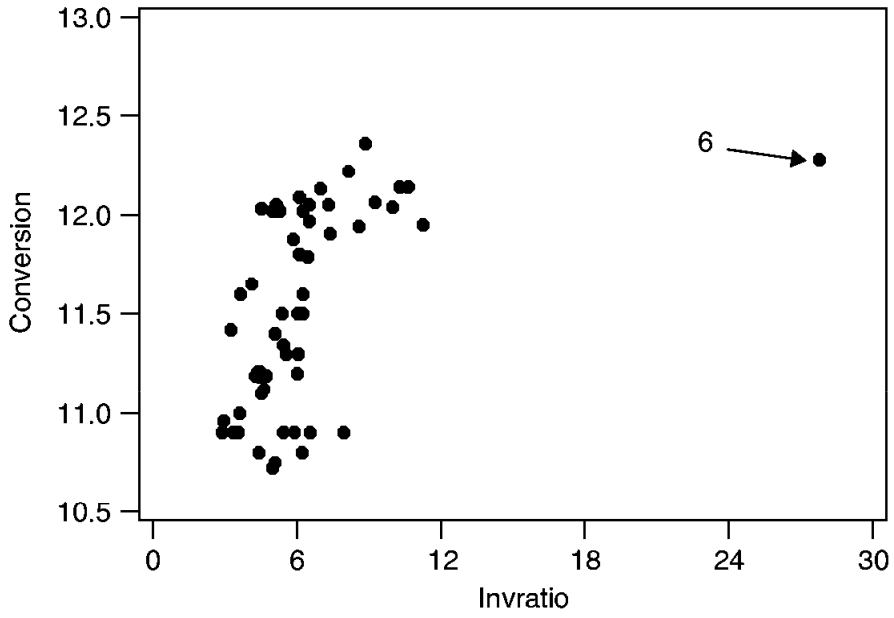


Figure 18.4 Scatterplot of conversion and invratio.

choosing to delete observations, if a decision is to be based solely on outlier diagnostics and if the invratio is to be included in the fitted model, observation 6 should be deleted. There is dramatic evidence that observation 6 is affecting the fit. If the invratio is ultimately removed as a predictor variable, one can reinsert observation 6 into the data set and reevaluate its influence on the fit.

An underlying principle in this decision is that a single observation should not dictate the fit to a data set.

When three or more predictor variables are used in a regression model, extreme combinations of the predictors may not be apparent from inspecting response–predictor and predictor–predictor scatterplots. Unusual combinations of values on response and predictor variables might involve three or more variables, thereby not appearing extreme on any single variable or any combination of two variables. The computation of *leverage values* aids in the identification of unusual combinations of predictor–variable values (see Exhibit 18.5). The calculation of leverage values is described using matrix algebra in the appendix to this chapter.

EXHIBIT 18.5 LEVERAGE VALUES

Leverage values h_{ii} (or h_i) are constants calculated from the values of the predictor variables. Leverage values have the following properties:

(a) $\hat{y}_i = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j, i = 1, 2, \ldots, n$, where the h_{ij} are also constants calculated from the predictor variables;

(b) $0 \leqslant h_{ii} \leqslant 1, -1 \leqslant h_{ij} \leqslant 1$;

(c) $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$; and

(d) h_{ij} measures the distance of the ith set of predictor-variable values from the point of averages of all $p + 1$ predictors (including the constant).

The algebraic properties described above can be summarized as follows. Property (a) states that a predicted response is a multiple of its corresponding observed response, added to a linear combination of the other responses. The importance of this property is that the multiplier on the observed response is the leverage value. Together, properties (b) and (c) state that leverage values are fractions between 0 and 1, and the closer the leverage value is to 1, the closer the h_{ij} are to 0. In the limit, if $h_{ii} = 1$, then $h_{ij} = 0$ $(j \neq i)$. Thus, properties (a) to (c) state that if a leverage value is close to 1, a predicted response is almost completely determined by its observed response, regardless of the adequacy of the fit of the model or the value of the observed response.

An implication of these properties is that leverage values close to 1 force the fit to pass through the observed response for the ith observation. Geometrically, the fitted line or plane must pass through the plotted point for the ith observation, regardless of any trend in the bulk of the remaining observations. Property (d) states that this will occur whenever the predictor-variable values for the ith observation are sufficiently extreme from the remaining observations.

Leverage values are important not only because of the above interpretations, but also because they can be calculated for any number of predictor variables. Leverage values can identify extreme observations when they occur because of an unusual large value on one variable or because of an unusual combination of values on two or more variables.

Figure 18.5 shows a sequence plot of the leverage values h_i for the chemical process example. The cutoff value typically used to identify outliers in the predictors is $h_i > 2(p+1)/n$ for models with intercepts and $h_i > 2p/n$ for no-intercept models. For this example,

$$2\frac{p+1}{n} = 2\frac{6}{58} = 0.21.$$

A scan of either Table 18.3 or Figure 18.5 reveals that observation 6 clearly stands out as an outlier with a leverage value of 0.83. This information reinforces the conclusions made from the scatterplots and the outlier diagnostics discussed earlier in this section and in the last one. Figure 18.5 also shows that observations 7 and 8 are potentially influential ($h_7 = 0.25$ and $h_8 = 0.21$).

Table 18.4 shows a comparison of the regression coefficients from a fit of (1) all the data, (2) the data with observation 6 removed, and (3) the data with observations 6, 7, and 8 removed. Dramatic changes occur with the least-squares estimate of the invratio when observation 6 is removed. Only small changes occur when the other two observations are removed. This finding is consistent with the suggestions of the scatterplots, leverage values, and DFBETAS.

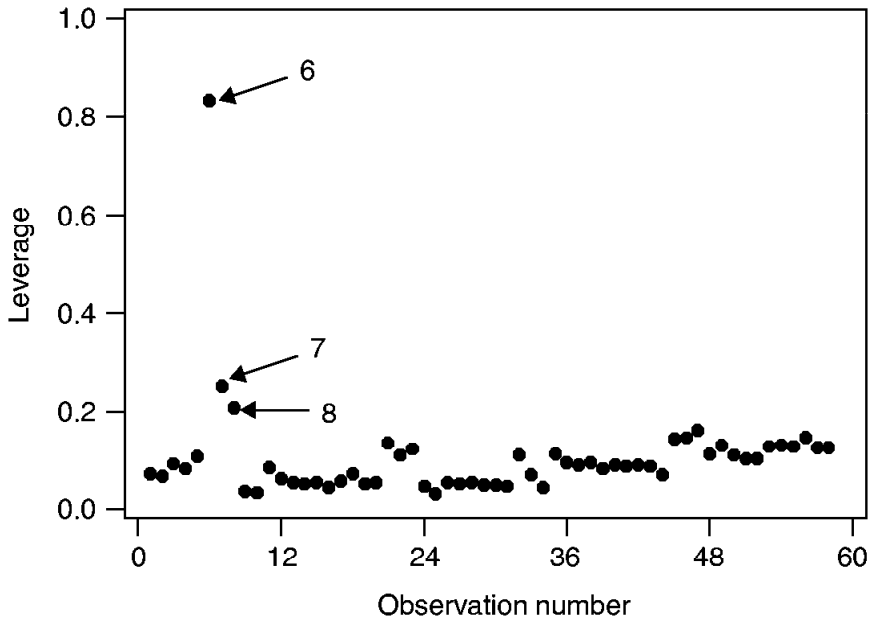


Figure 18.5 Sequence plot of leverage values.

TABLE 18.4 Comparison of Regression Coefficients with Influential Observations Removed: Chemical-Yield Data

Variable*	Complete Data Set	Observation 6 Deleted	Observations 6, 7, 8 Deleted
Intercept	51.097	51.171	51.217
Catalyst	−2.202	−1.173	−1.877
Conversion	−2.170	−2.565	−2.486
Flow	−1.107	−0.620	−0.703
Conversion × flow	−0.046	−0.165	−0.281
Invratio	−1.088	0.636	0.508

*Normal-deviate standardization for conversion, flow, and invratio.

18.2 EVALUATING MODEL ASSUMPTIONS

The procedures presented in this section emphasize techniques for verifying several model assumptions that are commonly used in statistical inference procedures. We focus on the evaluation of a specific statistical model for which assumptions of normality and correct specification of response and predictor variables is of concern.

18.2.1 Normally Distributed Errors

One of the central assumptions used in many statistical procedures is that the model errors are normally distributed. Normally distributed errors were assumed for single-sample models, ANOVA models, and regression models in earlier chapters of this book. While many statistical procedures (e.g., nonparametric tests) do not require an assumption of normality and others are fairly insensitive to such assumptions, the assumption of normality is required for the exact sampling distributions of many statistics to be valid. Some statistics (e.g., F-tests for the equality of two population variances) are extremely sensitive to departures from normality.

The error terms in statistical models are unobservable. Most statistical procedures for assessing the assumption of normally distributed errors thus rely on using residuals in place of the model errors.

One of the easiest and most direct evaluations of the assumption of normality is the normal quantile–quantile plot. This plotting technique was introduced in Chapter 5 and discussed further in Section 18.1 in the context of outlier identification. A normal quantile–quantile plot consists of plotting the ordered residuals from a fitted model against the corresponding ordered quantiles from a standard normal reference distribution. Approximate linearity indicates that the sample data are consistent with an assumption of normally distributed

TABLE 18.5 Predicted Responses for Torque Study

Obs.	Sleeve	Lubricant	Torque	Predicted Torque
		Alloy: Steel		
1	Porous	1	82	79.0
2	Porous	1	76	79.0
3	Porous	2	75	77.5
4	Porous	2	80	77.5
5	Porous	3	77	77.0
6	Porous	3	77	77.0
7	Porous	4	76	74.5
8	Porous	4	73	74.5
9	Nonporous	1	78	78.5
10	Nonporous	1	79	78.5
11	Nonporous	2	65	67.0
12	Nonporous	2	69	67.0
13	Nonporous	3	79	77.5
14	Nonporous	3	76	77.5
15	Nonporous	4	81	79.0
16	Nonporous	4	77	79.0
		Alloy: Aluminum		
17	Porous	1	79	78.0
18	Porous	1	77	78.0
19	Porous	2	72	72.0
20	Porous	2	72	72.0
21	Porous	3	77	78.5
22	Porous	3	80	78.5
23	Porous	4	72	73.5
24	Porous	4	75	73.5
25	Nonporous	1	71	69.5
26	Nonporous	1	68	69.5
27	Nonporous	2	73	70.5
28	Nonporous	2	68	70.5
29	Nonporous	3	69	69.0
30	Nonporous	3	69	69.0
31	Nonporous	4	70	70.0
32	Nonporous	4	70	70.0

errors. These plots can be supplemented with other data displays such as relative-frequency histograms (Section 1.3) of the residuals.

In Section 18.1 normal quantile-quantile plots for regression models were introduced. In order to demonstrate their use with ANOVA models, we reconsider

the torque study that was designed in Section 5.1 (Table 5.3). The complete data set, including the duplicate responses for each factor–level combination, is shown in Table 18.5.

Because this is a balanced, complete factorial experiment and a complete three-factor interaction model was used to fit the data, the predicted responses are simply the averages of the duplicate measurements for each alloy–sleeve–lubricant combination. These averages are listed next to the observed-torque values in Table 18.5. The residuals are listed in order in Table 18.6 and are plotted in Figure 18.6. The residual plot exhibits reasonable fidelity to a straight line. Thus, the assumption of normally distributed errors for these data appears to be reasonable.

Because graphical evaluations are subjective, it is useful to supplement normal quantile plots with appropriate tests of statistical hypotheses. Generally regarded as the most powerful statistical test of normality is the Shapiro–Wilk test (see Exhibit 18.6).

The assumptions underlying the use of the Shapiro–Wilk statistic (18.8) require independent observations. The statistic is appropriate for use when a single sample of independent observations is to be tested for normality. When observations are not independent, as with the use of residuals from analysis of variance and regression models, the statistic provides only an approximate test

TABLE 18.6 Ordered Torque Residuals and Constants for the Shapiro–Wilk Test for Normally Distributed Errors

Ordered Residual	a_i	Ordered Residual	a_i
−3.0	−0.4188	0.0	0.0068
−2.5	−0.2898	0.0	0.0206
−2.5	−0.2463	0.0	0.0344
−2.0	−0.2141	0.0	0.0485
−2.0	−0.1878	0.5	0.0629
−1.5	−0.1651	1.0	0.0777
−1.5	−0.1449	1.5	0.0931
−1.5	−0.1265	1.5	0.1093
−1.5	−0.1093	1.5	0.1265
−1.5	−0.0931	1.5	0.1449
−1.0	−0.0777	1.5	0.1651
−0.5	−0.0629	2.0	0.1878
0.0	−0.0485	2.0	0.2141
0.0	−0.0344	2.5	0.2463
0.0	−0.0206	2.5	0.2898
0.0	−0.0068	3.0	0.4188

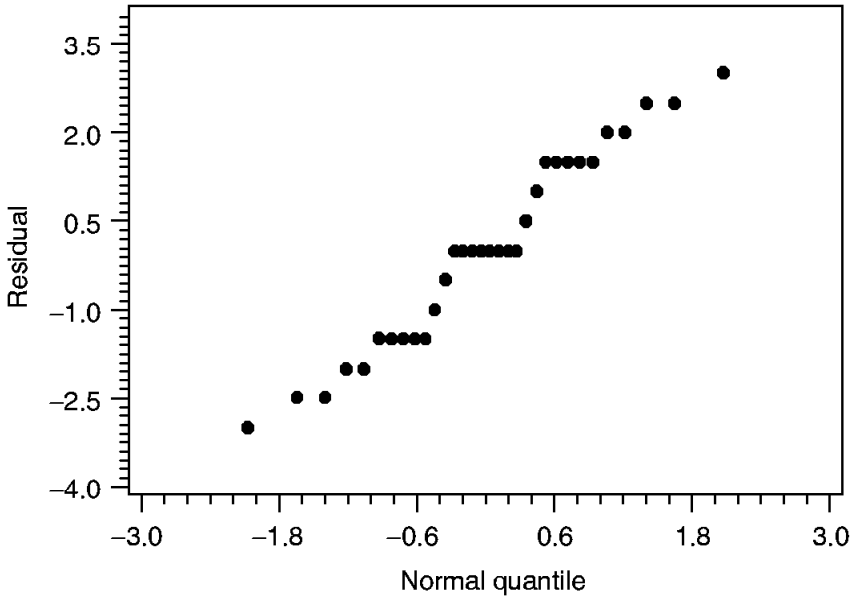


Figure 18.6 Normal quantile plot for torque residuals.

for normality. Its use is still recommended as a valuable statistical measure to augment the use of residual plots.

EXHIBIT 18.6 SHAPIRO–WILK TEST FOR NORMALITY

1. Let $y_{(i)}$ denote the ordered observations, either response values or residuals, as appropriate:
$$y_{(1)} \leqslant y_{(2)} \leqslant \cdots \leqslant y_{(n)}.$$
2. Compute the numerator of the sample variance of the observations:
$$s^2 = \sum (y_{(i)} - y)^2.$$
3. Using the values of a_i found in Table A12 of the appendix, calculate
$$b = \sum a_i y_{(i)}.$$
4. Compute the Shapiro–Wilk statistic
$$W = b^2/s^2. \tag{18.8}$$
5. Reject the assumption of normality if W is less than the value in Table A13 of the appendix.

Table 18.6 lists the constants from Table A12 needed to compute the Shapiro–Wilk statistic (18.8) for the torque data. For this data set, $b = 8.9359$, $s^2 = \text{SS}_E = 84.00$, and $W = 0.9506$. From Table A13, this value of W is not statistically significant ($0.10 < p < 0.50$). Hence, the hypothesis of

normality is not rejected, and statistical inference procedures that require the assumption of normally distributed observations may be used.

When the sample size exceeds 50 several alternatives to the Shapiro–Wilk test are available. Two popular tests are the Kolmogorov–Smirnov test and the Anderson–Darling test. Information on these tests can be found in the references listed at the end of the chapter.

As with most statistical procedures, sample-size considerations have an important effect on the performance of statistical tests for normality. Very small sample sizes may lead to nonrejection of tests for normality because of little power in the test statistics. So too, very large sample sizes may lead to rejection when only minor departures from normality are exhibited by a data set. This occurs because of the extremely high power levels for the tests. These test procedures should always be accompanied by residual plots, especially quantile–quantile plots (or histograms). An assessment using all of these procedures will lead to an informed evaluation of the departure, if any, from normal assumptions and the need for accommodation due to nonnormality.

We mention in closing this section that when ANOVA models are saturated (i.e., contain all main effects and interactions), as in the torque example, and only two repeat tests are available, the normal quantile–quantile plot and tests for normality will not be sensitive to many types of skewed error distributions. This is because the residuals for each factor–level combination will be equal in magnitude and opposite in sign. Thus, the distribution of the calculated residuals will be symmetric, regardless of the true error distribution. We used the torque data to illustrate the above procedures for ANOVA models, not to draw a definitive conclusion about its error distribution.

18.2.2 Correct Variable Specification

Correct model specification implies that all relevant design factors, covariates, or predictor variables are included in a statistical model and that all model variables, including the response, are expressed in an appropriate functional form. Model specification is most acutely a problem in regression models, especially when the functional form of an appropriate theoretical model is unknown.

The numerical values of least-squares residuals can be shown to be uncorrelated with the predictor-variable values in a regression model. For this reason, a plot of residuals against any predictor variable should result in a random scatter of points about the line $r = 0$. Any systematic trend in the plot indicates the need for some reexpression of either the response variable or the predictor variable.

The most frequently occurring patterns indicating a need for reexpression of variables in residual plots are either wedge-shaped or curvilinear trends. A wedge shape in a plot of the residuals versus the predictor variables usually

indicates that the error standard deviation is not constant for all values of the predictor variable. The question of whether to transform the response or the predictor variable cannot be unequivocally answered. A useful guideline is that if the wedge shape occurs in only one of the plots, the predictor variable should be transformed. If it occurs in two or more plots, the response should probably be transformed.

A curvilinear trend in a plot of residuals versus a predictor variable often indicates the need for additional variables in the model or a reexpression of one or more of the current model variables. For example, a quadratic trend in a plot may indicate the need for a quadratic term in x_j. Such a trend might also be indicative of the need to add a variable to the model because the error terms are reflecting a systematic, not random, pattern. The latter would especially be true if patterns were present in several of the residual plots.

Residual plots of this type also are helpful in visually assessing whether an interaction term $x_j x_k$ should be added to the model. A plot of the residuals from a fit excluding the interaction term $x_j x_k$ versus the interaction term should be a random scatter about zero. Any systematic departure from a random scatter suggests the need for an interaction variable in the model.

Another useful residual plot for detecting model misspecification is termed the *partial-residual plot.* A partial residual is defined in Exhibit 18.7. Partial residuals measure the linear effect of a predictor variable relative to the random error component of the model. As suggested by Equation (18.9), if a predictor variable has a strong linear effect on a response variable, the partial residuals for that predictor should be dominated by the linear term $b_j x_{ij}$. If the linear effect is weak or negligible, the randomness present in the least-squares residuals should dominate the partial residual. If some nonlinear function of the predictor variable is needed, the partial residual will often reflect the nonlinear function because the least-squares residuals will contain that portion of the misspecification.

EXHIBIT 18.7 PARTIAL RESIDUALS

Partial residuals corresponding to the predictor variable x_j adjust the least-squares residuals for the portion of the regression fit attributable to the linear effect of x_j:

$$r_i^* = r_i + b_j x_{ij}. \tag{18.9}$$

By plotting the partial residuals versus the predictor-variable values x_{ij}, the direction and strength of the linear effect of x_j on the response can be visually gauged. The regression of r_i^* versus x_{ij} is a line with an intercept of zero and a slope equal to b_j, the least-squares estimate of the coefficient of x_j in the full model. This is in contrast to an ordinary residual plot of r_i versus x_{ij},

in which the slope is always zero. This property of partial residuals enhances their ability to portray the extent of nonlinearity in a predictor variable.

An example of an ordinary residual plot is given in Figure 18.7. The residuals plotted in this figure are from a straight-line fit of the solvent weights to the production rates using the data in Table 15.8. Observe that the first few plotted points are all positive in Figure 18.7 and most of the last several are also positive. In the middle, the predominance of plotted points have negative residuals. This type of curvilinear trend suggests that a quadratic term in the rate variable might need to be added to the model.

Figure 18.8 is a partial-residual plot for the fit to the solvent weights. Observe the strong linear component to the plot and the variation of the points around the linear trend. The clarity of the linear trend relative to the variation around it attests to the strength of the linear effect of production rate on solvent weights.

It is also clear from Figure 18.8 that the plotted points are curving upward. The upward trend is not as strong as the linear trend; nevertheless, a linear trend does not adequately capture the entire relationship between solvent weights and production rates. The ability to visually assess the strengths of nonlinear and linear effects relative to one another and to the random error variation is the primary reason for the use of partial-residual plots.

A careful examination of Figures 18.7 and 18.8 indicates that not only a nonlinear trend but also increasing variability of the residuals as a function of the production rates is evident in the plots. This characteristic, like a wedge-shaped plot of residuals, suggests the need for a transformation of the response variable. We confirm this need in Section 18.3.

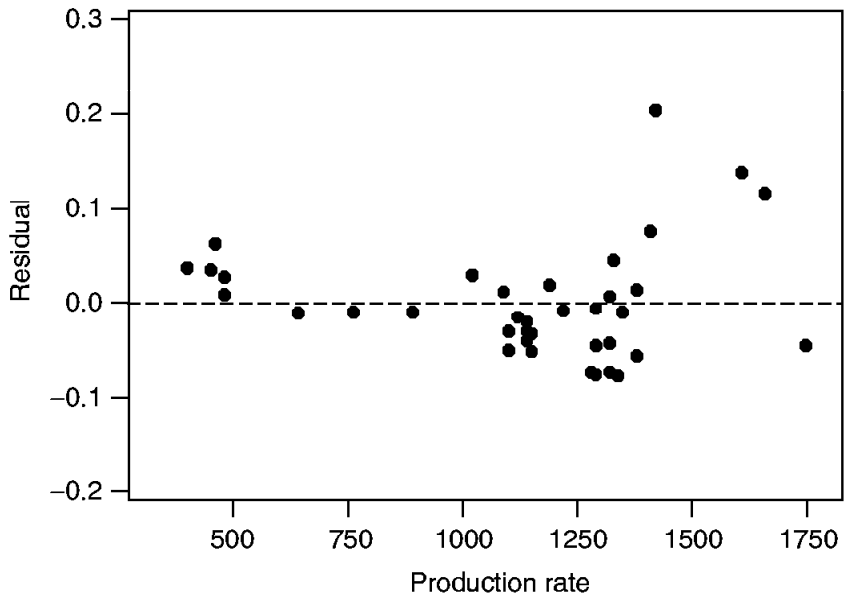


Figure 18.7 Residual plot of linear fit to solvent weights.

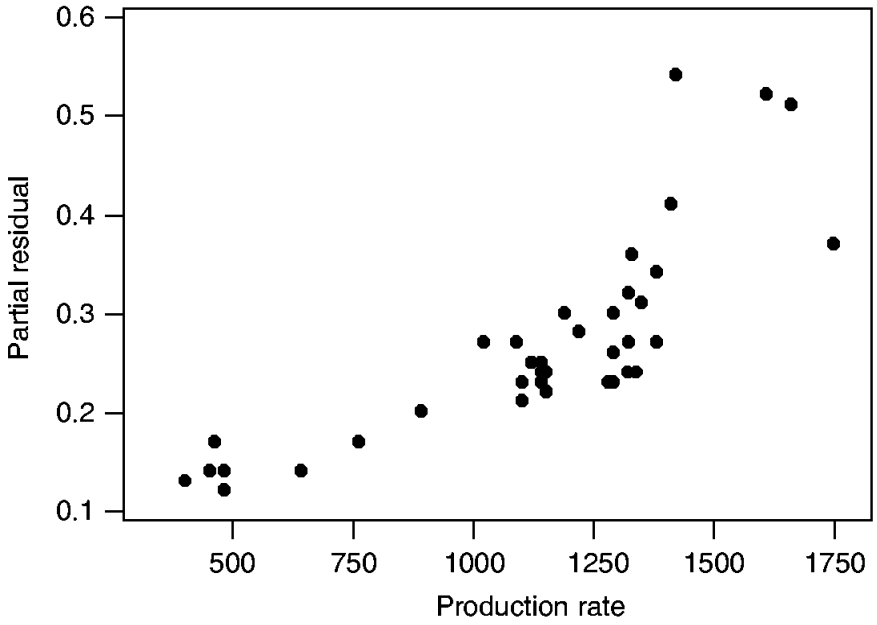


Figure 18.8 Partial-residual plot of solvent weight data.

A type of plot that is closely related to a partial residual plot is termed a *partial-regression* or *partial-regression leverage* plot. In this type of plot the response variable and one of the predictor variables are regressed on the other $p - 1$ predictor variables in two separate regressions. The response residuals are then plotted against the residuals of the predictor variable. A least-squares fit to the plotted points, as with the partial-residual plot, has a zero intercept and a slope equal to the least-squares estimate b_j from a fit to the complete data set.

18.2.3 Nonstochastic Predictor Variables

Random values for predictor variables arise in many experiments in which regression models are used. Random predictor variables occur in observational studies in which the predictor variables are not controlled and in designed experiments in which the predictor variables are subject to measurement errors. Measurement errors can occur, for example, when the values of the predictor variables are obtained from equipment settings or chemical analyses.

As an illustration, consider the design layout given in Figure 18.9. The design was constructed for a study of the effects of fuel properties on vehicle emissions. Illustrated are the target design properties (the circles of the cube) and the measured fuel properties achieved by the blending process (the points connected by the dashed lines). Although this was to be a designed experiment with controlled observations, there was variation in the factor levels due to the blending process and the measurement of the fuel properties.

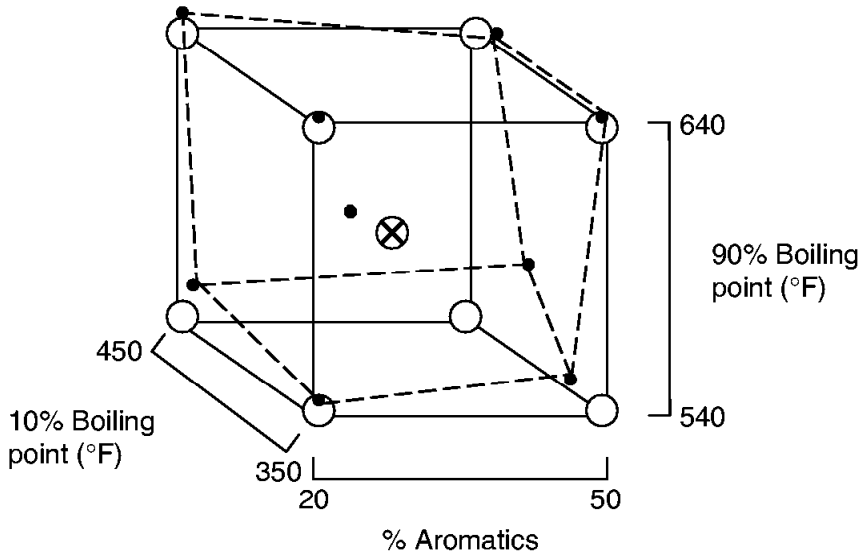


Figure 18.9 Target and measured factor–level values for fuel blends.

Accommodation of random factor or predictor variables is highly problem-dependent. We offer a few suggestions, realizing that they may not be appropriate in all circumstances.

In the analysis of fixed-effect ANOVA models, key considerations are how much measured factor levels differ from target values and the size of the likely effect on the response variable. If the difference between measured and target values is small relative to the difference between two adjacent target values, the effect of using the target factor levels instead of the measured ones in a statistical analysis is likely to be very small. Likewise, if changes anticipated in the response variable between the target and actual factor levels are likely to be small relative to changes expected between two target levels, use of the target levels should not substantially affect inferences. Two-level factorial and fractional factorial experiments are known to be robust with respect to measurement errors in the predictor variables.

In both of the settings described in the last paragraph, one could use the measured factor levels in an ANOVA model or in a regression model, as appropriate. The recommendation to use the target values is based on a presumption that the experiment was designed and, therefore, that the target factor levels were chosen in part to provide a suitable analysis with unconfounded factors. Use of measured factor levels may introduce confounding (Section 7.1) or the need to use the principle of reduction in sums of squares (Section 8.1) to analyze the data, whereas use of target factor levels would not add this complexity.

In regression analyses, measured predictor-variable values are most often used instead of target values. This is especially true when random predictor variables are a result of data being collected in an observational study. If

the observations on the predictor variables represent typical values for the predictors, then the random nature of the predictor variables is not a problem. Inferences are simply conditioned on the observed predictor values. To use this line of argument, the burden is on the researcher to ensure that these predictor-variable values are indeed representative of the phenomenon under study.

18.3 MODEL RESPECIFICATION

A necessary phase of a comprehensive regression analysis (PISEAS, Sec tion 14.2) is the formulation and specification of the regression model. Two basic assumptions (see Table 14.1) about the specified model are that the functional relationship between the response and the predictor variables is linear in the unknown regression coefficients and that the model errors are additive and normally distributed with mean zero and constant standard deviation. The two previous sections of this chapter present techniques for assessing the validity of these model assumptions. Information from such an assessment or known behavior of the phenomenon being studied may indicate that an initial model specification violates one or both of these assumptions.

Models that exhibit nonlinear relationships between the response and the predictor variables sometimes can be respecified as a linear model by suitable reexpressions of the response or (a subset of) the predictor variables or both. One value of such reexpressions is that they allow the procedures developed for linear models to be applied to nonlinear relationships between a response and a set of predictors. Reexpression of the model may also be necessary when the model is assumed to be linear but the errors do not have constant standard deviations (for example, when the magnitude of the error standard deviation is proportional to the mean of the response), or the error distribution is believed to be nonnormal.

Respecification of a nonlinear model is not, however, always a viable option. Reexpression of a nonlinear model might result in a complicated linear form for the deterministic component of the model and loss of the original measurement metric. Thus, at times a nonlinear model is the candidate model of choice because it provides a more parsimonious description of the relationship between response and predictor variables. Likewise, nonlinear models are sometimes the models chosen for fitting purposes because of the physical nature of the problem. For example, in engineering experiments the response of interest may be represented by the solution of a differential equation that is nonlinear in the model parameters.

In this chapter we discuss reexpressions of response and predictor variables. We consider the effects these reexpressions have on inferential procedures and interpreting the regression results.

18.3.1 Nonlinear-Response Functions

Regression models of the form (14.1) and (15.1) are termed linear because the unknown regression coefficients are required to appear as either additive constants or multipliers of the predictor variables. Nonlinear regression models are another important and useful family of regression models. Of particular interest in this section are a class of nonlinear models for which the response variable is a nonlinear function of the predictor variables. Such nonlinear response functions are defined in Exhibit 18.8.

EXHIBIT 18.8 NONLINEAR RESPONSE FUNCTIONS

Nonlinear response functions relate a response variable y to a set of predictor variables $x_1, x_2, \ldots, x_p$ using a functional form that is not linear in the unknown regression coefficients:

$$y_i = f(x_{ij}, \beta_k, e_i), \quad i = 1, 2, \ldots n, \quad j = 1, 2, \ldots p, \quad k = 0, 1, \ldots q, \quad (18.10)$$

where β_k is the kth unknown regression coefficient, x_{ij} is the ith value of the jth predictor variable, $f(x_{ij}, \beta_k, e_i)$ cannot be expressed as in Equation (15.1), and e_i is the ith random error component of the model.

For notational convenience, the model (18.10) is symbolically represented as a function of only a typical predictor-variable value x_{ij} and a typical regression coefficient β_k. It should be understood, however, that the response y_i is a function of the ith values of all p predictor variables, all $q + 1$ regression coefficients, and the ith error.

Knowledge of theoretical relationships between the response and the predictor variables may indicate what reexpressions would be helpful in linearizing a response function $f(x_{ij}, \beta_k, e_i)$. In addition, scatterplots of the response versus the predictor variables may suggest forms of reexpression.

Figure 18.10 shows plots of several common functional forms of a response y versus a single predictor variable x. The different shapes displayed in Figure 18.10 are rich in variety. Minor modifications to the expressions, such as location changes, allow even more flexibility. Figures 18.10*a* and *d* are linear in the regression coefficients, while *b* and *c* are nonlinear. Both of the nonlinear relationships can be linearized by taking logarithms of both sides of the equation.

Logarithmic reexpression of response or predictor variables is perhaps the most frequently used transformation of nonlinear models. While the natural logarithm (base e) is often used, the common logarithm (base 10) or a logarithm to any convenient base can be used, since all such transformations are multiples of one another:

$$\log_a y = \frac{\ln y}{\ln a},$$

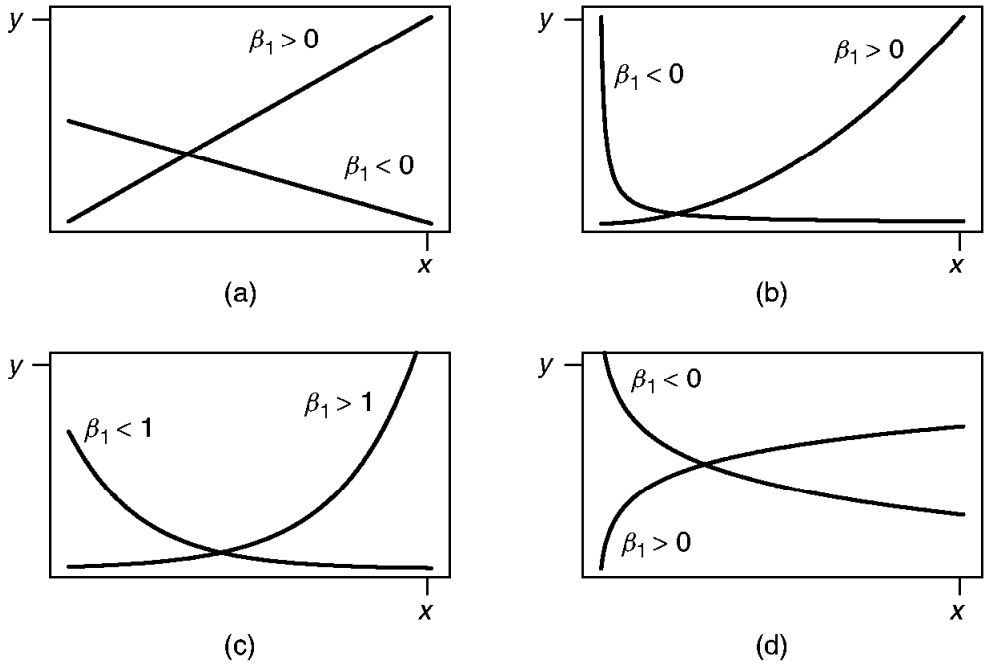


Figure 18.10 Common linear and nonlinear functional relationships. (a) Linear: $y = \beta_0 + \beta_1 x$. (b) Power: $y = \beta_0 x^{\beta_1}$. (c) Exponential: $y = \beta_0 \beta_1^x$. (d) Logarithmic: $y = \beta_0 + \beta_1 \ln x$.

where $\log_a y$ denotes the logarithm to the base a and ln y denotes the natural logarithm. Thus, any logarithmic reexpression has the same ability to linearize as does the natural logarithm; the investigator can use any base that is convenient.

There is a connection between reexpression and plotting on logarithmic graph paper. Engineering and scientific data are often plotted on either semilog graph paper (one of the axes has a logarithmic scale and the other has an arithmetic scale) or log–log paper (both axes have a logarithmic scale). The intent of such plots is to determine a combination of axis scales that result in a linear relationship. This is equivalent to reexpressing the variables using logarithmic transformations of one or both of the variables and plotting them on axes that have arithmetic scales. Working with the reexpressed values makes graph paper with special scales unnecessary and facilitates using computer-constructed scatterplots.

Logarithmic reexpressions are not only used to linearize a nonlinear relationship. Another frequent use of them is to stabilize (make constant) the error standard deviation. ANOVA and regression models sometimes have error components whose variability increases with the size of the response. Plots of standard deviations versus average responses for ANOVA models having repeat observations can reveal such trends. Residual plots showing increased variability in the residuals as a function of the size of the predicted response is an indication of nonconstant error standard deviations for regression models.

Reexpressing one or more of the predictor variables is the preferred method (when possible) of linearizing a response function in a regression model. This approach leaves the original scale of the response variable and the error structure of the model intact. When linearizing $f(x_{ij}, \beta_k, e_i)$ involves reexpressing the response y, a word of caution is in order.

If the true form of the relationship between y and $x_1, x_2, \ldots, x_p$ has an additive error component, then any reexpression of y that linearizes the deterministic portion of the model will result in a transformed model that may not have an additive error structure. For example, if the relationship between a response and a predictor variable is

$$y = \beta_0 x^{\beta_1} + e,$$

a logarithmic transformation will result in a nonlinear relationship between y and both x and the error component e. This violates the linearity assumption of a linear regression model.

In practice, the form of the model error structure is usually not known. Thus, when the response is reexpressed and a linear regression model is used to fit the data, it is incumbent on the investigator to check carefully that the assumptions made on the model errors are reasonable. The residual-based techniques discussed earlier in this chapter can be used for this purpose.

18.3.2 Power Reexpressions

Logarithmic transformations are special cases of more general power families of transformations. One of the simplest power families for positive response values can be expressed as

$$z = \begin{cases} y^{\lambda}, & \lambda \neq 0, \\ \ln y, & \lambda = 0. \end{cases} \tag{18.11}$$

This power family includes the reciprocal ($\lambda = -1$), square-root ($\lambda = \frac{1}{2}$), and logarithmic ($\lambda = 0$) reexpressions as special cases. It can be made a continuous function of the parameter λ [e.g., (18.12)]. As noted in the previous section, logarithms to any base are constant multiples of the natural logarithms used in Equation (18.11), so any convenient logarithm can be used in place of the natural logarithm.

For estimation purposes, it is preferable to use a power family that is expressed on a common scale for all values of the parameter λ. The following power family, often referred to as the *Box–Cox* family of transformations, is generally used when estimation of the parameter λ is to be included in the modeling procedure:

$$z = \begin{cases} (y^{\lambda} - 1)/(\lambda h^{\lambda - 1}), & \lambda \neq 0, \\ (\ln y)h, & \lambda = 0, \end{cases} \tag{18.12}$$

where h is the geometric mean of the responses,

$$h = (y_1 y_2 \cdots y_n)^{1/n}.$$

Once λ is estimated using Equation (18.12), the responses are reexpressed using Equation (18.11) and the estimated value for λ. The regression coefficients are then estimated using ordinary least squares with reexpressed responses (see Exhibit 18.9).

EXHIBIT 18.9 REGRESSION ESTIMATION WITH POWER-FAMILY REEXPRESSIONS

1. Initially choose several values of λ in the interval $[-2, 2]$.
2. For each value of λ reexpress the response values using Equation (18.12).
3. Fit the reexpressed responses z_i to a linear regression model, and plot the residual sum of squares SS_E from each of the fits versus λ. Denote the value of λ that yields the minimum value of SS_E by λ^*. (If the interval $[-2, 2]$ or the increment of λ used to determine the SS_E values is too imprecise, refine the interval or increment values as needed.)
4. Determine a $100(1 - \alpha)\%$ approximate confidence interval for λ by drawing a horizontal line on the plot of SS_E versus λ from the vertical axis at

$$SS_E(\lambda^*)\left(1 + \frac{t^2_{\alpha/2}}{v}\right),$$

   where $t_{\alpha/2}$ is a value from the Student-t table corresponding to v degrees of freedom and v equals the number of the degrees of freedom for SS_E. The values of λ where this line intersects the graph are the approximate $100(1 - \alpha)\%$ lower and upper confidence limits for λ.
5. Using λ^* or a nearby rounded value, fit the regression model using the reexpressed responses (18.11) to obtain final estimates of the regression coefficients.

A modification of the transformation (18.12) changes the origin of the reexpressed variable. If y and y_i are replaced by $y - \theta$ and $y_i - \theta$, the origin of the transformation is shifted to θ. Through this two-parameter reexpression, responses that have negative values can be respecified. A change of origin is also useful when the original responses are all several orders of magnitude greater than zero.

The estimation of λ is accomplished by fitting linear regression models to transformed responses (18.12) for several values of λ. The residual sums of squares, SS_E, are then plotted versus the corresponding values of λ. The minimum value of SS_E determines the value of the transformation parameter λ. Usually the value of λ that minimizes SS_E is rounded to a nearby convenient

decimal, fraction, or integer in order to facilitate a reasonable interpretation of the transformation.

When plotting SS_E versus λ, scaling may be needed because SS_E can change by several orders of magnitude as λ is varied. When such large changes occur, either a replotting of only those SS_E values near the minimum or a scaling of SS_E can be used. A convenient scaling is to plot $L = -(n/2)\ln(SS_E/n)$. The values of L are *likelihood* values corresponding to λ; the value of λ that minimizes SS_E also maximizes L.

Data collected on a synthetic-rubber process were introduced in Table 15.8. A fit of a quadratic polynomial model to the data, summarized in Tables 15.9 and 15.11, appears to adequately describe the curvilinear trend between the solvent weights and the production rates shown in Figure 15.5. However, a plot of the studentized deleted residuals in Figure 18.11 from the quadratic fit displays a wedge shape. This pattern is suggestive of a nonconstant standard deviation. The power transformation (18.11) is now investigated in an effort to find a transformation that will both linearize the relationship between the solvent weights and the production rates and also stabilize the standard deviations.

The quadratic response function is only one of many that could possibly characterize the curvilinear relationship between the response and the predictor variable. A range of transformations using Equation (18.12) and values of λ from -2 to 2 in increments of 0.25 were constructed. After plotting the SS_E values, a refined interval of values from -0.5 to 1.0 in increments of 0.1 was used. Figure 18.12 shows a plot of the SS_E values versus λ over this latter range with the 95% confidence interval noted.

Because $\lambda = 1$ is well outside the confidence interval shown in Figure 18.12, the need for reexpression of the response variable is confirmed. Even though

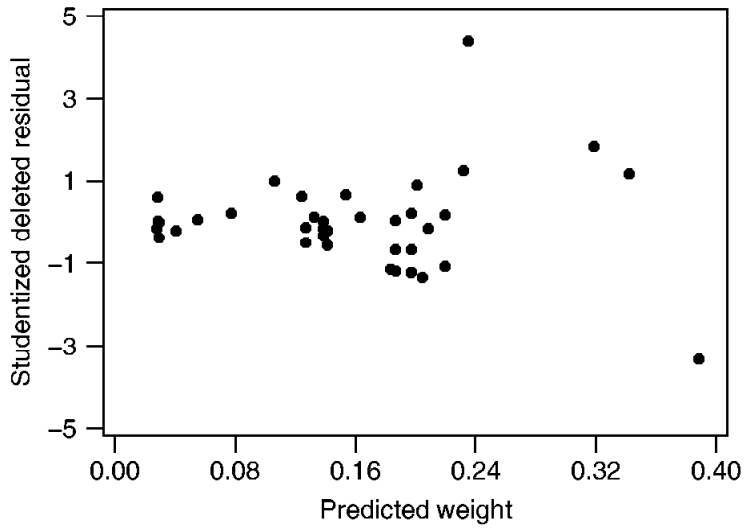


Figure 18.11 Residual plot for a quadratic fit to the solvent weight data.

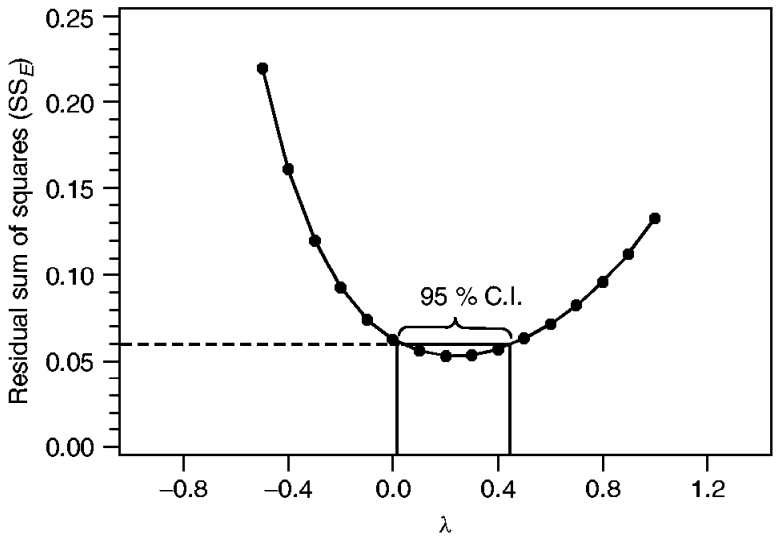


Figure 18.12 Estimation of λ for solvent-weight data. $\lambda^* = 0.2$; 95% confidence interval: $0.04 \leqslant \lambda \leqslant 0.36$.

the logarithmic transformation ($\lambda = 0$) is slightly outside the approximate confidence interval, the experimenter opted to use the logarithmic reexpression for ease of interpretation and justification.

Figure 18.13 displays a scatterplot of the reexpressed solvent weights using the logarithmic transformation $z = \ln y$. No substantive departure from a straight line is apparent. The regression fit is summarized in Table 18.7 along with a lack-of-fit test (Section 15.2). The lack-of-fit test does not indicate significant misspecification. Apart from the presence of an outlier, the plot of the studentized deleted residuals for the reexpressed model in Figure 18.14 shows

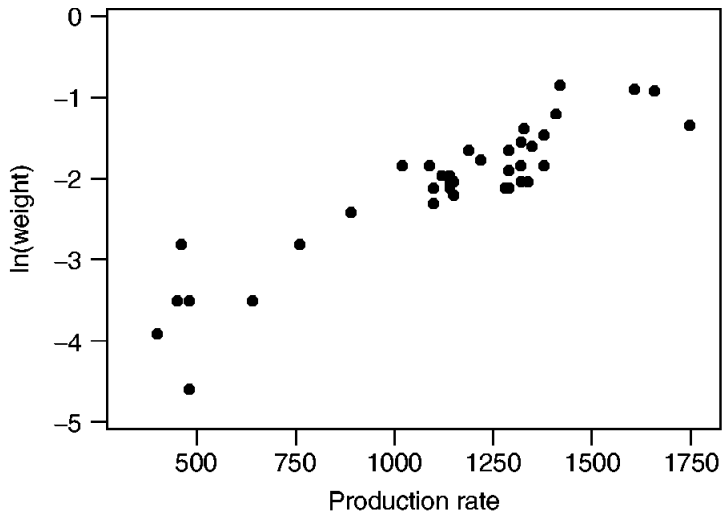


Figure 18.13 Scatterplot of reexpressed solvent weights.

TABLE 18.7 Summary of Linear Fit of ln (Weight) for Solvent Weight Data

Predictor Variable	Estimated Coefficient
Constant	−4.5841
Rate	0.0022

ANOVA

Source	df	Sum of Squares	Mean Squares	F	p-Value
Regression	1	20.61	20.61	173.19	0.000
Error	35	4.17	0.12		
Lack of Fit	25	3.24	0.13	1.40	0.297
Pure	10	0.93	0.09		
Total	36				

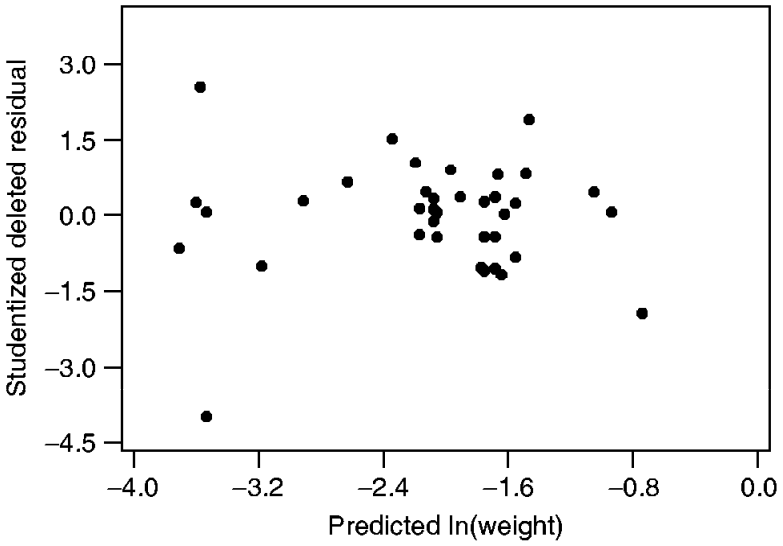


Figure 18.14 Residual plot for reexpressed solvent weights.

no anomalies; in particular, the wedge shape in Figure 18.11 has been eliminated. Thus, in this example the reexpression of the solvent weights using the power transformation (18.11) results in a parsimonious model fit that satisfies the regression-model assumptions better than the quadratic model fit of the original responses.

Power reexpressions can be extended to include applying the same reexpression to both the response variable y and the response function $f(x_j, \beta_k)$.

The following model transformations result:

$$z = \begin{cases} y^{\lambda} = [f(x_j,\ \beta_k)]^{\lambda} + e, & \lambda \neq 0, \\ \ln y = \ln[f(x_j,\ \beta_k)] + e, & \lambda = 0. \end{cases} \tag{18.13}$$

This extension is referred to as *reexpressing both sides* and can be used to preserve linearity or a meaningful physical model.

Suppose that the response function $f(x_j,\ \beta_k)$ has been specified, based on the nature of the process or phenomenon being studied, and that the model parameters have meaningful interpretations to the investigator. When the original model is respecified using Equation (18.13) to remedy a violation of the error-distribution assumptions, the form of the resulting prediction equation is

$$\hat{z} = [f(x_j,\ b_k)]^{l},$$

where l is the estimated value of λ and $l \neq 0$. Because this transformation is a unique function, the inverse reexpression can be applied to both sides of the prediction equation. In doing so, the original response metric is retained, and the original form and parametrization of the response function is preserved. Note that if $f(x_j,\ \beta_k)$ is linear in the unknown coefficients, then linearity is also preserved. Similar results are true for the logarithmic transformation using $l = 0$.

Estimation of λ when reexpressing both sides of the model requires modification similar to the use of Equation (18.12) instead of Equation (18.11). When reexpressing both sides, the following transformed response is fitted:

$$z = \begin{cases} \dfrac{[f(x_j,\ \beta_k)]^{\lambda} - 1}{\lambda h^{\lambda - 1}} + e, & \lambda \neq 0, \\ [\ln f(x_j,\ \beta_k)]h + e, & \lambda = 0. \end{cases} \tag{18.14}$$

APPENDIX: CALCULATION OF LEVERAGE VALUES AND OUTLIER DIAGNOSTICS

Using the matrix form of the multiple linear regression model given in the appendix to Chapter 15, the least-squares estimator of the regression coefficients can be written as

$$\mathbf{b}_c = (X_c' X_c)^{-1} X_c' \mathbf{y}, \tag{18.A.1}$$

where the subscript c is again used to denote the presence of a constant term in the model; that is, the first column of X_c is a column of ones. The predicted

values of the elements in the response vector $\mathbf{y}$ are given by

$$\hat{\mathbf{y}} = X_c\mathbf{b}_c = H\mathbf{y}, \tag{18.A.2}$$

where $H = X_c(X_c'X_c)^{-1}X_c'$. The matrix H is commonly called the *hat matrix* (notice that this is the matrix that when postmultiplied by $\mathbf{y}$ yields $\hat{\mathbf{y}}$). The leverage values, h_{ii} (or h_i), introduced in Section 18.1, are the diagonal elements of H.

The vector of raw residual values can be obtained by subtracting the predicted responses from the observed responses:

$$\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = (I - H)\mathbf{y}. \tag{18.A.3}$$

From equation (18.A.3) one can show that the variances and covariances of the elements of $\hat{\mathbf{y}}$ and $\mathbf{r}$ are contained on the diagonals and off-diagonals, respectively, of the following matrices:

$$\text{var}(\hat{\mathbf{y}}) = H\sigma^2 \tag{18.A.4}$$

and

$$\text{var}(\mathbf{r}) = (I - H)\sigma^2. \tag{18.A.5}$$

Studentized residuals t_i are the ratios of the raw residuals r_i and an estimate of the standard error of r_i. From (18.A.5), an estimated standard error of r_i is

$$\widehat{\text{SE}}(r_i) = (1 - h_i)^{1/2}s_e,$$

where $s_e = (\text{MS}_E)^{1/2}$ and MS_E is the mean squared error from the least-squares fit to the complete data set. Thus,

$$t_i = \frac{r_i}{(1 - h_i)^{1/2}s_e}. \tag{18.A.6}$$

Using matrix equations similar to Equation (18.A.1) to (18.A.5), one can derive expressions for least-squares estimators and prediction equations for a fit in which the ith observation is deleted. Straightforward matrix-algebra derivations leads to the expression for a deleted residual:

$$r_{(-i)} = \frac{r_i}{1 - h_i}. \tag{18.A.7}$$

From Equation (18.A.7) and the ith diagonal element of Equation (18.A.5), it follows that the standard error of a deleted residual is

$$\widehat{\text{SE}}(r_{(-i)}) = \frac{\sigma}{(1 - h_i)^{1/2}}. \tag{18.A.8}$$

Again using matrix algebra and the matrix expressions for residuals from the complete data set and the data set with the ith observation deleted, one can show that an estimator of the error standard deviation that is independent of the predicted responses and the residuals from both the complete and the reduced data sets is

$$s_{(-i)} = \left(\frac{\mathrm{SS}_E - r_i^2/(1 - h_i)}{\nu - 1} \right)^{1/2}, \tag{18.A.9}$$

where SS_E is the error sum of squares from the complete data set and is based on ν degrees of freedom.

REFERENCES

Text References

The references listed at the end of Chapters 4, 6, and 14 provide informative coverage on various aspects of statistical model assumptions and their assessment. Additional references include:

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*, New York: John Wiley & Sons, Inc.

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*, New York: John Wiley & Sons, Inc.

Freund, R. J. and Wilson, W. J. (1998). *Regression Analysis: Statistical Modeling of a Response Variable*, San Diego, Calif.: Academic Press.

Sen, A. and Srivastava, M. (1990). *Regression Analysis: Theory, Methods, and Applications*, New York: Springer-Verlag, Inc.

The following references provide additional information on univariate outlier-detection techniques. The second reference also provides a history of the study of outliers.

"Standard Practice for Dealing with Outlying Observations," in *Annual Book of ASTM Standards*, Designation E 178-80, Philadelphia: American Society for Testing Materials.

Beckman, R. J. and Cook, R. D. (1983). "Outlier .....s," *Technometrics*, **25**, 119–149.

Tietjen, G. L. and Moore, R. H. (1972). "An Extension of Some Grubbs-Type Statistics for the Detection of Several Outliers," *Technometrics*, **14**, 583–598.

The following texts provide specific coverage of the diagnostics for influential observations that were discussed in Section 18.1. All these texts are at an advanced mathematical level.

Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*, New York: John Wiley & Sons, Inc.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics*, New York: John Wiley and Sons, Inc. *This text provides extensive coverage of outlier diagnostics such as DFFITS and DFBETAS.*

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, New York: Chapman and Hall. *Excellent coverage of statistical basis for the identification of outliers. Extensive treatment of alternative outlier diagnostic procedures.*

Weatherill, G. B. (1986). *Regression Analysis with Applications*. New York: Chapman and Hall.

The following books are among many that comprehensively cover multivariate statistical procedures. These books can be consulted for inferential procedures when model errors are correlated.

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*, New York: John Wiley & Sons, Inc.

Johnson, R. A. and Wichern, D. W. (1982). *Applied Multivariate Statistical Analysis*, Englewood Cliffs, NJ: Prentice-Hall, Inc.

Morrison, D. F. (1976). *Multivariate Statistical Methods*, New York: McGraw-Hill Book Company.

Some useful articles on procedures for detecting model-assumption violations are given below. The first concerns measurement error in experimental design. The second discusses assumption violations in ANOVA models. The last two present various test procedures for normality.

Box, G. E. P. (1963). "The Effects of Errors in the Factor Levels and Experimental Design," *Technometrics*, **5**, 247–262.

Cochran, W. G. (1947). "Some Consequences When the Assumptions for the Analysis of Variance are not Satisfied," *Biometrics*, **3**, 22–38.

Shapiro, S. S. (1980). *How to Test Normality and Other Distributional Assumptions*, Volume 3, Milwaukee, WI: ASQC Basic References in Quality Control.

Shapiro, S. S. and Wilk, M. B. (1965). "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, **52**, 591–611.

Most references on nonlinear model fitting and inference are at an advanced mathematical level. This includes the following references:

Gallant, A. R. (1987). *Nonlinear Statistical Models*, New York: John Wiley & Sons, Inc.

Milliken, G. A. (1982). *Nonlinear Statistical Models*, Arlington, Virginia: The Institute for Professional Education.

Ratkowsky, D. A. (1983). *Nonlinear Regression Modeling*, New York: Marcel Dekker, Inc.

Additional information on reexpression is contained in the following articles:

Patterson, R. L. (1966). "Difficulties Involved in the Estimation of a Population Mean Using Transformed Sample Data," *Technometrics*, **8**, 535–537.

Snee, R. D. (1986). "An Alternative Approach to Fitting Models When Re-expression of the Response is Useful," *Journal of Quality Technology*, **18**, 211–225.

EXERCISES

1 A study of engine-induced structural-borne noise in a single-engine light aircraft was carried out to determine the level to which structural-borne noise influences interior levels of noise. Cabin noise was recorded for 20 ground tests for engine-attached aircraft configurations. Construct a quantile–quantile plot for this data set. Check suspect observations using Grubbs test. Do the data appear to contain any outliers?

Cabin Noise (decibels)

186	263	201	211	199	198	189	257	204	203
185	209	225	193	166	175	210	205	182	223

2 A study was conducted to identify factors that may contribute to tooth discoloration in humans. Twenty subjects were randomly chosen from different communities to participate in the year-long study. Dental examinations were given to each subject at the beginning of the study and again one year later. Percentage increases in tooth discoloration were measured. Data gathered from each individual included:

(a) Age (years).

(b) Fluoride level (ppm) of public water supply in the subject's community.

(c) Average number of tooth brushings per week.

The standardized regression equation for this data set is

$$\text{tooth discoloration} = 20.8 + 6.944 \times \text{fluoride} - 2.674 \times \text{age} - 4.297 \times \text{brush}.$$

Tooth Discoloration (%)	Fluoride Level (ppm)	Age (years)	Brushings per Week
12	0.7	46	5
10	1.3	44	8
14	1.5	34	7
20	2.4	72	6
18	2.6	22	9
18	2.9	40	11
25	3.0	52	5
21	3.6	37	10
36	3.8	11	7
44	4.0	10	6

Tooth Discoloration (%)	Fluoride Level (ppm)	Age (years)	Brushings per Week
11	0.5	24	8
15	1.8	39	7
22	2.1	42	6
21	0.4	6	5
27	3.4	68	5
25	2.9	59	7
18	2.1	18	10
21	3.4	32	12
18	3.0	15	10
20	1.7	59	9

Calculate the residuals and produce a scatterplot of residuals versus predicted values. Also plot the residuals versus each of the predictor variables. Comment on any interesting features of the plots.

3 Calculate the studentized deleted residuals from the fit to the data in Exercise 2. Calculate leverage, DFFITS, and DFBETAS values for each observation in the data set. From the plots and the statistics calculated in this exercise and the previous one, what conclusions can be drawn about the presence of outliers?

4 Refer to the videocassette sales data in Exercise 1, Chapter 14. Perform a comprehensive analysis of the possible presence of outliers in the raw data. Construct appropriate plots, and calculate the outlier diagnostics discussed in this chapter. What conclusions can be drawn about the presence of outliers?

5 Calculate outlier diagnostics for the synthetic-rubber process data in Table 15.7 for two models, one having only the linear term in the production rates and the other having both the linear and the quadratic terms. Which observations appear to be outliers? Redo this analysis using the natural logarithm of solvent weight as the response variable. How do the outlier diagnostics compare? Construct plots of the residuals to interpret any differences noted. Which model fit do you prefer? Why?

6 An experiment was conducted to study the effects of four factors on an engine-knock measurement.

(a) Investigate the observations on knock number using the techniques described in Section 18.1. Are any outliers identified in this analysis?

(b) Fit a regression model, linear in the four predictors, to the knock-number measurements. Are any observations identified as influential? If these results differ from those of (a), determine the reasons for the differences.

Spark Timing	Air–Fuel Ratio	Intake Temperature	Exhaust Temperature	Knock Number
13.3	13.9	31	697	84.4
13.3	14.1	30	697	84.1
13.4	15.2	32	700	88.4
12.7	13.8	31	669	84.2
14.4	13.6	31	631	89.8
14.4	13.8	30	638	84.0
14.5	13.9	32	643	83.7
14.2	13.7	31	629	84.1
12.2	14.8	36	724	90.5
12.2	15.3	35	739	90.1
12.2	14.9	36	722	89.4
12.0	15.2	37	743	90.2
12.9	15.4	36	723	93.8
12.7	16.1	35	649	93.0
12.9	15.1	36	721	93.3
12.7	15.9	37	696	93.1

7 Refer to the experiment described in Exercise 5, Chapter 5, regarding the study of glass-mold temperatures. Calculate the residuals from an analysis of variance using molten-glass temperature and glass type as factors. Include an interaction term in the analysis. Assess the assumption of normality using (a) graphical techniques and (b) the Shapiro–Wilk test. On the basis of these analyses, is the assumption of normally distributed errors reasonable?

8 Refer to Exercise 2 in Chapter 5, which describes an experiment in which temperatures and engine speeds are varied in order to study the effect they may have on engine stability. The data collected for this experiment are given in the accompanying table.

Engine Speed (rpm)	Temperature (°F)	Time (sec)
1000	70	34
1200	40	36
1200	70	20
1200	20	43
1200	10	44
1400	20	24
1600	10	19
1400	20	23

Engine Speed (rpm)	Temperature (°F)	Time (sec)
1600	40	15
1600	70	11
1000	40	37
1400	70	14
1200	70	18
1400	10	27
1600	20	18
1400	40	18
1000	20	37
1000	40	42
1000	20	39
1400	70	12
1600	70	10
1400	40	17
1200	20	40
1000	10	39
1200	40	35
1000	70	30
1000	40	32
1600	20	12
1600	40	10
1600	10	14
1200	10	40
1400	10	23

Calculate the residuals from a regression analysis in which engine speed and temperature are the only predictors. Plot the residuals against the predicted values. Is curvature evident in the plot? If so, reexpress the model by either adding additional model terms or transforming variables to obtain a fit in which the residual plot does not display curvature.

9 An investigation was carried out to study the hazards posed by chemical vapors released during cargo-tanker operations. Specifically, the vapor concentration at breathing height downwind from vapor vents on the ship's deck was measured. The data in the accompanying table represent vapor concentrations collected while the ship was carrying benzene in the cargo tank. Fit a regression model to the data using distance as the predictor variable. Construct a partial-residual plot. Interpret the results of this analysis. Reexpress the model, if appropriate, to rectify any difficulties noted in the plot.

In [Vapor Concentration(kg/m^3)]	Downwind Distance (m)
−2.0	2.0
−2.3	4.0
−1.9	2.5
−2.2	7.5
−2.3	6.0
−1.0	0.1
−3.0	5.5
−2.0	3.5
−2.3	3.0
−1.1	0.5
−2.3	9.0
−2.4	6.5
−2.1	4.5
−2.1	1.0
−2.4	5.0
−3.0	8.0
−2.0	1.5
−2.2	10.5
−2.9	7.0

10 The data in Exercise 3 were collected sequentially in time in the order they are listed. Plot the residuals versus the time sequence (observation number) after correcting for any curvature that was observed. From the plot, what conclusion can be drawn about the independence assumption of the model errors?

11 Assess the normality assumption for the VCR sales data in Exercise 1 of Chapter 14. If the normality assumption does not appear to be reasonable, identify a suitable reexpression of the model that does permit an assumption of normally distributed errors.

12 Assess the normality assumption for the nitrogen oxide data in Exercise 2 of Chapter 15.

13 Investigate the effects of outliers on graphical and numerical evaluations of normality by assessing the assumption of normality for the cabin-noise data of Exercise 1 (a) using the complete set, and (b) excluding the two largest observations. Is your conclusion about the reasonableness of an assumption of normally distributed errors affected by the removal of the two observations? Explain your conclusion.

14 Calculate and plot appropriate outlier diagnostics for the height data in Exercise 3 of Chapter 15. Use the plots to draw conclusions about the possible presence of influential data.

15 An experimental study was made to determine the frequency responses to longitudinal excitations of thin-walled hemispherical–cylindrical tanks containing liquid. The experimenter would like to express natural frequency in terms of the liquid height ratio in the tanks. The data in the accompanying table were collected from twelve experiments. Construct a scatterplot of these data using the liquid height ratio as the predictor variable. Reexpress the data, if necessary, to a linear form. Examine Box–Cox power transformations.

Liquid Height Ratio	Frequency (Hz)
0.2	2010
0.3	1876
0.4	1720
0.5	1650
0.6	1584
0.7	1532
0.8	1413
0.9	1482
1.0	1406
1.1	1385
1.2	1371
1.3	1350

16 The feasibility of using a portable instrument to determine the physical condition of fabric-reinforced polyurethane fuel tanks was studied. This instrument measured the surface resistivity of seven polyurethane samples in varying degrees of degradation due to humid aging. It is known that the data plot as a straight line on semilog paper. Confirm this fact by plotting the data. Replot the data on arithmetic paper by making the appropriate variable reexpression. Examine Box–Cox power transformations.

Exposure Time (days)	Surface Resistivity ($10^{10}\Omega/cm^2$)
0	4500
7	2050
15	3100
25	1000
42	225
57	32
65	21

17 Different fuels being used in certain military aircraft have caused problems in the engine's fuel pump by the presence of peroxides in the turbine fuel. To avoid these problems in the future, it would be advantageous to predict the potential peroxide content of a fuel before the fuel has any measurable peroxide content. A study was made of a particular fuel in which peroxides were measured as a function of the storage time of the fuel at 43°C. (See accompanying table.) Produce a scatterplot of the data and linearize them (if necessary) using appropriate reexpressions.

Storage Time (weeks)	Peroxides (ppm)	Storage Time (weeks)	Peroxides (ppm)
1	5	12	372
2	26	16	612
5	92	20	1105
8	151	24	1500
10	197	4	75
7	172	18	801
14	625	22	1401

18 Determine which of the nonlinear relationships below can be reexpressed as linear models. Show the reexpressions, where applicable.

(a) $y = ae^{bx}$

(b) $y = \dfrac{1}{a + bx}$

(c) $y = \dfrac{1}{1 + be^{x}}$

(d) $y = ax_1^b x_2^c$

(e) $y = ab^{x}$

(f) $y = a - be^{x}$

(g) $y = \dfrac{1}{1 + e^{a+bx}}$

(h) $y = a + bx_1 + c^{x_2}$

(i) $y = (a + b \sin x)^{-1}$

19 A heavily insulated and cooled diesel engine was used to measure the starting temperature of 15 test fuels. Multiple regression analysis will be applied to relate the minimum starting temperature of the diesel engine to several fuel properties: auto-ignition temperature, cloud point, flash point, and viscosity. Investigate the need for any reexpression among the data in the accompanying table.

Minimum Starting Temperature (°C)	Auto-Ignition Temperature (°C)	Cloud Point (°C)	Flash Point (°C)	Viscosity
−4.0	180	−6	77	1.40
−1.0	185	−46	64	1.50
−8.6	180	−21	83	1.95
2.8	190	−65	36	1.07
−9.4	200	−38	55	1.56
1.0	180	−68	−24	0.78
−4.2	191	−54	−21	1.12
−7.6	185	−67	43	1.39
−6.0	185	−48	64	2.07
−9.8	179	−16	84	2.57
9.0	202	−64	−21	0.76
12.0	210	3	32	0.23
−4.0	190	−60	−22	0.82
4.0	190	−59	−2	0.78
3.5	204	0	37	0.20

20 Examine the chemical-vapor data in Exercise 9. Use the techniques described in this chapter, as well as alternatives such as adding additional powers of the predictor variable, to obtain a satisfactory fit to the data. Seek a fitted model that is parsimonious (i.e., has a simple form) and for which the usual error assumptions appear reasonable.

21 Investigate whether second-order terms in the predictor variables for the engine-knock data in Exercise 6 improve the fit to the knock measurements. Assess the reasonableness of the usual error assumptions, perhaps after deleting influential observations.

CHAPTER 19

Variable Selection Techniques

Regression analyses often are conducted with a large set of candidate predictor variables, only a subset of which are useful for predicting the response. In this chapter techniques for identifying important predictor variables are discussed. Specific topics to be covered include:

- *criteria for the selection of important predictor variables,*
- *subset selection methods, including the assessment of all possible regressions and stepwise selection procedures, and*
- *collinearity effects on selection methods.*

One of the last components of a comprehensive regression analysis PISEAS, Section 14.2 is the selection of the most suitable predictor variables. This step in a comprehensive analysis usually is performed last because variable selection methods can be affected by model specification, influential observations, and collinearities. Variable selection is a concern in any regression analyses for which no complete theoretical model is available that stipulates the relationship between the response and the predictor variables.

Two competing goals face an investigator in the fitting of regression models that do not have theoretical bases for the specification of the model. On one hand is the desire for an adequate fit to explain the relationship between the response and the predictors. This usually is the primary concern, and it often mandates that many candidate predictors be included in an initial fit to the data.

The second concern is simplicity. One seeks the simplest expression of the relationship between the response and the predictors. Elimination of unnecessary predictor variables and a parsimonious expression of the relationship are sought. This goal must, however, be secondary to the first one cited above. For

Statistical Design and Analysis of Experiments, Edited by Robert L. Mason, Richard F. Gunst, and James L. Hess. ISBN 0-471-37216-1 © 2003 Wiley-Interscience.

this reason, we have delayed consideration of variable selection until discussions of model specification and the accommodation of influential observations have been completed.

In this chapter we consider variable selection techniques. In the next section several criteria for identifying important subsets of variables are introduced. In Sections 19.2 and 19.3, methods for selecting these subsets using the criteria of Section 19.1 are discussed. The final section of this chapter illustrates the effects that collinearities can have on least-squares estimates and on variable selection procedures.

19.1 COMPARING FITTED MODELS

In the general discussions of variable selection criteria presented in this section, two models are compared: the complete model and a model with a reduced number of predictor variables. For convenience we write these two models as follows:

$$M_1: \quad y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i \tag{19.1}$$

and

$$M_2: \quad y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i^*. \tag{19.2}$$

The first model, M_1, contains the complete set of p predictor variables and is assumed to adequately relate the response variable to the predictor variables, although there may be some predictors that are unnecessary. Specifically, the model is assumed to be correctly specified, but there might be some predictor variables included in the model for which $\beta_j = 0$. The intercept term can be excluded from the models (19.1) and (19.2) with obvious changes in the discussions which follow; however, if the reduced model (19.2) contains an intercept term, the complete model (19.1) must also contain one.

The second model, M_2, consists of only the first $k < p$ predictor variables. The first k of the p variables are chosen, rather than an arbitrary subset of k, only for ease of presentation. The variable selection methods presented in subsequent sections do not require that the first k variables be the subset of interest. Note that on reducing the model the error term may change to reflect the exclusion of important predictor variables. If important predictor variables, those for which $\beta_j \neq 0$, are erroneously deleted from the model, their effects on the response are included in the model error terms. Thus, we change the notation on the error term from e_i to e_i^* in the reduced model to reflect the bias which may occur. Likewise, coefficient estimates may change dramatically when important predictors are eliminated, reflecting biases incurred by eliminating these variables.

Many criteria have been proposed for assessing whether a reduced model is an adequate representation of the relationship between the response and predictor variables. Most of these criteria involve a comparison of the complete and the reduced models, because it is assumed that the complete model does adequately relate the response and predictors. The F-statistic in (15.18) provides a statistical test of the significance of the deleted parameters:

$$F = \frac{\text{MSR}(M_1|M_2)}{\text{MSE}_1}. \tag{19.3}$$

This statistic is based on the principle of reduction in error sums of squares (Section 8.1) and can be calculated from the error sums of squares for the two models:

$$\text{MSR}(M_1|M_2) = \frac{\text{SSE}_2 - \text{SSE}_1}{p-k} \quad \text{and} \quad \text{MSE}_1 = \frac{\text{SSE}_1}{n-p-1}.$$

This statistic is used to test the hypotheses

$$H_0: \beta_{k+1} = \cdots = \beta_p = 0$$

vs

$$H_a: \text{at least one } \beta_j \neq 0, \quad j = k+1, \ldots, p.$$

Under the null hypothesis the F-statistic (19.3) follows an F probability distribution with $\nu_1 = p - k$ and $\nu_2 = n - p - 1$ degrees of freedom.

An often used criterion for assessing the predictive ability of a reduced model relative to the complete model is a comparison of the coefficients of determination, R^2. Because the reduced models do not contain the same number of predictor variables as the complete model, adjusted coefficients of determination, equation (15.11), should be compared:

$$R_a^2 = 1 - \frac{n-1}{n-k-1}\frac{\text{SSE}_2}{\text{TSS}}. \tag{19.4}$$

For the complete model, k is replaced by p and SSE_2 by SSE_1. If the adjusted coefficients of determination are approximately equal for the full and a reduced model, the two fits are regarded as having equal predictive ability.

A third criterion for assessing reduced models relative to the complete model is the following statistic:

$$C_m = \frac{\text{SSE}_2}{\text{MSE}_1} - (n - 2m), \tag{19.5}$$

where $m = k + 1$. An adequate reduced model is one for which C_m is approximately equal to m. When predictors for which $\beta_j = 0$ are correctly eliminated from the model, SSE_2 estimates $(n - k - 1)\sigma^2$ and MSE_1 estimates σ^2. Thus, C_m is approximately equal to m.

C_m can be much greater than m when important predictor variables are deleted from the model. In such circumstances, SSE_2 contains the bias due to the incorrect deletion of the predictors and it overestimates $(n - k - 1)\sigma^2$, while MSE_1 still estimates σ^2. When unimportant predictors remain in a reduced model, C_m can be much less than m. In such cases there are usually subsets with fewer variables for which C_m is approximately equal to m.

Use of the statistic (19.5) is generally accompanied by a plot of the C_m values versus m for the better subsets, that is, those with small values of m and for which $C_m \leqslant m$. Such plots allow a visual comparison of alternative subsets.

There are many other alternative criteria for the assessment of reduced models. The above three criteria are presented because they are among the most widely used and they illustrate the important features of subset selection criteria. In applications they are combined with a methodology for selecting candidate subsets. In the next two sections we present some of the more popular subset selection methods.

It is important to note that the material in this section and the next two are based solely on statistical criteria. As with all other statistical procedures, they are not intended to replace expert judgment or subject-matter theoretical considerations. In many applications variable selection procedures are not even appropriate, since theoretical considerations define both the important predictors and the functional relationship between the response and the predictors.

19.2 ALL-POSSIBLE-SUBSET COMPARISONS

The most comprehensive assessment of candidate reduced models is a comparison of the selection criteria (19.3) to (19.5) for all possible subsets of predictor variables. By comparing all possible subsets, an investigator can not only determine the "best" reduced model(s) according to the above criteria, but also identify alternatives to the "best" ones.

The term "best" is placed in quotes because of the frequent misconception that there is a single reduced model that outperforms all others. When using the above selection criteria, one should realize that they are all random variables and therefore subject to variability. The reduced model that has the highest R_a^2, for example, is not necessarily the best one if there are other reduced models that have R_a^2 close to the maximum. The selection criteria should be used as guides to select all of the better subsets, and whenever possible, engineering or

scientific judgment should be used to select from among the better candidate reduced models.

To illustrate these points, we return to the chemical-yield data introduced in Table 18.1. In Chapter 18 this data set was used to illustrate the detection of influential observations. Prior to the calculation of subset selection criteria, we examined diagnostics for influential observations and, after several fits to the model with various observations deleted, decided to eliminate five observations from the data set: observations 6, 11, 16, 18, and 45.

Table 19.1 exhibits subset selection criteria for all possible reduced models using the data set in Table 18.1 with the five influential observations mentioned above eliminated. The five candidate predictor variables are catalyst (CAT), conversion (CON), flow of raw materials (FLOW), the inverse of a ratio of reactants (INV), and the interaction of conversion and flow (INT). The interaction was again calculated after CON, FLOW, and INV were standardized using normal deviate standardization (see Section 15.4) to reduce possible effects of collinearities (see Section 19.4).

We focus attention in Table 19.1 on the use of the adjusted coefficient of determination (19.4) and the C-statistic (19.5), because computer programs that perform all subset regressions generally print out these statistics but do not always print out the F-statistics, equation (19.3). Once screening for the better subsets is completed, the F-statistics should be evaluated for the selected subsets.

Screening for the better subsets on the basis of large R_a^2-values leads to several candidate reduced models. If one screens by using a cutoff of, say, $R_a^2 > 0.35$, then one subset with two predictors, four subsets with three predictors, and four subsets with four predictors will be selected. The choice of $R_a^2 > 0.35$ is arbitrary but is reasonably close to the maximum value of $R_a^2 = 0.433$. The important issue here is that several candidates have been identified, not simply the one with the highest value for the coefficient of determination. We also would not hesitate to include one or more of the reduced models that have slightly smaller values of R_a^2 if these models have physical or experimental meaning to the investigator.

A further delineation of the better reduced models occurs when the criterion $C_m \approx m$ is examined. A plot of C_m versus k, the number of predictor variables (excluding the constant term, which is included in all the models), is shown in Figure 19.1 for all reduced models having C_m less than 10. Code numbers are used in the figure to identify the subsets of predictors. Again, somewhat arbitrarily, we choose a cutoff of $C_m = m + 2$ for identification of the better subsets. With this criterion, two subsets with three predictors and three subsets with four predictors remain as candidates for reduced models.

If one had to select a single subset solely on the basis of a statistical analysis of this data set, one of the two subsets with $C_m < m + 2 = 6$ having three predictors would be selected. Each of these subsets has an acceptable

TABLE 19.1 Subset Selection Criteria for All Possible Reduced Models: Chemical-Yield Data

Subset	R_a^2	m	C_m
INT	0*	2	39.5
INV	0*	2	39.1
FLOW	0.021	2	35.9
CAT	0.031	2	35.1
CON	0.238	2	17.2
INV, INT	0*	3	41.0
FLOW, INV	0.009	3	37.4
CAT, INT	0.012	3	37.1
CAT, INV	0.012	3	37.1
FLOW, INT	0.014	3	36.9
CAT, FLOW	0.045	3	34.2
CON, INT	0.224	3	19.1
CON, FLOW	0.302	3	12.4
CAT, CON	0.347	3	8.6
CON, INV	0.386	3	5.3
CAT, INV, INT	0*	4	39.0
FLOW, INV, INT	0.002	4	38.3
CAT, FLOW, INT	0.031	4	35.8
CAT, FLOW, INV	0.060	4	33.4
CON, FLOW, INT	0.295	4	13.8
CAT, CON, INT	0.344	4	9.7
CON, FLOW, INV	0.374	4	7.2
CON, INV, INT	0.375	4	7.1
CAT, CON, FLOW	0.402	4	4.8
CAT, CON, INV	0.433	4	2.3
CAT, FLOW, INV, INT	0.045	5	35.0
CON, FLOW, INV, INT	0.363	5	9.0
CAT, CON, FLOW, INT	0.390	5	6.8
CAT, CON, INV, INT	0.421	5	4.3
CAT, CON, FLOW, INV	0.425	5	4.0
CAT, CON, FLOW, INV, INT	0.412	6	6.0

*Negative values set to zero.

R_a^2-value, their C_m-values are approximately equal to m, and each subset contains one fewer variable than the three better reduced models having four predictors. The choice between the two three-predictor reduced models is not unequivocal on the sole basis of statistical criteria. However, in the four-variable reduced model that has all four variables that are included in these

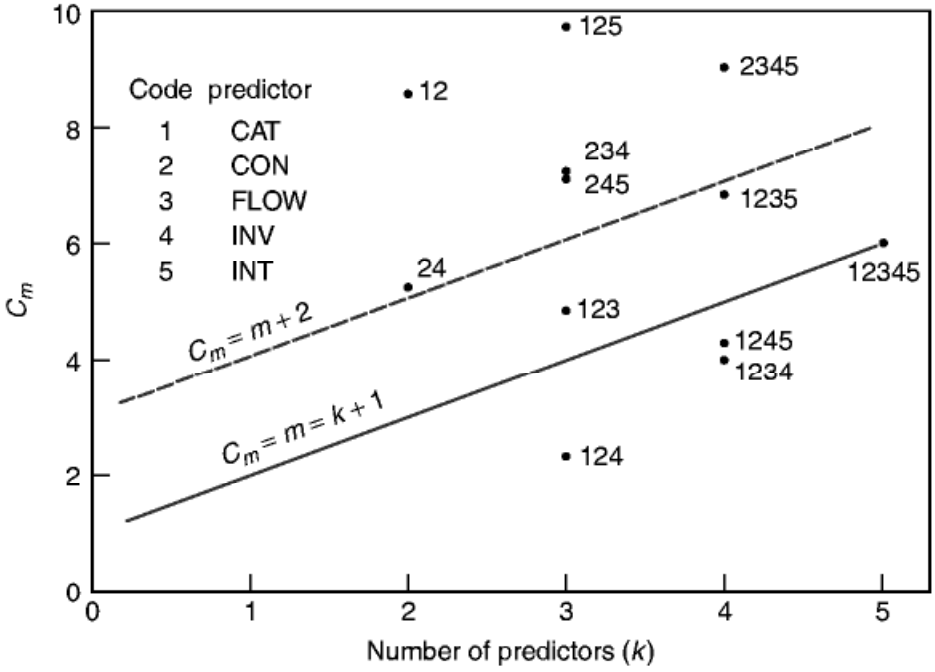


Figure 19.1 Plot of C_m statistics ($C_m \leqslant 10$), chemical-yield data.

two three-variable models (CAT, CON, FLOW, INV), FLOW is not statistically significantly ($p = 0.570$) at the recommended significance level of $\alpha = 0.25$ (see Sections 14.5 and 15.2). Based on this information, the subset CAT, CON, and INV would be selected.

The examination of all possible reduced models allows the greatest scrutiny of alternative candidates. Statistical criteria can be used to screen the better subset models and, with engineering and scientific judgment, to select one or more for use as final prediction equations. This approach to the identification of reduced models is recommended whenever the number of candidate predictor variables is not prohibitively large. When it is, the procedures discussed in the next section should be used.

19.3 STEPWISE SELECTION METHODS

Stepwise selection methods sequentially add or delete predictor variables to the prediction equation, generally one at a time. These methods involve many fewer model fits than all possible subsets, since each step in the procedure leads directly to the next one. Three popular subset selection methods will be discussed in this section: forward selection, backward elimination, and stepwise iteration.

The selection methods presented in this section are especially useful when a large number of candidate predictors are available for possible inclusion

in the final reduced model. The major disadvantage of these methods is that they generally isolate only one reduced model and do not identify alternative candidate subsets of the same size. These methods also do not necessarily identify the reduced model that is optimal according to the selection criteria discussed in Section 19.1.

Stepwise selection methods are implemented using F statistics as in Equation (19.3) (or the equivalent t statistics). A critical issue in using stepwise selection methods is the choice of the significance level to use in the tests of the significance of the candidate predictor variables. Because several tests are to be made on the same data set, care must be taken in selecting the Type-I error rate. This problem is similar to the one encountered in Chapter 6 when applying multiple comparison procedures. In stepwise selection, excellent fits might occur using a small significance level, such as $\alpha = 0.05$, and the current data set. However, there is no guarantee that the excellent fit will result in excellent predictions for future data. This can occur because of large Type-II error rates; that is, a high probability of not including important predictors. To protect against this outcome, a large significance level, such as $\alpha = 0.25$, is recommended when applying stepwise selection methods.

19.3.1 Forward Selection

The forward selection (FS) procedure for variable selection begins with no predictor variables in the model. Variables are added one at a time until a satisfactory fit is achieved or until all predictors have been added. The decision when to terminate the procedure can be made using any of the selection criteria discussed in Section 19.1; however, F-statistics generally are used. Because these F-statistics are based on the principle of reduction in error sums of squares, they measure the incremental contribution of a predictor variable above that provided by the variables already in the model. When the addition of a predictor does not result in a statistically significant F-statistic, the procedure is terminated (see Exhibit 19.1).

The F-statistic used in step 4 of the FS procedure tests the hypothesis

$$H_0: \beta_{r+1} = 0 \quad \text{vs} \quad H_a: \beta_{r+1} \neq 0$$

in the model

$$M_1: \qquad y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{r+1} x_{i,r+1} + e_i^*. \tag{19.6}$$

The procedure selects as the $(r+1)$st predictor variable the one in the remaining $p - r$ that has the largest F-value. If this predictor is not statistically significant, the procedure terminates and no further predictors are added to the model. If it is statistically significant, the procedure continues with the remaining $p - r - 1$ candidate predictors.

EXHIBIT 19.1 FORWARD SELECTION PROCEDURE

1. Fit p single-variable regression models, each having one of the predictor variables in it. Calculate the overall model F-statistic for each of the p fits. Select the fit with the largest F-statistic.
2. Using a significance level of $\alpha = 0.25$, test the significance of the predictor variable. If it is not significant, terminate the procedure and conclude that none of the predictors are useful in predicting the response. If it is significant, retain the predictor variable, set $r = 1$, and proceed to step 3.
3. Fit $p - r$ reduced models, each having the r predictor variables from the previous stage of the selection process and one of the remaining candidate predictors. Calculate the overall model F-statistic for each of the fits. Select the fit with the largest F-statistic.
4. Using a significance level of $\alpha = 0.25$, test the significance of the additional predictor variable using the F-statistic (19.3) with the model M_1 containing the $r + 1$ predictors and the reduced model M_2 containing the r predictors from the previous stage of the selection process. The degrees of freedom for this F-statistic are $\nu_1 = 1$ and $\nu_2 = n - r - 2$.
5. If the F-statistic is not significant, terminate the procedure and retain the r predictors from the previous stage. If it is significant, add the additional predictor to the r previously selected, increment r by 1, and return to step 3.

Table 19.2 shows the application of the forward selection procedure to the chemical-yield data of Table 18.1 with the five outliers mentioned in the last section deleted. At stage 1, conversion is added to the model. At stage 2, the inverse of the reactant ratio variable is added. At stage 3, catalyst is added. The procedure terminates at stage 4. In this instance, the method produced the optimal subset of size $r = 3$, the one ultimately selected in the examination of all possible reduced models in the last section. We again note that selection of the optimal subset cannot be guaranteed with this procedure; however, the optimal subset is often selected when predictor variables are not highly collinear.

When there are many candidate predictors, application of FS is frequently allowed to proceed two or three steps beyond the one that suggests termination. This is because the estimate of the uncontrolled error variance, MS_E, is heavily biased in the early steps of this procedure. The effects of any predictors that should be in the model but have not yet been included are contained in the error term e_i^*. When possible, the final reduced model should be compared with the complete model using the F-statistic (19.3) to ensure that the procedure did not terminate prematurely due to an inflated error component.

Interpretation of the results of FS should be done with some care. The order of entry of the predictor variables should not be interpreted as indicating an order of importance of the variables. Interrelationships among the predictors,

TABLE 19.2 Forward Selection Procedure for the Chemical-Yield Data

Stage	Reduced Model	F-Statistic for Added Predictor	p-Value
1	CAT	2.67	0.109
	CON*	17.19	0.000
	FLOW	2.13	0.150
	INV	0.23	0.633
	INT	0.01	0.915
2	CON, CAT	9.54	0.003
	CON, FLOW	5.72	0.021
	CON, INV*	13.33	0.001
	CON, INT	0.10	0.754
3	CON, INV, CAT*	5.10	0.028
	CON, INV, FLOW	0.01	0.915
	CON, INV, INT	0.14	0.712
4	CON, INV, CAT, FLOW	0.33	0.570
	CON, INV, CAT, INT	0.03	0.869

*Variable selected

synergistic effects of the variables, and the biases due to poor model specification in the early steps of the procedure can all affect the order of entry of the predictors.

19.3.2 Backward Elimination

The backward elimination (BE) procedure for variable selection begins with all the predictor variables in the model. Variables are deleted one at a time until an unsatisfactory fit is encountered (see Exhibit 19.2). As with forward selection, the decision of when to terminate the procedure is most frequently based on the F-statistic (19.3).

At each step of BE the statistical significance of the predictor variable having the smallest t or F statistic is examined. Only if the predictor is not statistically significant is it deleted from the model. Unlike forward selection, the error term in the denominator of the F-statistic (19.3) should not be heavily biased in backward elimination, because the analysis begins with what is assumed to be a correctly specified model, apart from possibly some extraneous predictor variables. At each step of the selection procedure only one model is fitted: the reduced model containing all predictors not yet deleted.

EXHIBIT 19.2 BACKWARD ELIMINATION PROCEDURE

1. Fit the complete model having all p predictor variables in it. Select the variable having the smallest t or F statistic. The F-statistic (19.3) is the square of the t-statistic (15.15) used to test the significance of β_j in the complete model.
2. Using a significance level of $\alpha = 0.25$, test the significance of the predictor variable. If it is significant terminate the procedure and retain all the predictor variables. If it is not significant delete the predictor variable, set $r = 1$, and proceed to step 3.
3. Fit the reduced model having $p - r$ predictor variables remaining from the previous step. Calculate F-statistics for testing the significance of each of the remaining variables. Select the predictor variable having the smallest F-statistic. The degrees of freedom of this F-statistic are $v_1 = 1$ and $v_2 = n - p + r - 1$.
4. Using a significance level of $\alpha = 0.25$, test the significance of the selected predictor variable. If the F-statistic is significant, terminate the procedure and retain all of the remaining $p - r$ predictor variables from the previous stage of the selection process. If it is not significant, delete the selected variable, increment r by 1, and return to step 3.

Table 19.3 illustrates the application of BE to the chemical-yield data. At stage 1 the interaction between conversion and flow is deleted. At stage 2, flow is deleted. No further predictors are deleted at the next stage, so the procedure terminates. The optimal subset that was obtained from an examination of all possible reduced models has also been selected by BE.

As with forward selection, backward elimination cannot guarantee the selection of optimal subsets. Likewise, only one reduced model is selected; no alternative subsets are identified. Backward elimination begins with the complete model and a good estimate of the error standard deviation. For these reasons, BE is preferred to FS unless the number of predictors is so large that BE is inefficient or, if $n < p + 1$, BE cannot calculate estimates of the regression coefficients. We again caution that no order of importance should be attached to the predictor variables based on the size of the t or F statistics.

Most computer programs that compute backward-elimination F statistics recalculate the mean-squared error after each predictor variable is eliminated. If several predictor variables are eliminated, there is a danger that the mean-squared error could become biased. Because only small F statistics cause predictors to be eliminated, the sums of squares that are added to the error sum of squares can cause the mean-squared error to underestimate the error variance. If this is a concern, the mean-squared error from the original fit to the complete data set with all the predictors included should be used in the denominator of Equation (19.3).

TABLE 19.3 Backward Elimination Procedure for the Chemical-Yield Data

Stage	Candidate Variable for Deletion	F-Statistic for Candidate Variable	p-Value
1	CAT	5.03	0.030
	CON	31.02	0.000
	FLOW	0.29	0.590
	INV	2.84	0.099
	INT*	0.00	0.978
2	CAT	5.35	0.025
	CON	32.04	0.000
	FLOW*	0.33	0.570
	INV	2.91	0.095
3	CAT	5.10	0.028
	CON	38.05	0.000
	INV	8.55	0.005

*Variable deleted

19.3.3 Stepwise Iteration

The stepwise iteration (SI) procedure adds predictor variables one at a time like a forward selection procedure, but at each stage of the procedure the deletion of variables is permitted (see Exhibit 19.3). It combines features of both forward selection and backward elimination.

EXHIBIT 19.3 STEPWISE ITERATION PROCEDURE

1. Initiate the selection procedure as in forward selection.
2. After a predictor variable is added, apply steps 3 and 4 of the backward elimination procedure to the predictor variables in the model.
3. Continue the selection procedure as in forward selection. At each stage of the process, if a predictor variable is added, return to step 2. Terminate the procedure when no predictor variables are added at a stage of forward selection.

Table 19.4 illustrates the application of the SI procedure. Stages 1 and 2 are identical to the same stages of forward selection shown in Table 19.2. Once two predictors are in the model, backward elimination is performed as each variable is entered. Thus, stage 3 is a backward-elimination stage for the reduced model containing the two variables entered at stage 2. Because each of the significance probabilities at stage 3 is less than 0.25, both predictors are retained. At stage 4 forward selection is again performed with the remaining

predictors. Stage 5 is a backward elimination procedure for the reduced model containing the three predictors resulting from stage 4. Again, none of the predictors is deleted, so stage 6 is the next stage of forward selection. Because no further variables are entered at stage 6, the process terminates with the optimal subset of three predictors.

More computational effort is involved with stepwise iteration than with either forward selection or backward elimination. The tradeoff for the additional computational effort is the ability to delete nonsignificant predictors as variables are added and the ability to add new predictors following deletion. In this sense it is a compromise between the strict addition or deletion of FS and BE and the major expenditure of effort involved with the evaluation of all possible reduced models. We again stress that the advantages of being able to assess all possible subsets of predictors outweighs the additional computational effort involved unless the number of predictor variables renders this procedure impractical.

TABLE 19.4 Stepwise Iteration Procedure for the Chemical-Yield Data

	Predictor Variables				
Stage	In	Added	Deleted	F-Statistic	p-Value
1	None	CAT		2.67	0.109
	None	CON*		17.19	0.000
	None	FLOW		2.13	0.150
	None	INV		0.23	0.633
	None	INT		0.01	0.915
2	CON	CAT		9.54	0.003
	CON	FLOW		5.72	0.021
	CON	INV*		13.33	0.001
	CON	INT		0.11	0.754
3	CON, INV		CON	34.30	0.000
	CON, INV		INV	13.33	0.001
4	CON, INV	CAT*		5.10	0.028
	CON, INV	FLOW		0.01	0.915
	CON, INV	INT		0.14	0.712
5	CON, INV, CAT		CAT	5.10	0.028
	CON, INV, CAT		CON	38.05	0.000
	CON, INV, CAT		INV	8.55	0.005
6	CON, INV, CAT	FLOW		0.33	0.570
	CON, INV, CAT	INT		0.03	0.869

*Variable added.

19.4 COLLINEAR EFFECTS

In the last section it was stressed that different subset selection methods need not result in the selection of the same subset of predictors, nor need the subsets selected be optimal according to the selection criteria used. The chemical-yield data were used to illustrate the stepwise procedures in the last section in part because they do yield the optimal subset regardless of which of the methods is employed. Suitably well-behaved data sets can be expected to yield consistent results on the various subset selection methods. Many designed experiments, notably complete factorials, yield identical subsets regardless of the selection method used.

One property of data sets that is known often to produce results depending on the subset selection method is collinearity among predictor variables. Collinearities (redundancies) were introduced in Section 15.4 in the context of fitting polynomial regression models. In this section we illustrate some of the effects of collinearities, especially their effects on subset selection methods.

Data on the calibration of a cryogenic flowmeter consist of 120 observations on a response variable (PD) and seven predictor variables: TEMP, PRESS, DENS, COOL, TIME, PULSE, and ULLAGE. A description of these variables and a listing of this data set is appended to this chapter. After reviewing residual plots, there appears to be a need to add a quadratic term for temperature. We add such a term, TEMP2, after first standardizing the temperature using the correlation-form standardization, as in equation (15.26). Both the standardized temperature and TEMP2 are included in the model.

An initial assessment of the resulting fit, ignoring the questions of model specification that were raised in Section 18.2, suggests that the fit is satisfactory. The coefficient of determination calculated from the sums of squares in Table 19.5 is 78%; the estimated error standard deviation is 0.112. Several of the t-statistics are nonsignificant, leading to a consideration of variable selection to reduce the model.

Forward selection selects six predictor variables, all those listed in Table 19.5 except TEMP and TIME. Backward elimination also selects six predictors, deleting DENS and TIME. Stepwise iteration agrees with forward selection. An examination of all possible reduced models indicates that both of these six-variable subsets are acceptable, the FS subset having $R^2 = 0.775$ and $C_7 = 6.08$, and the BE subset having the same R^2 with $C_7 = 6.17$. What is disturbing about these results is that the t-statistics in Table 19.5 suggest that both DENS and TEMP can be deleted but neither of the above subsets deletes both.

The reason for these apparently contradictory results is that DENS and TEMP are highly collinear. The correlation coefficient (Pearson's r) between these two predictors is -0.999. As explained in the reference from which these data are taken, the process under study is such that the density measurements of the liquid nitrogen used in the experiment are solely a function of temperature

TABLE 19.5 Regression Fit to Cryogenic-Flowmeter Data

ANOVA

Source	df	Sums of Squares	Mean Squares	F	p-Value
Regression	8	4.908	0.614	48.48	0.000
Error	111	1.405	0.013		
Total	119	6.313			

Predictor Variable*	Estimate	Standard Error	t	p-Value	VIF
Constant	−8.300	23.514	−0.35	0.725	
TEMP	−1.863	4.416	−0.42	0.674	1,541.1
TEMP2	4.258	1.157	3.68	0.000	2.3
PRESS	0.023	0.009	2.62	0.010	101.3
DENS	1.128	3.558	0.32	0.752	1,255.7
COOL	−0.150	0.061	−2.44	0.016	145.3
TIME	−0.000	0.000	−0.99	0.325	1.2
PULSE	0.120	0.003	5.88	0.000	1.6
ULLAGE	−0.002	0.001	−2.94	0.004	1.2

*Both TEMP and TEMP2 use the correlation-form standardization of TEMP.

and pressure, and the dependence on pressure was thought to be minimal throughout the experimental region.

Collinearities such as this one often lead to inconsistent variable-selection results when different methods are applied. Other problems are also associated with collinearities. The standard errors of least-squares coefficient estimators, Section 15.2, are proportional to their *variance inflation factors* (VIF):

$$\text{VIF}_j = (1 - R_j^2)^{-1}, \tag{19.7}$$

where R_j^2 is the coefficient of determination of the regression of x_j on the constant term and the other predictor variables in the model. A predictor variable that is highly collinear with the constant term or one or more other predictors will have a R_j^2 value close to 1. This in turn will produce a large variance inflation factor and, unless the estimated error standard deviation s_e is compensatingly small, a large estimated standard error for the least-squares estimate b_j.

The dramatic effect of the collinearity between DENS and TEMP on their variance inflation factors can be seen in the last column of Table 19.5. Each of their variance inflation factors is over 1000. This means that the estimators

standard errors are over 30 times larger than they would be if the two predictor variables were completely uncorrelated with all other predictors in the model ($R_j^2 = 0$, VIF = 1). It is interesting to note that there appear to be other collinear variables in this data set, because both PRESS and COOL have variance inflation factors that exceed 100.

Among the effects known to result from collinearities are the following. Coefficient estimates often tend to be too large in magnitude and occasionally of the wrong sign. Variance inflation factors can be orders of magnitude larger than their minimum value of 1; consequently, standard errors of coefficient estimators tend to be much larger than for estimators of noncollinear predictors. Finally, variable selection procedures, including t-statistics from the full model fit, can erroneously indicate addition or deletion of collinear variables.

Although strong collinearities cause difficulties such as these, their effects can sometimes be minimized. Proper consideration of variables, including the elimination of variables known to be always redundant, and standardization prior to the formation of polynomial terms can reduce the occurrence of collinearities. Variable selection techniques can be used at times to screen for collinear predictors. Biased estimators of the regression coefficients are another alternative. These topics are explored more fully in the references.

REFERENCES

Text References

Most of the regression texts referenced at the end of Chapters 14 to 16 contain discussions devoted to variable selection techniques and the effects of collinearities. Three additional articles that specifically address this topic are:

Gunst, R. F. (1983). "Regression Analysis with Multicollinear Predictor Variables: Definition, Detection, and Effects," *Communications in Statistics*, **12**, 2217–2260.

Henderson, H. V. and Velleman, P. F. (1981). "Building Multiple Regression Models Interactively," *Biometrics*, **37**, 391–441.

Hocking, R. R. (1976). "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, **32**, 1–49.

APPENDIX: CRYOGENIC-FLOWMETER DATA

This data set contains 120 observations collected during an experiment to evaluate the accuracy and precision of a new facility for calibrating cryogenic flowmeters. The variable definitions are:

PD Percent difference between the weight recorded by the flowmeter under test and the weight provided by an independent weighing process.
TEMP Temperature of the liquid nitrogen at the meter (K).
PRESS Pressure at the flow meter (lb/in.^{2}).
DENS Density of the liquid nitrogen (lb/gal).
COOL Subcooling temperature (amount below the boiling point of the liquid) (K).
TIME Elapsed time of the accumulation at the meter (sec).
PULSE Number of pulses of the photoelectric cell counting revolutions of the flowmeter shaft.
ULLAGE Ullage temperature of the gaseous helium above the liquid nitrogen in the weighing tank (K).

Cryogenic-Flowmeter Data

TEMP	PRESS	DENS	COOL	TIME	PULSE	ULLAGE	PD
78.88	82.3	6.69	17.27	78.93	59	120.55	−0.4861
79.02	83.0	6.68	17.24	85.79	65	106.15	−0.4252
79.00	83.2	6.68	17.29	84.92	65	103.07	−0.1933
79.35	86.7	6.67	17.47	84.73	65	84.84	−0.3660
78.51	85.2	6.70	18.08	84.79	65	88.64	−0.3951
77.89	83.7	6.73	18.48	83.22	64	89.25	−0.4165
78.00	80.0	6.72	17.79	78.38	58	85.03	−0.5655
78.07	82.0	6.72	18.03	74.64	58	87.98	−0.5542
78.10	82.4	6.72	18.07	75.26	58	87.87	−0.5382
78.95	89.7	6.69	18.29	72.50	56	82.82	−0.4643
79.07	90.5	6.68	18.28	73.51	57	86.80	−0.3763
79.21	90.8	6.68	18.18	72.35	56	87.54	−0.4361
87.31	97.0	6.35	10.89	92.59	74	95.46	0.3343
87.77	98.6	6.33	10.63	93.71	74	98.38	0.4132
87.89	98.0	6.32	10.44	95.24	75	97.39	0.4157
87.75	106.3	6.33	11.55	159.48	74	99.41	0.4501
87.36	109.1	6.35	12.25	157.83	74	96.54	0.3557
87.13	108.5	6.36	12.42	158.85	74	99.55	0.3706
79.76	104.0	6.66	19.27	100.14	70	99.87	−0.3618
79.42	105.7	6.67	19.82	101.44	70	101.82	−0.2609
79.39	106.5	6.67	19.93	101.60	70	103.06	−0.2493
79.51	107.0	6.67	19.87	79.08	55	102.46	−0.3516
79.53	107.6	6.67	19.91	80.41	56	100.39	−0.3213
79.49	107.5	6.67	19.95	80.25	56	96.75	−0.2731
79.51	107.3	6.67	19.90	99.88	69	90.74	−0.2688
79.46	108.2	6.67	20.05	98.50	68	96.52	−0.2587

TEMP	PRESS	DENS	COOL	TIME	PULSE	ULLAGE	PD
79.46	107.9	6.67	20.01	99.21	69	92.53	−0.2354
79.51	108.0	6.67	19.98	99.91	69	91.38	−0.2296
79.76	108.3	6.66	19.76	95.99	67	87.91	−0.2166
81.13	109.5	6.60	18.52	88.92	68	83.64	−0.2512
79.56	104.8	6.67	19.57	88.60	67	85.15	−0.2688
78.61	102.2	6.70	20.22	88.24	67	85.85	−0.3154
79.69	94.1	6.66	18.13	91.09	67	115.55	−0.4985
79.81	94.3	6.65	18.04	90.76	67	109.69	−0.3540
79.90	94.5	6.65	17.98	89.79	66	102.59	−0.4536
81.22	87.8	6.60	15.75	87.29	67	143.33	−0.4524
81.50	88.2	6.59	15.53	88.58	68	131.99	−0.3122
81.73	88.6	6.58	15.36	87.49	67	122.64	−0.2805
82.03	90.4	6.57	15.31	212.45	68	110.55	−0.3796
81.82	90.5	6.57	15.53	199.38	67	107.80	−0.2091
81.87	91.0	6.57	15.55	210.06	68	104.87	−0.1859
81.80	88.4	6.57	15.26	108.84	68	101.56	−0.1738
81.64	88.4	6.58	15.42	109.81	68	100.05	−0.2340
81.59	88.7	6.58	15.51	109.40	68	98.90	−0.1695
81.59	92.4	6.58	16.02	151.85	68	86.51	−0.0828
81.71	94.2	6.58	16.14	148.84	67	90.39	−0.0902
81.78	94.9	6.58	16.16	151.62	68	93.55	−0.0587
81.78	98.6	6.58	16.62	98.48	67	85.08	−0.0431
82.05	101.0	6.57	16.63	94.06	68	89.86	−0.0608
82.19	101.3	6.56	16.53	93.71	68	89.41	−0.0730
82.26	101.3	6.56	16.46	92.57	68	89.68	−0.0811
81.24	73.8	6.59	13.43	88.62	69	99.26	−0.3291
81.41	73.9	6.59	13.29	88.92	69	98.79	−0.3254
81.52	73.8	6.58	13.15	87.48	68	97.69	−0.3425
78.65	75.6	6.70	16.36	91.27	69	89.23	−0.3474
78.51	75.9	6.70	16.55	91.02	69	91.82	−0.3491
78.37	76.2	6.71	16.75	91.41	69	92.15	−0.3235
78.98	81.9	6.68	17.11	92.21	69	83.65	−0.2837
79.09	82.9	6.68	17.15	92.32	69	87.15	−0.2155
79.18	83.5	6.68	17.15	91.02	68	87.51	−0.2624
82.33	90.5	6.55	15.02	90.72	70	131.25	−0.2146
82.56	90.5	6.54	14.79	90.98	70	118.27	−0.1794
82.63	90.9	6.54	14.78	90.75	70	111.47	−0.1204
80.97	83.5	6.61	15.37	88.53	69	86.94	−0.2040
81.08	84.3	6.60	15.38	88.59	69	88.95	−0.1725
81.13	85.0	6.60	15.44	89.83	70	89.51	−0.1974
81.66	87.1	6.58	15.21	89.15	70	84.50	−0.0589
81.82	88.4	6.57	15.24	88.55	70	87.86	−0.0766

TEMP	PRESS	DENS	COOL	TIME	PULSE	ULLAGE	PD
82.01	89.5	6.57	15.21	88.59	70	88.10	−0.0748
82.88	94.6	6.53	15.01	90.77	70	86.50	−0.0712
83.23	92.3	6.52	14.36	90.99	70	99.83	−0.0397
81.54	93.1	6.59	16.15	91.91	69	106.76	−0.2226
81.54	93.1	6.59	16.15	93.22	70	106.76	−0.2038
81.57	93.0	6.58	16.11	93.46	70	102.15	−0.1725
81.57	94.5	6.58	16.31	94.82	70	85.86	−0.1255
81.59	95.3	6.58	16.39	95.21	70	88.64	−0.0863
81.68	96.0	6.58	16.38	95.25	70	88.68	−0.0432
82.49	99.4	6.55	16.00	94.35	70	85.31	0.0217
82.61	100.7	6.54	16.04	91.14	70	87.62	0.0099
82.63	101.1	6.54	16.06	91.14	70	88.10	0.0692
84.80	115.8	6.46	15.57	112.44	71	87.90	0.2890
85.12	120.0	6.45	15.75	97.59	72	91.02	0.4289
85.21	105.4	6.44	13.98	95.60	72	98.78	0.2603
85.84	109.8	6.41	13.84	95.42	72	88.88	0.4350
86.00	111.3	6.41	13.84	95.09	72	92.06	0.4909
79.97	92.7	6.65	17.67	90.89	69	111.00	0.0493
80.37	92.0	6.63	17.18	92.92	70	106.57	−0.1090
80.37	91.7	6.63	17.14	91.92	69	102.66	−0.2092
80.62	88.6	6.62	16.46	91.01	69	85.34	−0.1009
80.67	89.3	6.62	16.51	92.19	70	88.50	−0.0781
80.78	89.6	6.62	16.44	92.42	70	88.91	−0.1347
81.61	92.7	6.58	16.02	91.16	69	84.70	−0.0199
81.78	94.4	6.58	16.09	92.08	70	88.11	−0.0020
81.91	95.5	6.57	16.09	93.57	71	89.06	0.0349
83.09	99.2	6.52	15.37	92.91	71	87.05	0.1232
83.23	99.8	6.52	15.30	92.64	71	91.03	0.1351
83.32	99.8	6.52	15.21	92.99	71	90.60	0.0646
77.47	68.9	6.74	16.25	88.86	68	147.56	−0.1911
77.61	68.9	6.74	16.11	89.03	68	143.52	−0.1736
77.65	69.3	6.73	16.15	89.00	68	142.36	−0.1105
77.98	73.6	6.72	16.66	90.82	69	88.38	0.0859
78.17	74.7	6.72	16.69	88.38	68	90.19	−0.0376
78.19	75.3	6.71	16.78	89.54	69	90.22	−0.0703
78.98	77.3	6.68	16.35	89.73	69	82.64	−0.0333
77.93	75.7	6.72	17.10	88.44	68	84.31	−0.1601
77.17	74.6	6.75	17.67	88.62	68	85.35	−0.1908
77.00	79.6	6.76	18.72	88.76	68	82.01	0.0020
77.07	80.7	6.76	18.83	88.94	68	85.01	−0.1671
77.10	81.4	6.76	18.91	90.09	69	85.59	0.0097
77.91	65.0	6.72	14.99	89.46	68	111.92	−0.2960

TEMP	PRESS	DENS	COOL	TIME	PULSE	ULLAGE	PD
77.98	65.7	6.72	15.07	90.10	68	107.46	−0.2922
78.05	66.2	6.72	15.11	89.44	68	103.85	−0.2805
76.91	69.4	6.76	16.90	87.41	67	85.88	−0.4197
76.80	69.7	6.77	17.08	89.07	68	86.64	−0.3098
76.84	77.9	6.77	18.59	88.90	68	87.54	−0.4386
77.21	78.3	6.75	18.29	88.72	68	82.52	−0.3418
77.21	79.4	6.75	18.47	88.99	68	86.01	−0.3202
77.31	80.0	6.75	18.47	87.54	67	86.30	−0.3887
78.19	84.2	6.72	18.25	88.60	68	85.42	−0.2432
78.30	84.5	6.71	18.18	89.41	68	86.35	−0.2571

Data from Joiner, B. (1977), "Evaluation of Cryogenic Flow Meters: An Example in Non-standard Experimental Design and Analysis," *Technometrics*, **19**, 353–380. Copyright American Statistical Association, Alexandria, VA. Reprinted by permission.

EXERCISES

1 The effects of fuel properties on startability were evaluated in a diesel engine using 18 test fuels. The minimum starting temperature was recorded as each test fuel flowed through the engine. Analysis of the fuel yielded five fuel properties: cetane number, viscosity, distillation temperature at 90% boiling point, auto-ignition temperature, and flash point (see accompanying table). Perform all possible subset comparisons using the five fuel properties as predictor variables. Produce a C_m-plot of the C_m-statistics. Which subset(s) do you feel may be candidate models for the engine starting temperature?

Minimum Starting Temperature	Cetane Number	Viscosity	90% Distillation	Auto-ignition Temperature	Flash Point
−9.6	57	1.95	304.8	252	82.6
−2.0	45	1.50	257.8	257	63.6
0.0	35	0.78	246.7	252	−24.0
8.0	28	0.76	236.7	209	−21.0
11.0	35	3.74	361.0	217	32.0
1.8	41	1.07	221.3	262	36.0
−10.4	47	1.56	258.4	272	55.0
−5.0	50	3.00	343.0	252	77.0
−5.2	43	1.12	481.2	198	−21.0

Minimum Starting Temperature	Cetane Number	Viscosity	90% Distillation	Auto-ignition Temperature	Flash Point
−8.6	49	1.39	474.0	192	43.0
−7.0	48	2.07	469.6	192	64.0
−10.8	60	2.57	471.8	186	84.0
−5.0	39	0.82	250.5	197	−22.0
3.0	40	0.78	234.0	197	−2.0
2.5	37	3.73	357.0	211	37.0
−10.4	42	1.46	251.0	192	55.0
−5.5	39	1.80	329.3	197	50.0
−5.5	44	1.49	281.4	190	60.5

2 Analyze the data in Exercise 1 using the forward selection, backward elimination, and stepwise iteration procedures with a level of significance $\alpha = 0.25$. Compare the results of these procedures with the conclusions drawn in Exercise 1.

3 An experiment was conducted to determine which factors affect the knock intensity of an engine during the combustion process. Twelve different fuels were chosen in this study, and varying levels of speed, load, fuel flow, manifold pressure, and brake horsepower were applied as each fuel flowed through the engine. A listing of the resultant knock-intensity values is given in the table on the next page. Identify one or more better subsets of the predictor variables by summarizing all possible reduced models in a C_m plot. If all subsets can't be computed, see Exercise 4.

Response variable

KNOCK (knock intensity, psi/angular degree)

Predictor variables

SPEED (rpm)
LOAD (lb/ft)
FUELFLOW (fuel flow, lb/hr)
EXTEMP (exhaust temperature, °F)
MANPRESS (manifold pressure, inches of mercury)
BHP (brake horsepower, hp)
GRAVITY (specific gravity, PAPI)
VISCOS (viscosity, cS, at 40°C)
D50 (distillation, 50% recovery, °C)
FBP (final boiling point, °C)

4 Perform the forward selection, backward elimination, and stepwise iteration procedures on the knock-intensity data in Exercise 3. If all three procedures do not result in the same preferred subset, examine the data

Knock-Intensity Values (Exercise 3)

OBS	SPEED	LOAD	FUELFLOW	EXTEMP	MANPRESS	BHP	GRAVITY	VISCOS	D50	FBP	KNOCK
1	1404	885.9	90.2	1278	4.0	236.8	30.9	2.20	256.3	418.3	148.90
2	1800	1010.9	123.6	1217	9.0	346.5	30.9	2.20	256.3	418.3	130.50
3	2200	1016.8	154.8	1175	14.0	444.8	30.9	2.20	256.3	418.3	121.60
4	2600	981.8	165.0	1239	15.8	486.0	30.9	2.20	256.3	418.3	123.70
5	1400	869.4	84.5	1237	4.0	231.8	37.5	2.51	278.9	486.4	123.80
6	1800	990.0	110.2	1148	8.0	339.3	37.5	2.51	278.9	486.4	111.20
7	2200	1032.8	137.8	1095	13.0	432.6	37.5	2.51	278.9	486.4	95.81
8	2600	954.3	157.5	1133	15.0	472.4	37.5	2.51	278.9	486.4	95.48
9	1400	871.8	87.3	1223	4.0	232.4	33.7	1.76	249.4	421.3	228.30
10	1800	969.8	109.9	1158	8.0	332.4	33.7	1.76	249.4	421.3	175.70
11	2200	1000.4	132.7	1097	12.5	419.1	33.7	1.76	249.4	421.3	150.50
12	2600	930.4	150.8	1137	15.0	460.6	33.7	1.76	249.4	421.3	156.60
13	1401	886.4	90.8	1295	4.0	236.5	35.8	3.20	283.3	481.2	90.03
14	1800	1013.8	124.4	1212	9.0	347.5	35.8	3.20	283.3	481.2	86.25
15	2200	1070.3	153.7	1152	13.9	448.3	35.8	3.20	283.3	481.2	87.50
16	2600	1018.9	178.9	1234	15.5	504.4	35.8	3.20	283.3	481.2	84.32
17	1403	860.8	88.8	1213	4.0	230.0	35.9	1.56	239.1	421.3	232.70
18	1803	955.2	105.5	1144	8.0	327.9	35.9	1.56	239.1	421.3	183.90
19	2201	994.2	139.5	1105	13.0	416.6	35.9	1.56	239.1	421.3	161.60
20	2601	908.9	147.1	1126	14.5	450.1	35.9	1.56	239.1	421.3	162.20
21	1401	840.1	85.8	1204	4.0	224.1	38.8	1.29	222.8	421.3	266.70
22	1801	940.2	110.8	1136	8.0	322.4	38.8	1.29	222.8	421.3	191.40
23	2206	980.2	130.2	1097	12.3	411.7	38.8	1.29	222.8	421.3	168.80
24	2604	870.7	159.4	1101	14.0	431.7	38.8	1.29	222.8	421.3	174.30
25	1402	829.4	86.4	1168	3.5	221.4	39.5	1.33	224.8	414.7	356.00

OBS	SPEED	LOAD	FUELFLOW	EXTEMP	MANPRESS	BHP	GRAVITY	VISCOS	D50	FBP	KNOCK
26	1805	930.5	108.5	1129	9.0	319.8	39.5	1.33	224.8	414.7	246.20
27	2201	965.4	130.4	1100	12.0	404.6	39.5	1.33	224.8	414.7	212.50
28	2600	873.2	146.2	1112	13.8	432.3	39.5	1.33	224.8	414.7	233.60
29	1404	830.4	84.1	1182	3.5	222.0	35.6	1.65	241.3	421.3	304.60
30	1801	949.1	111.7	1148	8.0	325.5	35.6	1.65	241.3	421.3	216.20
31	2202	960.3	123.2	1095	12.0	402.6	35.6	1.65	241.3	421.3	176.20
32	2602	874.8	145.3	1123	14.5	433.4	35.6	1.65	241.3	421.3	194.60
33	1402	813.8	84.1	1132	3.5	217.2	41.4	1.19	213.9	422.6	386.80
34	1801	927.8	106.9	1119	8.0	318.2	41.4	1.19	213.9	422.6	255.00
35	2202	943.2	122.4	1079	12.2	395.5	41.4	1.19	213.9	422.6	241.90
36	2602	850.7	144.9	1097	13.8	421.5	41.4	1.19	213.9	422.6	251.60
37	1400	847.2	85.0	1230	2.5	225.8	43.2	1.08	197.3	412.9	238.90
38	1800	940.4	104.8	1151	8.9	322.3	43.2	1.08	197.3	412.9	376.60
39	2200	990.2	140.7	1127	13.8	414.8	43.2	1.08	197.3	412.9	331.70
40	2600	851.2	142.5	1125	14.4	421.4	43.2	1.08	197.3	412.9	401.10
41	1400	804.8	90.7	1360	3.3	214.5	40.1	2.00	271.1	475.2	113.40
42	1800	893.7	115.2	1342	6.0	306.3	40.1	2.00	271.1	475.2	74.00
43	2200	965.1	153.8	1315	9.0	404.3	40.1	2.00	271.1	475.2	78.11
44	2600	879.6	154.3	1302	10.0	435.5	40.1	2.00	271.1	475.2	93.30
45	1400	824.9	86.0	1312	4.0	219.9	42.7	1.60	263.0	466.0	80.71
46	1800	952.6	111.9	1245	7.6	326.5	42.7	1.60	263.0	466.0	67.05
47	2200	1004.1	145.7	1208	11.5	420.6	42.7	1.60	263.0	466.0	79.17
48	2600	890.6	156.5	1202	12.4	440.9	42.7	1.60	263.0	466.0	81.20

further to identify the reason for the differences. Which subset(s) do you recommend? Why?

5 Identify the better subsets for the complete quadratic model for the oxides of nitrogen data in Exercise 2 of Chapter 15. Do not center or scale the raw predictors. Compare the results of the stepwise selection methods (FS, BE, SI). Do the strong collinearities among the polynomial terms have a noticeable effect on the selection procedures?

6 Repeat Exercise 5 with standardized predictors. How do the results of the stepwise methods change?

7 Identify the better subsets for the height data in Exercise 3, Chapter 15. Next, add two predictors, the brachial index and the tibiofemural index defined in Exercise 7 of Chapter 15. Redo the variable selection process. Are any of the previous results altered? Are either of the two additional predictors included in the better subsets?

8 Standardize the predictors for the engine-knock data in Exercise 6 of Chapter 18. Form a complete second-order model from the standardized predictors. Investigate whether the second-order terms aid in the ability to predict the knock measurements.

9 Determine the variance inflation factors and standard errors for the coefficient estimators of the second-order fit in Exercise 4 in Chapter 15. Based on these results and any other information available in the data (e.g., predictor-variable pairwise correlations), comment on any possible collinearities among the predictor variables.

10 The engine-startability experiment described in Exercise 9 in Chapter 15 was expanded to include twelve predictor variables. The revised data set is given below. Identify one or more better subsets of the predictor variables by summarizing all possible reduced models in a C_m plot.

Response variable

START TIME (seconds)

Predictor variables

CRANK (cranking speed, rrm)
POUR (pour point, °C)
ANILINE (aniline point, °C)
CETANE (cetane number)
GRAVITY (specific gravity at 15.6°C)
FLASH (flash point, °C)
VISC (viscosity at 40°C)
ARO (total aromatics, vol%)
CLOUD (cloud point, °C)
AUTOTP (auto-ignition temperature, °C)
BP30 (30% boiling point, °C)
BP70 (70% boiling point, °C)

FUEL	START	CRANK	POUR	ANILINE	CETANE	GRAVITY	FLASH	VISC	ARO	CLOUD	AUTOTP	BP30	BP70
A	26.5	181	−17	66.1	48.7	0.8484	77	2.75	26.3	−13	189	246	304
A	15.7	198	−17	66.1	48.7	0.8484	77	2.75	26.3	−13	189	246	304
A	8.4	206	−17	66.1	48.7	0.8484	77	2.75	26.3	−13	189	246	304
A	4.0	232	−17	66.1	48.7	0.8484	77	2.75	26.3	−13	189	246	304
A	8.3	216	−17	66.1	48.7	0.8484	77	2.75	26.3	−13	189	246	304
A	9.0	218	−17	66.1	48.7	0.8484	77	2.75	26.3	−13	189	246	304
A	5.0	218	−17	66.1	48.7	0.8484	77	2.75	26.3	−13	189	246	304
B	20.0	227	−13	62.7	46.2	0.8519	75	2.67	28.9	−13	186	241	302
B	17.2	227	−13	62.7	46.2	0.8519	75	2.67	28.9	−13	186	241	302
B	29.1	210	−13	62.7	46.2	0.8519	75	2.67	28.9	−13	186	241	302
B	22.8	194	−13	62.7	46.2	0.8519	75	2.67	28.9	−13	186	241	302
B	30.4	195	−13	62.7	46.2	0.8519	75	2.67	28.9	−13	186	241	302
C	38.3	194	−17	62.8	45.2	0.8545	74	2.53	33.5	−14	185	238	298
C	26.3	202	−17	62.8	45.2	0.8545	74	2.53	33.5	−14	185	238	298
C	33.0	193	−17	62.8	45.2	0.8545	74	2.53	33.5	−14	185	238	298
D	48.2	195	−19	55.6	43.6	0.8565	73	2.44	35.4	−16	189	236	294
D	31.3	200	−19	55.6	43.6	0.8565	73	2.44	35.4	−16	189	236	294
E	39.3	190	−14	63.0	47.0	0.8540	76	2.77	28.4	−14	185	247	304
E	35.8	193	−14	63.0	47.0	0.8540	76	2.77	28.4	−14	185	247	304
F	42.0	190	−14	60.5	44.9	0.8597	76	2.75	29.7	−12	188	248	305
F	39.1	186	−14	60.5	44.9	0.8597	76	2.75	29.7	−12	188	248	305
G	30.3	196	−17	61.7	45.6	0.8565	76	2.75	31.2	−13	188	247	305
H	23.4	203	−14	64.5	48.1	0.8514	76	2.74	26.4	−12	191	245	304
I	27.5	203	−16	61.9	46.3	0.8499	68	2.53	30.1	−12	188	238	297
I	30.1	206	−16	61.9	46.3	0.8499	68	2.53	30.1	−12	188	238	297
J	21.5	208	−16	64.6	48.4	0.8493	71	2.62	29.1	−13	188	241	303
J	21.8	206	−16	64.6	48.4	0.8493	71	2.62	29.1	−13	188	241	303
K	35.7	202	−21	53.6	41.8	0.8524	59	2.09	38.0	−14	186	221	293
K	32.9	201	−21	53.6	41.8	0.8524	59	2.09	38.0	−14	186	221	293

11 Analyze the data in Exercise 10 using the forward selection, backward elimination, and stepwise iteration procedures with a level of significance of $\alpha = 0.25$. Compare the results of these procedures with the conclusions drawn in Exercise 10. If all three procedures do not result in the same preferred subset, examine the data further to identify the reasons for the differences. Which subset(s) do you recommend? Why?

12 The experiment described in Exercise 13 in Chapter 15 was expanded to include 10 predictor variables. The revised data set is given below. Analyze the data using the forward selection and stepwise iteration procedures with a level of significance of $\alpha = 0.25$. Identify one or more better subsets of the predictor variables. Why is it not appropriate to use backward elimination or the all-possible procedures in finding the best subset? Explain your answer.

y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
16.20	1.1858	0.3019	1.2826	1.7109	0.4752	0.7037	0.9726	10.1199	40.8237	102.2159
16.81	1.0517	0.2667	1.2129	1.8275	0.4100	0.6945	0.9420	9.9607	40.6181	102.6866
14.41	1.3362	0.3337	1.2707	1.6452	0.5450	0.7085	0.9742	10.0532	40.3893	102.0402
14.37	1.3552	0.3291	1.2396	1.6818	0.5637	0.7058	0.9610	9.9900	40.3389	102.1992
16.96	1.1567	0.2471	1.1810	2.0106	0.5716	0.6820	0.9014	10.1431	41.8941	103.2274
15.22	1.2698	0.2485	1.1248	2.0722	0.7016	0.6800	0.8770	9.9861	41.5360	103.4979
12.49	1.1096	0.1819	0.9935	2.3053	0.7237	0.6575	0.8178	9.6277	40.7045	104.3782
10.16	1.3384	0.2898	1.1225	1.8793	0.5522	0.6862	0.8893	9.8593	40.5520	102.9172
23.01	0.6008	0.0640	1.4032	3.6711	0.5766	0.6428	0.8811	12.0174	52.9478	107.6988
23.70	1.9174	0.4184	1.4014	1.4909	0.9543	0.7246	1.0251	10.3560	40.7563	101.2789

13 Identify the better subsets for a complete quadratic model for the data given in Exercise 20 in Chapter 15. Do not center or scale the raw predictors. Compare the results of the stepwise selection methods (FS, BE, and SI).

14 Repeat Exercise 13 with standardized predictors. How do the results of the stepwise methods change?

15 Determine the variance inflation factors and standard errors for the coefficient estimators of the fit obtained in Exercise 9 in Chapter 15. Based on these results, comment on any possible collinearities among the predictor variables.

16 Data were gathered on a set of diesel engines having similar technology in order to determine the NO_x emissions produced by 27 different fuels. The data set is given below. Identify one or more better subsets of the predictor variables by summarizing all possible reduced models in a C_m plot.

Response variable

LNNOX (natural logarithm of NO_x emissions, gr/hp-hr)

Predictor variables (i.e, fuel properties)

NATCET (natural cetane number of a fuel with no cetane improver)

CETDIFF (difference between cetane number of a fuel with cetane improver and the corresponding natural cetane number of the fuel without cetane improver)

TAROM (total aromatics, vol%)

SULFUR (sulfur, ppm)

SPGRAV (specific gravity)

T10 (10% boiling point, °F)

T50 (50% boiling point, °F)

T90 (90% boiling point, °F)

LNNOX	NATCET	CETDIFF	TAROM	SULFUR	SPGRAV	T10	T50	T90
0.970779	42.3	0	34.306	457	0.8599	450	489	579
0.951658	42.3	0	34.306	457	0.8599	450	489	579
0.966984	42.3	5.8	34.306	457	0.8602	450	489	579
0.963174	42.3	5.8	34.306	457	0.8602	450	489	579
0.978326	42.3	10.4	34.306	457	0.8605	450	489	579
0.966984	42.3	10.4	34.306	457	0.8605	450	489	579
0.845868	42.1	0	13.238	92	0.8289	430	494	582
0.854415	42.1	0	13.238	92	0.8289	430	494	582
0.837248	42.1	5.8	13.238	92	0.8291	430	494	582
0.871293	42.1	5.8	13.238	92	0.8291	430	494	582
0.858662	42.1	10.1	13.238	92	0.8297	430	494	582
0.858662	42.1	10.1	13.238	92	0.8297	430	494	582
0.841567	53.4	0	12.8716	53	0.8301	433	497	581
0.850151	53.4	0	12.8716	53	0.8301	433	497	581
0.936093	42.1	5.8	26.8864	120	0.8569	413	494	574
0.932164	42.1	5.8	26.8864	120	0.8569	413	494	574
0.850151	42.8	0	19.5584	473	0.8279	448	493	577
0.875469	42.8	0	19.5584	473	0.8279	448	493	577
0.867100	42.8	5.2	19.5584	473	0.8282	448	493	577
0.875469	42.8	5.2	19.5584	473	0.8282	448	493	577
0.862890	42.8	10.4	19.5584	473	0.8287	448	493	577
0.891998	42.8	10.4	19.5584	473	0.8287	448	493	577
0.896088	42.2	0	12.78	69	0.8593	461	494	583
0.896088	42.2	0	12.78	69	0.8593	461	494	583
0.891998	42.4	5.3	12.78	69	0.8595	461	494	583
0.912283	42.4	5.3	12.78	69	0.8595	461	494	583
0.916291	42.4	10.6	12.78	69	0.8600	461	494	583

LNNOX	NATCET	CETDIFF	TAROM	SULFUR	SPGRAV	T10	T50	T90
0.900161	42.4	10.6	12.78	69	0.8600	461	494	583
0.883768	42.8	0	26.7032	143	0.8285	410	494	565
0.871293	42.8	0	26.7032	143	0.8285	410	494	565
0.875469	42.8	0	26.7032	143	0.8285	410	494	565
0.879627	42.8	5.3	26.7032	143	0.8289	410	494	565
0.883768	42.8	5.3	26.7032	143	0.8289	410	494	565
0.883768	42.8	5.3	26.7032	143	0.8289	410	494	565
0.883768	48	0	25.8788	159	0.8292	414	495	577
0.875469	48	0	25.8788	159	0.8292	414	495	577
0.900161	42.8	9.8	26.7032	143	0.8293	410	494	565
0.920283	42.8	9.8	26.7032	143	0.8293	410	494	565
0.936093	46.9	0	30.9168	325	0.8481	433	508	592
0.936093	46.9	0	30.9168	325	0.8481	433	508	592
0.951658	46.9	0	30.9168	325	0.8481	433	508	592
0.932164	46.9	0	30.9168	325	0.8481	433	508	592
0.993252	42.1	5.8	13.238	92	0.8291	430	494	582
1.085189	42.1	5.8	26.8864	120	0.8569	413	494	574
1.036737	42.4	5.3	12.78	69	0.8595	461	494	583
1.026042	42.8	5.3	26.7032	143	0.8289	410	494	565
0.770108	42.1	5.8	13.238	92	0.8291	430	494	582
0.858662	42.1	5.8	26.8864	120	0.8569	413	494	574
0.806476	42.4	5.3	12.78	69	0.8595	461	494	583
0.815365	42.8	5.3	26.7032	143	0.8289	410	494	565
0.912283	45.1	0	31.19	401	0.8483	422.6	503.6	596.3
0.936093	45.1	0	31.19	401	0.8483	422.6	503.6	596.3
0.908259	45.1	0	31.19	401	0.8483	422.6	503.6	596.3
0.841567	41	0	12.69	13.2	0.8272	347.9	423.5	546.8
0.858662	40.2	0	19.83	28.8	0.8336	357.8	435.2	544.1
0.912283	40.2	1.8	30.09	3	0.8384	367.7	436.1	574.7

17 Analyze the data in Exercise 16 using the forward selection, backward elimination, and stepwise iteration procedures with a level of significance of $\alpha = 0.25$. Compare the results of these procedures with the conclusions drawn in Exercise 16. If all three procedures do not result in the same preferred subset, examine the data further to identify the reason for the differences. Which subset(s) do you recommend? Why?

18 Reanalyze the data in Exercise 16 using standardized predictor variables and 8 additional squared terms, one for each fuel property. This should yield 16 candidate linear and quadratic terms for the fuel properties. Identify one or more better subsets of the predictor variables by summarizing all possible reduced models in a C_m plot.

19 Suppose there was interest in examining the 28 two-way interactions between the 8 fuel properties listed in Exercise 16, in addition to examining the 8 linear and 8 squared fuel terms. Devise a strategy for identifying

the better subsets using a level of significance of $\alpha = 0.25$. In doing this, be sure to address the issues of how to (a) maintain model hierarchy so that a quadratic or interaction term does not enter the final equation unless the corresponding linear terms are already in the model, and (b) reduce the effects of collinearity induced by the many related terms and small sample size.

Appendix: Statistical Tables

TABLE A1 Table of Random Numbers

46	96	85	77	27	92	86	26	45	21	89	91	71	42	64	64	58	22	75	81	74	91	48	46	18
44	19	15	32	63	55	87	77	33	29	45	00	31	34	84	05	72	90	44	27	78	22	07	62	17
34	39	80	62	24	33	81	67	28	11	34	79	26	35	34	23	09	94	00	80	55	31	63	27	91
74	97	80	30	65	07	71	30	01	84	47	45	89	70	74	13	04	90	51	27	61	34	63	87	44
22	14	61	60	86	38	33	71	13	33	72	08	16	13	50	56	48	51	29	48	30	93	45	66	29
40	03	96	40	03	47	24	60	09	21	21	18	00	05	86	52	85	40	73	73	57	68	36	33	91
52	33	76	44	56	15	47	75	78	73	78	19	87	06	98	47	48	02	62	03	42	05	32	55	02
37	59	20	40	93	17	82	24	19	90	80	87	32	74	59	84	24	49	79	17	23	75	83	42	00
11	02	55	57	48	84	74	36	22	67	19	20	15	92	53	37	13	75	54	89	56	73	23	39	07
10	33	79	26	34	54	71	33	89	74	68	48	23	17	49	18	81	05	52	85	70	05	73	11	17
67	59	28	25	47	89	11	65	65	20	42	23	96	41	64	20	30	89	87	64	37	93	36	96	35
93	50	75	20	09	18	54	34	68	02	54	87	23	05	43	36	98	29	97	93	87	08	30	92	98
24	43	23	72	80	64	34	27	23	46	15	36	10	63	21	59	69	76	02	62	31	62	47	60	34
39	91	63	18	38	27	10	78	88	84	42	32	00	97	92	00	04	94	50	05	75	82	70	80	35
74	62	19	67	54	18	28	92	33	69	98	96	74	35	72	11	68	25	08	95	31	79	11	79	54
91	03	35	60	81	16	61	97	25	14	78	21	22	05	25	47	26	37	80	39	19	06	41	02	00
42	57	66	76	72	91	03	63	48	46	44	01	33	53	62	28	80	59	55	05	02	16	13	17	54
06	36	63	06	15	03	72	38	01	58	25	37	66	48	56	19	56	41	29	28	76	49	74	39	50
92	70	96	70	89	80	87	14	25	49	25	94	62	78	26	15	41	39	48	75	64	69	61	06	38
91	08	88	53	52	13	04	82	23	00	26	36	47	44	04	08	84	80	07	44	76	51	52	41	59
68	85	97	74	47	53	90	05	90	84	87	48	25	01	11	05	45	11	43	15	60	40	31	84	59
59	54	13	09	13	80	42	29	63	03	24	64	12	43	28	10	01	65	62	07	79	83	05	59	61
39	18	32	69	33	46	58	19	34	03	59	28	97	31	02	65	47	47	70	39	74	17	30	22	65
67	43	31	09	12	60	19	57	63	78	11	80	10	97	15	70	04	89	81	78	54	84	87	83	42
61	75	37	19	56	90	75	39	03	56	49	92	72	95	27	52	87	47	12	52	54	62	43	23	13
78	10	91	11	00	63	19	63	74	58	69	03	51	38	60	36	53	56	77	06	69	03	89	91	24
93	23	71	58	09	78	08	03	07	71	79	32	25	19	61	04	40	33	12	06	78	91	97	88	95
37	55	48	82	63	89	92	59	14	72	19	17	22	51	90	20	03	64	96	60	48	01	95	44	84
62	13	11	71	17	23	29	25	13	85	33	35	07	69	25	68	57	92	57	11	84	44	01	33	66
29	89	97	47	03	13	20	86	22	45	59	98	64	53	89	64	94	81	55	87	73	81	58	46	42
16	94	85	82	89	07	17	30	29	89	89	80	98	36	25	36	53	02	49	14	34	03	52	09	20
04	93	10	59	75	12	98	84	60	93	68	16	87	60	11	50	46	56	58	45	88	72	50	46	11
95	71	43	68	97	18	85	17	13	08	00	50	77	50	46	92	45	26	97	21	48	22	23	08	32
86	05	39	14	35	48	68	18	36	57	09	62	40	28	87	08	74	79	91	08	27	12	43	32	03
59	30	60	10	41	31	00	69	63	77	01	89	94	60	19	02	70	88	72	33	38	88	20	60	86
05	45	35	40	54	03	98	96	76	27	77	84	80	08	64	60	44	34	54	24	85	20	85	77	32
71	85	17	74	66	27	85	19	55	56	51	36	48	92	32	44	40	47	10	38	22	52	42	29	96
80	20	32	80	98	00	40	92	57	51	52	83	14	55	31	99	73	23	40	07	64	54	44	99	21
13	50	78	02	73	39	66	82	01	28	67	51	75	66	33	97	47	58	42	44	88	09	28	58	06
67	92	65	41	45	36	77	96	46	21	14	39	56	36	70	15	74	43	62	69	82	30	77	28	77
72	56	73	44	26	04	62	81	15	35	79	26	99	57	28	22	25	94	80	62	95	48	98	23	86
28	86	85	64	94	11	58	78	45	36	34	45	91	38	51	10	68	36	87	81	16	77	30	19	36
69	57	40	80	44	94	60	82	94	93	98	01	48	50	57	69	60	77	69	60	74	22	05	77	17
71	20	03	30	79	25	74	17	78	34	54	45	04	77	42	59	75	78	64	99	37	03	18	03	36
89	98	55	98	22	45	12	49	82	71	57	33	28	69	50	59	15	09	25	79	39	42	84	18	70
58	74	82	81	14	02	01	05	77	94	65	57	70	39	42	48	56	84	31	59	18	70	41	74	60
50	54	73	81	91	07	81	26	25	45	49	61	22	88	41	20	00	15	59	93	51	60	65	65	63
49	33	72	90	10	20	65	28	44	63	95	86	75	78	69	24	41	65	86	10	34	10	32	00	93
11	85	01	43	65	02	85	69	56	88	34	29	64	35	48	15	70	11	77	83	01	34	82	91	04
34	22	46	41	84	74	27	02	57	77	47	93	72	02	95	63	75	74	69	69	61	34	31	92	13
05	57	23	06	26	23	08	66	16	11	75	28	81	56	14	62	82	45	65	80	36	02	76	55	63
37	78	16	06	57	12	46	22	90	97	78	67	39	06	63	60	51	02	07	16	75	12	90	41	16
23	71	15	08	82	64	87	29	01	20	46	72	05	80	19	27	47	15	76	51	58	67	06	80	54
42	67	98	41	67	44	28	71	45	08	19	47	76	30	26	72	33	69	92	51	95	23	26	85	76
05	83	03	84	32	62	83	27	48	83	09	19	84	90	20	20	50	87	74	93	51	62	10	23	30

TABLE A1 Table of Random Numbers

60	46	18	41	23	74	73	51	72	90	40	52	95	41	20	89	48	98	27	38	81	33	83	82	94
32	80	64	75	91	98	09	40	64	89	29	99	46	35	69	91	50	73	75	92	90	56	82	93	24
79	86	53	77	78	06	62	37	48	82	71	00	78	21	65	65	88	45	82	44	78	93	22	78	09
45	13	23	32	01	09	46	36	43	66	37	15	35	04	88	79	83	53	19	13	91	59	81	81	87
20	60	97	48	21	41	84	22	72	77	99	81	83	30	46	15	90	26	51	73	66	34	99	40	60
67	91	44	83	43	25	56	33	28	80	99	53	27	56	19	80	76	32	53	95	07	53	09	61	98
86	50	76	93	86	35	68	45	37	83	47	44	92	57	66	59	64	16	48	39	26	94	54	66	40
66	73	38	38	23	36	10	95	16	01	10	01	59	71	55	99	24	88	31	41	00	73	13	80	62
55	11	50	29	17	73	97	04	20	39	20	22	71	11	43	00	15	10	12	35	09	11	00	89	05
23	54	33	87	92	92	04	49	73	96	57	53	57	08	93	09	69	87	83	07	46	39	50	37	85
41	48	67	79	44	57	40	29	10	34	58	63	51	18	07	41	02	39	79	14	40	68	10	01	61
03	97	71	72	43	27	36	24	59	88	82	87	26	31	11	44	28	58	99	47	83	21	35	22	88
90	24	83	48	07	41	56	68	11	14	77	75	48	68	08	90	89	63	87	00	06	18	63	21	91
98	98	97	42	27	11	80	51	13	13	03	42	91	14	51	22	15	48	67	52	09	40	34	60	85
74	20	94	21	49	96	51	69	99	85	43	76	55	81	36	11	88	68	32	43	08	14	78	05	34
94	67	48	87	11	84	00	85	93	56	43	99	21	74	84	13	56	41	90	96	30	04	19	68	73
58	18	84	82	71	23	66	33	19	25	65	17	90	84	24	91	75	36	14	83	86	22	70	86	89
31	47	28	24	88	49	28	69	78	62	23	45	53	38	78	65	87	44	91	93	91	62	76	09	20
45	62	31	06	70	92	73	27	83	57	15	64	40	57	56	54	42	35	40	93	55	82	08	78	87
31	49	87	12	27	41	07	91	72	64	63	42	06	66	82	71	28	36	45	31	99	01	03	35	76
69	37	22	23	46	10	75	83	62	94	44	65	46	23	65	71	69	20	89	12	16	56	61	70	41
93	67	21	56	98	42	52	53	14	86	24	70	25	18	23	23	56	24	03	86	11	06	46	10	23
77	56	18	37	01	32	20	18	70	79	20	85	77	89	28	17	77	15	52	47	15	30	35	12	75
37	07	47	79	60	75	24	15	31	63	25	93	27	66	19	53	52	49	98	45	12	12	06	00	32
72	08	71	01	73	46	39	60	37	58	22	25	20	84	30	02	03	62	68	58	38	04	06	89	94
55	22	48	46	72	50	14	24	47	67	84	37	32	84	82	64	97	13	69	86	20	09	80	46	75
69	24	98	90	70	29	34	25	33	23	12	69	90	50	38	93	84	32	28	96	03	65	70	90	12
01	86	77	18	21	91	66	11	84	65	48	75	26	94	51	40	51	53	36	39	77	69	06	25	07
51	40	94	06	80	61	34	28	46	28	11	48	48	94	60	65	06	63	71	06	19	35	05	32	56
58	78	02	85	80	29	67	27	44	07	67	23	20	28	22	62	97	59	62	13	41	72	70	71	07
33	75	88	51	00	33	56	15	84	34	28	50	16	65	12	81	56	43	54	14	63	37	74	97	59
58	60	37	45	62	09	95	93	16	59	35	22	91	78	04	97	98	80	20	04	38	93	13	92	30
72	13	12	95	32	87	99	32	83	65	40	17	92	57	22	68	98	79	16	23	53	56	56	07	47
22	21	13	16	10	52	57	71	40	49	95	25	55	36	95	57	25	25	77	05	38	05	62	57	77
97	94	83	67	90	68	74	88	17	22	38	01	04	33	49	38	47	57	61	87	15	39	43	87	00
09	03	68	53	63	29	27	31	66	53	39	34	88	87	04	35	80	69	52	74	99	16	52	01	65
29	95	61	42	65	05	72	27	28	18	09	85	24	59	46	03	91	55	38	62	51	71	47	37	38
81	96	78	90	47	41	38	36	33	95	05	90	26	72	85	23	23	30	70	51	56	93	23	84	80
44	62	20	81	21	57	57	85	00	47	26	10	87	22	45	72	03	51	75	23	38	38	56	77	97
68	91	12	15	08	02	18	74	56	79	21	53	63	41	77	15	07	39	87	11	19	25	62	19	30
29	33	77	60	29	09	25	09	42	28	07	15	40	67	56	29	58	75	84	06	19	54	31	16	53
54	13	39	19	29	64	97	73	71	61	78	03	24	02	93	86	69	76	74	28	08	98	84	08	23
75	16	85	64	64	93	85	68	08	84	15	41	57	84	45	11	70	13	17	60	47	80	10	13	00
36	47	17	08	79	03	92	85	18	42	95	48	27	37	99	98	81	94	44	72	06	95	42	31	17
29	61	08	21	91	23	76	72	84	98	26	23	66	54	86	88	95	14	82	57	17	99	16	28	99

Source: The RAND Corporation, *A Million Random Degits with 100,100 Normal Deviates*, New York: The Free Press, 1965. Copyright 1955 and 1983, the RAND Corporation. Used by permission.

TABLE A2 Standard Normal Cumulative Probabilities*

z_c	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

z_c	0.675	1.282	1.645	1.960	2.326	2.576
Probability	0.75	0.90	0.95	0.975	0.99	0.995
Upper Tail	0.25	0.10	0.05	0.025	0.01	0.005

*The entries in this table show the probability that a standard normal variate is less than or equal to z_c.

TABLE A3 Student t Cumulative Probabilities*

	Cumulative Probability					
ν	0.75	0.90	0.95	0.975	0.99	0.995
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947
16	0.690	1.337	1.746	2.120	2.583	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.508	2.819
23	0.685	1.319	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.313	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
60	0.679	1.296	1.671	2.000	2.390	2.660
90	0.678	1.291	1.662	1.987	2.368	2.632
120	0.677	1.289	1.658	1.980	2.358	2.617
∞	0.674	1.282	1.645	1.960	2.326	2.576
Upper Tail	0.25	0.10	0.05	0.025	0.01	0.005

*The entries in this table show the value of the student t variate with ν degrees of freedom corresponding to the designated cumulative probability.

TABLE A4 Chi-Square Cumulative Probabilities*

	Cumulative Probability													
ν	0.005	0.010	0.025	0.050	0.100	0.250	0.500	0.750	0.900	0.950	0.975	0.990	0.995	0.999
1					0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.96	26.12
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.73	26.76	31.26
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00	45.32
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40	46.80
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80	48.27
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18	49.73
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56	51.18
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93	52.62
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29	54.05
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64	55.48
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99	56.89
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34	58.30
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.81	55.76	59.34	63.69	66.77	73.40
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.42	104.22	112.32
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.88	106.63	112.33	116.32	124.84
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.65	107.56	113.14	118.14	124.12	128.30	137.21
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.14	118.50	124.34	129.56	135.81	140.17	149.45
Upper Tail	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050	0.025	0.010	0.005	0.001

Source: *Biometrika Table for Statisticians* (1966). Edited by E. S. Pearson and H. O. Hartley, Volume I. Copyright Biometrika Trustees. Reprinted with permission.

*The entries in this table are values of a chi-square variate with ν degrees of freedom corresponding to the designated cumulative probability.

TABLE A5(*a*) *F* Cumulative Probabilities: 0.75 (Upper Tail: 0.25)*

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32	9.41	9.49	9.58	9.63	9.67	9.71	9.76	9.80	9.85
2	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38	3.39	3.41	3.43	3.43	3.44	3.45	3.46	3.47	3.48
3	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44	2.45	2.46	2.46	2.46	2.47	2.47	2.47	2.47	2.47
4	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08
5	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.88	1.88	1.88	1.88	1.87	1.87	1.87
6	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.77	1.76	1.76	1.75	1.75	1.75	1.74	1.74	1.74
7	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.69	1.69	1.68	1.68	1.67	1.67	1.66	1.66	1.65	1.65	1.65
8	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.63	1.62	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.58
9	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.54	1.53	1.53
10	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.53	1.52	1.52	1.51	1.51	1.50	1.49	1.48
11	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.47	1.46	1.45
12	1.46	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.45	1.44	1.43	1.42
13	1.45	1.55	1.55	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.42	1.41	1.40
14	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.41	1.41	1.40	1.39	1.38
15	1.43	1.52	1.52	1.51	1.49	1.48	1.47	1.46	1.46	1.45	1.44	1.43	1.41	1.41	1.40	1.39	1.38	1.37	1.36
16	1.42	1.51	1.51	1.50	1.48	1.47	1.46	1.45	1.44	1.44	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34
17	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33
18	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33	1.32
19	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41	1.40	1.38	1.37	1.36	1.35	1.34	1.33	1.32	1.30
20	1.40	1.49	1.48	1.47	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.32	1.31	1.29
21	1.40	1.48	1.48	1.46	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28
22	1.40	1.48	1.47	1.45	1.44	1.42	1.41	1.40	1.39	1.39	1.37	1.36	1.34	1.33	1.32	1.31	1.30	1.29	1.28
23	1.39	1.47	1.47	1.45	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28	1.27
24	1.39	1.47	1.46	1.44	1.43	1.41	1.40	1.39	1.38	1.38	1.36	1.35	1.33	1.32	1.31	1.30	1.29	1.28	1.26
25	1.39	1.47	1.46	1.44	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.34	1.33	1.32	1.31	1.29	1.28	1.27	1.25
26	1.38	1.46	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.37	1.35	1.34	1.32	1.31	1.30	1.29	1.28	1.26	1.25
27	1.38	1.46	1.45	1.43	1.42	1.40	1.39	1.38	1.37	1.36	1.35	1.33	1.32	1.31	1.30	1.28	1.27	1.26	1.24
28	1.38	1.46	1.45	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.34	1.33	1.31	1.30	1.29	1.28	1.27	1.25	1.24
29	1.38	1.45	1.45	1.43	1.41	1.40	1.38	1.37	1.36	1.35	1.34	1.32	1.31	1.30	1.29	1.27	1.26	1.25	1.23
30	1.38	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.36	1.35	1.34	1.32	1.30	1.29	1.28	1.27	1.26	1.24	1.23
40	1.36	1.44	1.42	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.31	1.30	1.28	1.26	1.25	1.24	1.22	1.21	1.19
60	1.35	1.42	1.41	1.38	1.37	1.35	1.33	1.32	1.31	1.30	1.29	1.27	1.25	1.24	1.22	1.21	1.19	1.17	1.15
120	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28	1.26	1.24	1.22	1.21	1.19	1.18	1.16	1.13	1.10
∞	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25	1.24	1.22	1.19	1.18	1.16	1.14	1.12	1.08	1.00

(*a*)

*The entries in this table are values of the F variate having ν_1 numerator and ν_2 denominator degrees of freedom corresponding to the designated cumulative probability.

TABLE A5(*b*) *F* Cumulative Probabilities: 0.90 (Upper Tail: 0.10)*

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.75
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

(*b*)

* The entries in this table are values of the F variate having ν_1 numerator and ν_2 denominator degrees of freedom corresponding to the designated cumulative probability.

TABLE A5(*c*) *F* Cumulative Probabilities: 0.95 (Upper Tail: 0.05)*

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

(*c*)

* The entries in this table are values of the F variate having ν_1 numerator and ν_2 denominator degrees of freedom corresponding to the designated cumulative probability.

TABLE A5(*d*) *F* Cumulative Probabilities: 0.975 (Upper Tail: 0.025)*

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

(*d*)

* The entries in this table are values of the F variate having ν_1 numerator and ν_2 denominator degrees of freedom corresponding to the designated cumulative probability.

TABLE A5(*e*) *F* Cumulative Probabilities: 0.99 (Upper Tail: 0.01)*

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999.5	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

(*e*)

* The entries in this table are values of the F variate having ν_1 numerator and ν_2 denominator degrees of freedom corresponding to the designated cumulative probability.

TABLE A5(f) F Cumulative Probabilities: 0.995 (Upper Tail: 0.005)*

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	16211	20000	21615	22500	23056	23437	23715	23925	24091	24224	24426	24630	24836	24940	25044	25148	25253	25359	25465
2	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4	199.4	199.4	199.4	199.4	199.5	199.5	199.5	199.5	199.5	199.5
3	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43.69	43.39	43.08	42.78	42.62	42.47	42.31	42.15	41.99	41.83
4	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97	20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.32
5	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	12.14
6	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.88
7	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.65	7.53	7.42	7.31	7.19	7.08
8	14.59	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	5.95
9	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	5.19
10	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75	4.64
11	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.24	5.05	4.86	4.76	4.65	4.55	4.44	4.34	4.23
12	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01	3.90
13	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.64	4.46	4.27	4.17	4.07	3.97	3.87	3.76	3.65
14	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.43	4.25	4.06	3.96	3.86	3.76	3.66	3.55	3.44
15	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25	4.07	3.88	3.79	3.69	3.58	3.48	3.37	3.26
16	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	4.10	3.92	3.73	3.64	3.54	3.44	3.33	3.22	3.11
17	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	3.97	3.79	3.61	3.51	3.41	3.31	3.21	3.10	2.98
18	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.86	3.68	3.50	3.40	3.30	3.20	3.10	2.99	2.87
19	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.76	3.59	3.40	3.31	3.21	3.11	3.00	2.89	2.78
20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.69
21	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	3.60	3.43	3.24	3.15	3.05	2.95	2.84	2.73	2.61
22	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70	3.54	3.36	3.18	3.08	2.98	2.88	2.77	2.66	2.55
23	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	3.47	3.30	3.12	3.02	2.92	2.82	2.71	2.60	2.48
24	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	2.43
25	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	3.37	3.20	3.01	2.92	2.82	2.72	2.61	2.50	2.38
26	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49	3.33	3.15	2.97	2.87	2.77	2.67	2.56	2.45	2.33
27	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45	3.28	3.11	2.93	2.83	2.73	2.63	2.52	2.41	2.29
28	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41	3.25	3.07	2.89	2.79	2.69	2.59	2.48	2.37	2.25
29	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38	3.21	3.04	2.86	2.76	2.66	2.56	2.45	2.33	2.21
30	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18	3.01	2.82	2.73	2.63	2.52	2.42	2.30	2.18
40	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	2.95	2.78	2.60	2.50	2.40	2.30	2.18	2.06	1.93
60	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.74	2.57	2.39	2.29	2.19	2.08	1.96	1.83	1.69
120	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.54	2.37	2.19	2.09	1.98	1.87	1.75	1.61	1.43
∞	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.36	2.19	2.00	1.90	1.79	1.67	1.53	1.36	1.00

(f)

* The entries in this table are values of the F variate having ν_1 numerator and ν_2 denominator degrees of freedom corresponding to the designated cumulative probability.

Source: E. S. Pearson and H. O. Hartley, eds., *Biometrika Table for Statisticians*, Volume I, (1966). Copyright Biometrika Trustees. Reprinted with permission.

TABLE A6 Factors for Determining One-sided Tolerance Limits

	$\gamma = 0.90$			$\gamma = 0.95$			$\gamma = 0.99$		
	p			p			p		
n	0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
2	10.253	13.090	18.500	20.581	26.260	37.094	103.029	131.426	185.617
3	4.258	5.311	7.340	6.155	7.656	10.553	13.995	17.370	23.896
4	3.188	3.957	5.438	4.162	5.144	7.042	7.380	9.083	12.387
5	2.742	3.400	4.666	3.407	4.203	5.741	5.362	6.578	8.939
10	2.066	2.568	3.532	2.355	2.911	3.981	3.048	3.738	5.074
15	1.867	2.329	3.212	2.068	2.566	3.520	2.521	3.102	4.222
20	1.765	2.208	3.052	1.926	2.396	3.295	2.276	2.808	3.832
25	1.702	2.132	2.952	1.838	2.292	3.158	2.129	2.633	3.601
30	1.657	2.080	2.884	1.777	2.220	3.064	2.030	2.515	3.447
35	1.624	2.041	2.833	1.732	2.167	2.995	1.957	2.430	3.334
40	1.598	2.010	2.793	1.697	2.125	2.941	1.902	2.364	3.249
45	1.577	1.986	2.761	1.669	2.092	2.898	1.857	2.312	3.180
50	1.559	1.965	2.735	1.646	2.065	2.862	1.821	2.269	3.125
60	1.532	1.933	2.694	1.609	2.022	2.807	1.764	2.202	3.038
70	1.511	1.909	2.662	1.581	1.990	2.765	1.722	2.153	2.974
80	1.495	1.890	2.638	1.559	1.964	2.733	1.688	2.114	2.924
90	1.481	1.874	2.618	1.542	1.944	2.706	1.661	2.082	2.883
100	1.470	1.861	2.601	1.527	1.927	2.684	1.639	2.056	2.850
150	1.433	1.818	2.546	1.478	1.870	2.611	1.566	1.971	2.740
200	1.411	1.793	2.514	1.450	1.837	2.570	1.524	1.923	2.679
250	1.397	1.777	2.493	1.431	1.815	2.542	1.496	1.891	2.638
300	1.386	1.765	2.477	1.417	1.800	2.522	1.475	1.868	2.608
350	1.378	1.755	2.466	1.406	1.787	2.506	1.461	1.850	2.585
400	1.372	1.748	2.456	1.398	1.778	2.494	1.448	1.836	2.567
450	1.366	1.742	2.448	1.391	1.770	2.484	1.438	1.824	2.553
500	1.362	1.736	2.442	1.385	1.763	2.475	1.430	1.814	2.540
550	1.358	1.732	2.436	1.380	1.757	2.468	1.422	1.806	2.530
600	1.355	1.728	2.431	1.376	1.752	2.462	1.416	1.799	2.520
650	1.352	1.725	2.427	1.372	1.748	2.456	1.411	1.792	2.512
700	1.349	1.722	2.423	1.368	1.744	2.451	1.406	1.787	2.505
750	1.347	1.719	2.420	1.365	1.741	2.447	1.401	1.782	2.499
800	1.344	1.717	2.417	1.363	1.737	2.443	1.397	1.777	2.493
850	1.343	1.714	2.414	1.360	1.734	2.439	1.394	1.773	2.488
900	1.341	1.712	2.411	1.358	1.732	2.436	1.390	1.769	2.483
950	1.339	1.711	2.409	1.356	1.729	2.433	1.387	1.766	2.479
1000	1.338	1.709	2.407	1.354	1.727	2.430	1.385	1.762	2.475
∞	1.282	1.645	2.326	1.282	1.645	2.326	1.282	1.645	2.326

Source: Adapted and reprinted from Odeh, R. E. and Owen, D. B. (1980). *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*. New York: Marcel Dekker, Inc., pp. 22–25 and 30–37, by courtesy of Marcel Dekker, Inc.

TABLE A7 Factors for Determining Two-sided Tolerance Limits

	$\gamma = 0.90$			$\gamma = 0.95$			$\gamma = 0.99$		
	p			p			p		
n	0.900	0.950	0.990	0.900	0.950	0.990	0.900	0.950	0.990
2	15.512	18.221	23.423	31.092	36.519	46.944	155.569	182.720	234.877
3	5.788	6.823	8.819	8.306	9.789	12.647	18.782	22.131	28.586
4	4.157	4.913	6.372	5.368	6.341	8.221	9.416	11.118	14.405
5	3.499	4.142	5.387	4.291	5.077	6.598	6.655	7.870	10.220
10	2.546	3.026	3.958	2.856	3.393	4.437	3.617	4.294	5.610
15	2.285	2.720	3.565	2.492	2.965	3.885	2.967	3.529	4.621
20	2.158	2.570	3.372	2.319	2.760	3.621	2.675	3.184	4.175
25	2.081	2.479	3.254	2.215	2.638	3.462	2.506	2.984	3.915
30	2.029	2.417	3.173	2.145	2.555	3.355	2.394	2.851	3.742
35	1.991	2.371	3.114	2.094	2.495	3.276	2.314	2.756	3.618
40	1.961	2.336	3.069	2.055	2.448	3.216	2.253	2.684	3.524
45	1.938	2.308	3.032	2.024	2.412	3.168	2.205	2.627	3.450
50	1.918	2.285	3.003	1.999	2.382	3.129	2.166	2.580	3.390
60	1.888	2.250	2.956	1.960	2.335	3.068	2.106	2.509	3.297
70	1.866	2.224	2.922	1.931	2.300	3.023	2.062	2.457	3.228
80	1.849	2.203	2.895	1.908	2.274	2.988	2.028	2.416	3.175
90	1.835	2.186	2.873	1.890	2.252	2.959	2.001	2.384	3.133
100	1.823	2.172	2.855	1.875	2.234	2.936	1.978	2.357	3.098
150	1.786	2.128	2.796	1.826	2.176	2.859	1.905	2.271	2.985
200	1.764	2.102	2.763	1.798	2.143	2.816	1.866	2.223	2.921
250	1.750	2.085	2.741	1.780	2.121	2.788	1.839	2.191	2.880
300	1.740	2.073	2.725	1.767	2.106	2.767	1.820	2.169	2.850
350	1.732	2.064	2.713	1.757	2.094	2.752	1.806	2.152	2.828
400	1.726	2.057	2.703	1.749	2.084	2.739	1.794	2.138	2.810
450	1.721	2.051	2.695	1.743	2.077	2.729	1.785	2.127	2.795
500	1.717	2.046	2.689	1.737	2.070	2.721	1.777	2.117	2.783
550	1.713	2.041	2.683	1.733	2.065	2.713	1.770	2.109	2.772
600	1.710	2.038	2.678	1.729	2.060	2.707	1.765	2.103	2.763
650	1.707	2.034	2.674	1.725	2.056	2.702	1.759	2.097	2.755
700	1.705	2.032	2.670	1.722	2.052	2.697	1.755	2.091	2.748
750	1.703	2.029	2.667	1.719	2.049	2.692	1.751	2.086	2.742
800	1.701	2.027	2.664	1.717	2.046	2.688	1.747	2.082	2.736
850	1.699	2.025	2.661	1.715	2.043	2.685	1.744	2.078	2.731
900	1.697	2.023	2.658	1.712	2.040	2.682	1.741	2.075	2.727
950	1.696	2.021	2.656	1.711	2.038	2.679	1.738	2.071	2.722
1000	1.695	2.019	2.654	1.709	2.036	2.676	1.736	2.068	2.718
∞	1.645	1.960	2.576	1.645	1.960	2.576	1.645	1.960	2.576

Source: Adapted and reprinted from Odeh, R. E. and Owen, D. B. (1980). *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*. New York: Marcel Dekker, Inc., pp. 90–93 and 98–105, by courtesy of Marcel Dekker, Inc.

TABLE A8 Upper-Tail Critical Values for the F-Max Test

		Critical Value									
ν	α	$k = 3$	4	5	6	7	8	9	10	11	12
2	0.10	42.48	69.13	98.18	129.1	161.7	195.6	230.7	266.8	303.9	341.9
	0.05	87.49	142.5	202.4	266.2	333.2	403.1	475.4	549.8	626.2	704.4
	0.01	447.5	729.2	1036	1362	1705	2063	2432	2813	3204	3604
3	0.10	16.77	23.95	30.92	37.73	44.40	50.94	57.38	63.72	69.97	76.14
	0.05	27.76	39.51	50.88	61.98	72.83	83.48	93.94	104.2	114.4	124.4
	0.01	84.56	119.8	153.8	187.0	219.3	251.1	282.3	313.0	343.2	373.1
4	0.10	10.38	13.88	17.08	20.06	22.88	25.57	28.14	30.62	33.01	35.33
	0.05	15.46	20.56	25.21	29.54	33.63	37.52	41.24	44.81	48.27	51.61
	0.01	36.70	48.43	59.09	69.00	78.33	87.20	95.68	103.8	111.7	119.3
5	0.10	7.68	9.86	11.79	13.54	15.15	16.66	18.08	19.43	20.71	21.95
	0.05	10.75	13.72	16.34	18.70	20.88	22.91	24.83	26.65	28.38	30.03
	0.01	22.06	27.90	33.00	37.61	41.85	45.81	49.53	53.06	56.42	59.63
6	0.10	6.23	7.78	9.11	10.30	11.38	12.38	13.31	14.18	15.01	15.79
	0.05	8.36	10.38	12.11	13.64	15.04	16.32	17.51	18.64	19.70	20.70
	0.01	15.60	19.16	22.19	24.89	27.32	29.57	31.65	33.61	35.46	37.22
7	0.10	5.32	6.52	7.52	8.41	9.20	9.93	10.60	11.23	11.82	12.37
	0.05	6.94	8.44	9.70	10.80	11.80	12.70	13.54	14.31	15.05	15.74
	0.01	12.09	14.55	16.60	18.39	20.00	21.47	22.82	24.08	25.26	26.37
8	0.10	4.71	5.68	6.48	7.18	7.80	8.36	8.88	9.36	9.81	10.23
	0.05	6.00	7.19	8.17	9.02	9.77	10.46	11.08	11.67	12.21	12.72
	0.01	9.94	11.77	13.27	14.58	15.73	16.78	17.74	18.63	19.46	20.24

TABLE A8 Upper-Tail Critical Values for the F-Max Test

		Critical Value									
ν	α	$k = 3$	4	5	6	7	8	9	10	11	12
9	0.10	4.26	5.07	5.74	6.31	6.82	7.28	7.70	8.09	8.45	8.78
	0.05	5.34	6.31	7.11	7.79	8.40	8.94	9.44	9.90	10.33	10.73
	0.01	8.49	9.93	11.10	12.11	12.99	13.79	14.52	15.19	15.81	16.39
10	0.10	3.93	4.63	5.19	5.68	6.11	6.49	6.84	7.16	7.46	7.74
	0.05	4.85	5.67	6.34	6.91	7.41	7.86	8.27	8.64	8.99	9.32
	0.01	7.46	8.64	9.59	10.39	11.10	11.74	12.31	12.84	13.33	13.79
12	0.10	3.45	4.00	4.44	4.81	5.13	5.42	5.68	5.92	6.14	6.35
	0.05	4.16	4.79	5.30	5.72	6.09	6.42	6.72	6.99	7.24	7.48
	0.01	6.10	6.95	7.63	8.20	8.69	9.13	9.53	9.89	10.23	10.54
15	0.10	3.00	3.41	3.74	4.02	4.25	4.46	4.65	4.82	4.98	5.13
	0.05	3.53	4.00	4.37	4.67	4.94	5.17	5.38	5.57	5.75	5.91
	0.01	4.93	5.52	5.99	6.37	6.71	7.00	7.27	7.51	7.73	7.93
20	0.10	2.57	2.87	3.10	3.29	3.46	3.60	3.73	3.85	3.96	4.06
	0.05	2.95	3.28	3.53	3.74	3.92	4.08	4.22	4.35	4.46	4.57
	0.01	3.90	4.29	4.60	4.85	5.06	5.25	5.42	5.57	5.70	5.83
30	0.10	2.14	2.34	2.50	2.62	2.73	2.82	2.90	2.97	3.04	3.10
	0.05	2.40	2.61	2.77	2.90	3.01	3.11	3.19	3.27	3.34	3.40
	0.01	2.99	3.23	3.41	3.56	3.68	3.79	3.88	3.97	4.04	4.12
60	0.10	1.71	1.82	1.90	1.96	2.02	2.07	2.11	2.14	2.18	2.21
	0.05	1.84	1.96	2.04	2.11	2.16	2.21	2.25	2.29	2.32	2.35
	0.01	2.15	2.26	2.35	2.42	2.47	2.52	2.57	2.61	2.64	2.67

Source: Nelson, L. (1987). "Upper 10%, 5%, and 1% Points of the Maximum F-Ratio," *Journal of Quality Technology*, **19**, 165–67. Copyright American Society for Quality Control, Inc., Milwaukee, WI. Reprinted by permission.

TABLE A9 Orthogonal Polynomial Coefficients

[Table not available in this electronic edition.]

Note: For $k \geqslant 9$, the remainder of the coefficients for column a_i are given by $a_{i,k-j+1} = -a_{i,j}$ for i odd, and $a_{i,k-j+1} = a_{i,j}$ for i even.

[Table not available in this electronic edition.]

TABLE A9 Orthogonal Polynomial Coefficients

TABLE A9 Orthogonal Polynomial Coefficients

[Table not available in this electronic edition.]

TABLE A9 Orthogonal Polynomial Coefficients

[Table not available in this electronic edition.]

Source: Taken from Table XXIII of Fisher, R. A. and Yates, F. (1974) *Statistical Tables for Biological, Agricultural, and Medical Research*, 6th ed. Reprinted by permission of Pearson Education Limited.

Note: The two values at the bottom of each column are the values of D, the sum of squares of the coefficients used in the orthogonal polynomial, and λ, the coefficient of the highest power of X_i in the orthogonal polynomial.

TABLE A10 Critical Values for Outlier Test Using L_k and S_k

		Critical Value $\times 10^3$																							
k	α	$n = 50$	45	40	35	30	25	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3
1	0.01	760	745	722	690	650	607	539	522	504	485	463	440	414	386	355	321	283	241	195	145	93	44	10	
	0.025	796	776	756	732	699	654	594	579	562	544	525	503	479	453	423	390	353	310	262	207	145	81	25	1
	0.05	820	802	784	762	730	692	638	624	610	593	576	556	534	510	482	451	415	374	326	270	203	127	49	3
	0.10	840	826	812	792	766	732	685	673	660	646	631	613	594	573	548	520	488	450	405	350	283	199	98	11
2	0.01	667	641	610	573	527	468	391	373	353	332	310	286	261	233	204	174	141	108	75	44	19	4		
	0.025	697	667	644	610	567	512	439	421	403	382	360	337	311	284	254	221	186	149	110	71	35	9		
	0.05	720	698	673	641	601	550	480	464	446	426	405	382	357	330	300	267	230	191	148	102	56	18	1	
	0.10	746	726	702	674	637	591	527	511	494	476	456	435	411	384	355	323	286	245	199	148	92	38	3	
3	0.01	592	558	522	484	434	377	300	272	260	237	219	194	172	147	120	98	70	48	28	10	2			
	0.025	622	592	561	527	479	417	341	321	299	282	261	239	214	184	162	129	100	73	45	21	5			
	0.05	646	618	588	554	506	450	377	354	337	322	300	276	250	224	196	162	129	99	64	32	10			
	0.10	673	648	622	586	523	489	420	398	384	364	342	322	298	270	240	208	170	134	95	56	20			
4	0.01	531	498	460	418	369	308	231	211	192	171	151	132	113	94	70	52	32	18	8					
	0.025	559	529	491	455	408	342	265	243	226	208	185	167	145	122	96	74	52	30	13					
	0.05	588	556	523	482	434	374	299	277	259	240	219	197	174	150	125	98	70	45	22					
	0.10	614	586	554	516	472	412	339	316	302	282	260	236	212	186	159	128	98	68	38					
5	0.01	483	444	408	364	312	246	175	154	140	126	108	90	72	56	38	26	12							
	0.025	510	473	433	398	352	282	209	189	171	151	135	113	95	77	57	40	23							
	0.05	535	502	468	424	376	312	238	217	200	181	159	140	122	98	76	54	34							
	0.10	562	533	499	458	411	350	273	251	236	216	194	172	150	126	103	74	51							
6	0.01	438	399	364	321	268	204	136	118	104	91	72	57	46	33	19									
	0.025	466	430	387	384	302	233	165	145	129	117	96	78	63	47	31									
	0.05	490	456	421	376	327	262	188	168	154	136	115	97	79	60	42									
	0.10	518	488	451	410	359	296	220	199	184	165	144	124	104	82	62									

TABLE A10 Critical Values for Outlier Test Using L_k and S_k

		Critical Value $\times 10^3$																						
k	α	$n = 50$	45	40	35	30	25	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3
7	0.01	400	361	324	282	229	168	104	88	76	64	49	37	27											
	0.025	428	391	348	308	261	192	128	108	95	82	65	51	38											
	0.05	450	417	378	334	283	222	150	130	116	100	82	66	50											
	0.10	477	447	408	365	316	251	176	158	142	125	104	86	68											
8	0.01	368	328	292	250	196	144	78	64	53	44	30													
	0.025	392	356	314	274	226	159	98	80	68	58	45													
	0.05	414	382	342	297	245	184	115	99	86	72	55													
	0.10	442	410	372	328	276	213	140	124	108	92	73													
9	0.01	336	296	262	220	166	112	58	46	36															
	0.025	363	325	283	242	193	132	73	59	48															
	0.05	383	350	310	264	212	154	88	74	62															
	0.10	410	378	338	294	240	180	110	94	80															
10	0.01	308	270	234	194	142	92	42																	
	0.025	334	295	257	213	165	108	54																	
	0.05	356	320	280	235	183	126	66																	
	0.10	380	348	307	262	210	152	85																	

Source: Adapted from Tietjen, G. L. and Moore, R. H. (1972). "Some Grubbs-Type Statistics for the Detection of Several Outliers," *Technometrics*, **14**, 583–598. Copyright American Statistical Association, Alexandria, VA. Reprinted by permission.

Adapted from Grubbs, F. E. (1950). "Sample Criteria for Testing Outlying Observations," *Annals of Mathematical Statistics*, **21**, 27–58. Used by permission of the Institute of Mathematical Statistics.

Adapted from Grubbs, F. E. and Beck, G. (1972). "Extension of Sample Sizes and Percentage Points for Significance Tests of Outlying Observations," *Technometrics*, **14**, 847–854. Copyright American Statistical Association, Alexandria, VA. Reprinted by permission.

TABLE A11 Critical Values for Outlier Test Using E_k

		Critical Value $\times 10^3$																							
k	α	$n = 50$	45	40	35	30	25	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3
1	0.01	748	728	704	669	624	571	499	484	459	440	422	404	374	337	311	274	235	197	156	110	68	29	4	
	0.05	796	776	756	732	698	654	594	579	562	544	525	503	479	453	423	390	353	310	262	207	145	81	25	1
	0.10	820	802	784	762	730	692	638	624	610	593	576	556	534	510	482	451	415	374	326	270	203	127	49	3
2	0.01	636	607	574	533	482	418	339	323	306	290	263	238	207	181	159	134	101	78	50	28	12	2		
	0.05	684	658	629	596	549	493	416	398	382	362	340	317	293	262	234	204	172	137	99	65	34	10	1	
	0.10	708	684	657	624	582	528	460	442	424	406	384	360	337	309	278	250	214	175	137	94	56	22	2	
3	0.01	550	518	480	435	386	320	236	219	206	188	166	146	123	103	83	64	44	26	14	6	1			
	0.05	599	567	534	495	443	381	302	287	267	248	227	206	179	156	133	107	83	57	34	16	4			
	0.10	622	593	562	523	475	417	338	322	304	284	263	240	216	189	162	138	108	80	53	27	9			
4	0.01	482	446	408	364	308	245	170	156	141	122	107	90	72	56	42	30	18	9	4					
	0.05	529	492	458	417	364	298	221	203	187	170	153	134	112	92	73	55	37	21	10					
	0.10	552	522	486	443	391	331	252	234	217	198	182	160	138	116	94	73	52	32	16					
5	0.01	424	386	347	299	250	188	121	108	94	79	68	54	42	31	20	12	6							
	0.05	468	433	395	351	298	236	163	146	132	116	102	84	68	53	39	26	14							
	0.10	492	459	422	379	325	264	188	172	156	140	122	105	86	68	52	36	22							
6	0.01	376	336	298	252	204	146	86	74	62	52	40	32	22	14	8									
	0.05	417	381	343	298	246	186	119	105	91	78	67	52	39	28	18									
	0.10	440	406	367	324	270	210	138	124	110	95	82	67	52	38	26									

TABLE A11 Critical Values for Outlier Test Using E_k

		Critical Value $\times 10^3$																						
k	α	$n = 50$	45	40	35	30	25	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3
7	0.01	334	294	258	211	166	110	58	50	41	32	24	18	12											
	0.05	373	337	297	254	203	146	85	74	62	50	41	30	21											
	0.10	396	360	320	276	224	168	102	89	76	64	53	40	29											
8	0.01	297	258	220	177	132	87	40	32	26	18	14													
	0.05	334	299	259	214	166	114	59	50	41	32	24													
	0.10	355	320	278	236	186	132	72	62	51	42	32													
9	0.01	264	228	190	149	108	66	26	20	14															
	0.05	299	263	223	181	137	89	41	33	26															
	0.10	319	284	243	202	154	103	51	42	34															
10	0.01	235	200	164	124	87	50	17																	
	0.05	268	233	195	154	112	68	28																	
	0.10	287	252	212	172	126	80	35																	

Source: Adapted from Tietjen, G. L. and Moore, R. H. (1972). "Some Grubbs-Type Statistics for the Detection of Several Outliers," *Technometrics*, **14**, 583–598. Copyright American Statistical Association, Alexandria, VA. Reprinted by permission.

Adapted from Grubbs, F. E. (1950). "Sample Criteria for Testing Outlying Observations," *Annals of Mathematical Statistics*, **21**, 27–58. Used by permission of the Institute of Mathematical Statistics.

TABLE A12 Coefficients Used in the Shapiro–Wilk Test for Normality*

	a_{n-i+1}											
i	$n = 3$	4	5	6	7	8	9	10	11	12	13	14
1	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739	0.5601	0.5475	0.5359	0.5251
2		0.1677	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291	0.3315	0.3325	0.3325	0.3318
3				0.0875	0.1401	0.1743	0.1976	0.2141	0.2260	0.2347	0.2412	0.2460
4						0.0561	0.0947	0.1224	0.1429	0.1586	0.1707	0.1802
5								0.0399	0.0695	0.0922	0.1099	0.1240
6										0.0303	0.0539	0.0727
7												0.0240
	$n = 15$	16	17	18	19	20	21	22	23	24	25	26
1	0.5150	0.5056	0.4968	0.4886	0.4808	0.4734	0.4643	0.4590	0.4542	0.4493	0.4450	0.4407
2	0.3306	0.3290	0.3273	0.3253	0.3232	0.3211	0.3185	0.3156	0.3126	0.3098	0.3069	0.3043
3	0.2495	0.2521	0.2540	0.2553	0.2561	0.2565	0.2578	0.2571	0.2563	0.2554	0.2543	0.2533
4	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085	0.2119	0.2131	0.2139	0.2145	0.2148	0.2151
5	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686	0.1736	0.1764	0.1787	0.1807	0.1822	0.1836
6	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334	0.1399	0.1443	0.1480	0.1512	0.1539	0.1563
7	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013	0.1092	0.1150	0.1201	0.1245	0.1283	0.1316
8		0.0196	0.0359	0.0496	0.0612	0.0711	0.0804	0.0878	0.0941	0.0997	0.1046	0.1089
9				0.0163	0.0303	0.0422	0.0530	0.0618	0.0696	0.0764	0.0823	0.0876
10						0.0140	0.0263	0.0368	0.0459	0.0539	0.0610	0.0672
11								0.0122	0.0228	0.0321	0.0403	0.0476
12										0.0107	0.0200	0.0284
13												0.0094

TABLE A12 Coefficients Used in the Shapiro–Wilk Test for Normality*

	$n = 27$	28	29	30	31	32	33	34	35	36	37	38
1	0.4366	0.4328	0.4291	0.4254	0.4220	0.4188	0.4156	0.4127	0.4096	0.4068	0.4040	0.4015
2	0.3018	0.2992	0.2968	0.2944	0.2921	0.2898	0.2876	0.2854	0.2834	0.2813	0.2794	0.2774
3	0.2522	0.2510	0.2499	0.2487	0.2475	0.2463	0.2451	0.2439	0.2427	0.2415	0.2403	0.2391
4	0.2152	0.2151	0.2150	0.2148	0.2145	0.2141	0.2137	0.2132	0.2127	0.2121	0.2116	0.2110
5	0.1848	0.1857	0.1864	0.1870	0.1874	0.1878	0.1880	0.1882	0.1883	0.1883	0.1883	0.1881
6	0.1584	0.1601	0.1616	0.1630	0.1641	0.1651	0.1660	0.1667	0.1673	0.1678	0.1683	0.1686
7	0.1346	0.1372	0.1395	0.1415	0.1433	0.1449	0.1463	0.1475	0.1487	0.1496	0.1505	0.1513
8	0.1128	0.1162	0.1192	0.1219	0.1243	0.1265	0.1284	0.1301	0.1317	0.1331	0.1344	0.1356
9	0.0923	0.0965	0.1002	0.1036	0.1066	0.1093	0.1118	0.1140	0.1160	0.1179	0.1196	0.1211
10	0.0728	0.0778	0.0822	0.0862	0.0899	0.0931	0.0961	0.0988	0.1013	0.1036	0.1056	0.1075
11	0.0540	0.0598	0.0650	0.0697	0.0739	0.0777	0.0812	0.0844	0.0873	0.0900	0.0924	0.0947
12	0.0358	0.0424	0.0483	0.0537	0.0585	0.0629	0.0669	0.0706	0.0739	0.0770	0.0798	0.0824
13	0.0178	0.0253	0.0320	0.0381	0.0435	0.0485	0.0530	0.0572	0.0610	0.0645	0.0677	0.0706
14		0.0084	0.0159	0.0227	0.0289	0.0344	0.0395	0.0441	0.0484	0.0523	0.0559	0.0592
15				0.0076	0.0144	0.0206	0.0262	0.0314	0.0361	0.0404	0.0444	0.0481
16						0.0068	0.0131	0.0187	0.0239	0.0287	0.0331	0.0372
17								0.0062	0.0119	0.0172	0.0220	0.0264
18										0.0057	0.0110	0.0158
19												0.0053

TABLE A12 Coefficients Used in the Shapiro–Wilk Test for Normality*

	$n = 39$	40	41	42	43	44	45	46	47	48	49	50
1	0.3989	0.3964	0.3940	0.3917	0.3894	0.3872	0.3850	0.3830	0.3808	0.3789	0.3770	0.3751
2	0.2755	0.2737	0.2719	0.2701	0.2684	0.2667	0.2651	0.2635	0.2620	0.2604	0.2589	0.2574
3	0.2380	0.2368	0.2357	0.2345	0.2334	0.2323	0.2313	0.2302	0.2291	0.2281	0.2271	0.2260
4	0.2104	0.2098	0.2091	0.2085	0.2078	0.2072	0.2065	0.2058	0.2052	0.2045	0.2038	0.2032
5	0.1880	0.1878	0.1876	0.1874	0.1871	0.1868	0.1865	0.1862	0.1859	0.1855	0.1851	0.1847
6	0.1689	0.1691	0.1693	0.1694	0.1695	0.1695	0.1695	0.1695	0.1695	0.1693	0.1692	0.1691
7	0.1520	0.1526	0.1531	0.1535	0.1539	0.1542	0.1545	0.1548	0.1550	0.1551	0.1553	0.1554
8	0.1366	0.1376	0.1384	0.1392	0.1398	0.1405	0.1410	0.1415	0.1420	0.1423	0.1427	0.1430
9	0.1225	0.1237	0.1249	0.1259	0.1269	0.1278	0.1286	0.1293	0.1300	0.1306	0.1312	0.1317
10	0.1092	0.1108	0.1123	0.1136	0.1149	0.1160	0.1170	0.1180	0.1189	0.1197	0.1205	0.1212
11	0.0967	0.0986	0.1004	0.1020	0.1035	0.1049	0.1062	0.1073	0.1085	0.1095	0.1105	0.1113
12	0.0848	0.0870	0.0891	0.0909	0.0927	0.0943	0.0959	0.0972	0.0986	0.0998	0.1010	0.1020
13	0.0733	0.0759	0.0782	0.0804	0.0824	0.0842	0.0860	0.0876	0.0892	0.0906	0.0919	0.0932
14	0.0622	0.0651	0.0677	0.0701	0.0724	0.0745	0.0765	0.0783	0.0801	0.0817	0.0832	0.0846
15	0.0515	0.0546	0.0575	0.0602	0.0628	0.0651	0.0673	0.0694	0.0713	0.0731	0.0748	0.0764
16	0.0409	0.0444	0.0476	0.0506	0.0534	0.0560	0.0584	0.0607	0.0628	0.0648	0.0667	0.0685
17	0.0305	0.0343	0.0379	0.0411	0.0442	0.0471	0.0497	0.0522	0.0546	0.0568	0.0588	0.0608
18	0.0203	0.0244	0.0283	0.0318	0.0352	0.0383	0.0412	0.0439	0.0465	0.0489	0.0511	0.0532
19	0.0101	0.0146	0.0188	0.0227	0.0263	0.0296	0.0328	0.0357	0.0385	0.0411	0.0436	0.0459
20		0.0049	0.0094	0.0136	0.0175	0.0211	0.0245	0.0277	0.0307	0.0335	0.0361	0.0386
21				0.0045	0.0087	0.0126	0.0163	0.0197	0.0229	0.0259	0.0288	0.0314
22						0.0042	0.0081	0.0118	0.0153	0.0185	0.0215	0.0244
23								0.0039	0.0076	0.0111	0.0143	0.0174
24										0.0037	0.0071	0.0104
25												0.0035

*$a_i = -a_{n-i+1}$ for $i = 1, 2, \ldots, k$ where $k = n/2$ if n is even and $k = (n-1)/2$ if n is odd.

Source: Shapiro, S. S. and Wilk, M. B. (1965). "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, **52**, 591–611. Copyright Biometrika Trustees. Reprinted with permission.

TABLE A13 Critical Values for the Shapiro–Wilk Test for Normality

	Critical Value				
n	$\alpha = 1\%$	2%	5%	10%	50%
3	0.753	0.756	0.767	0.789	0.959
4	0.687	0.707	0.748	0.792	0.935
5	0.686	0.715	0.762	0.806	0.927
6	0.713	0.743	0.788	0.826	0.927
7	0.730	0.760	0.803	0.838	0.928
8	0.749	0.778	0.818	0.851	0.932
9	0.764	0.791	0.829	0.859	0.935
10	0.781	0.806	0.842	0.869	0.938
11	0.792	0.817	0.850	0.876	0.940
12	0.805	0.828	0.859	0.883	0.943
13	0.814	0.837	0.866	0.889	0.945
14	0.825	0.846	0.874	0.895	0.947
15	0.835	0.855	0.881	0.901	0.950
16	0.844	0.863	0.887	0.906	0.952
17	0.851	0.869	0.892	0.910	0.954
18	0.858	0.874	0.897	0.914	0.956
19	0.863	0.879	0.901	0.917	0.957
20	0.868	0.884	0.905	0.920	0.959
21	0.873	0.888	0.908	0.923	0.960
22	0.878	0.892	0.911	0.926	0.961
23	0.881	0.895	0.914	0.928	0.962
24	0.884	0.898	0.916	0.930	0.963
25	0.888	0.901	0.918	0.931	0.964

TABLE A13 Critical Values for the Shapiro–Wilk Test for Normality

	Critical Value				
n	$\alpha = 1\%$	2%	5%	10%	50%
26	0.891	0.904	0.920	0.933	0.965
27	0.894	0.906	0.923	0.935	0.965
28	0.896	0.908	0.924	0.936	0.966
29	0.898	0.910	0.926	0.937	0.966
30	0.900	0.912	0.927	0.939	0.967
31	0.902	0.914	0.929	0.940	0.967
32	0.904	0.915	0.930	0.941	0.968
33	0.906	0.917	0.931	0.942	0.968
34	0.908	0.919	0.933	0.943	0.969
35	0.910	0.920	0.934	0.944	0.969
36	0.912	0.922	0.935	0.945	0.970
37	0.914	0.924	0.936	0.946	0.970
38	0.916	0.925	0.938	0.947	0.971
39	0.917	0.927	0.939	0.948	0.971
40	0.919	0.928	0.940	0.949	0.972
41	0.920	0.929	0.941	0.950	0.972
42	0.922	0.930	0.942	0.951	0.972
43	0.923	0.932	0.943	0.951	0.973
44	0.924	0.933	0.944	0.952	0.973
45	0.926	0.934	0.945	0.953	0.973
46	0.927	0.935	0.945	0.953	0.974
47	0.928	0.928	0.946	0.954	0.974
48	0.929	0.937	0.947	0.954	0.974
49	0.929	0.937	0.947	0.955	0.974
50	0.930	0.938	0.947	0.955	0.974

Source: Adapted from Shapiro, S. S. and Wilk, M. B. (1965), "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, **52**, 591–611. Copyright Biometrika Trustees. Reprinted with permission.

TABLE A14 Percentage Points of the Studentized Range

$\alpha = 0.05$

ν \ k	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	17.97	26.98	32.82	37.08	40.41	43.12	45.40	47.36	49.07	50.59	51.96	53.20	54.33	55.36	56.32	57.22	58.04	58.83
2	6.085	8.331	9.798	10.88	11.74	12.44	13.03	13.54	13.99	14.39	14.75	15.08	15.38	15.65	15.91	16.14	16.37	16.57
3	4.501	5.910	6.825	7.502	8.037	8.478	8.853	9.177	9.462	9.717	9.946	10.15	10.35	10.53	10.69	10.84	10.98	11.11
4	3.927	5.040	5.757	6.287	6.707	7.053	7.347	7.602	7.826	8.027	8.208	8.373	8.525	8.664	8.794	8.914	9.028	9.134
5	3.635	4.602	5.218	5.673	6.033	6.330	6.582	6.802	6.995	7.168	7.324	7.466	7.596	7.717	7.828	7.932	8.030	8.122
6	3.461	4.339	4.896	5.305	5.628	5.895	6.122	6.319	6.493	6.649	6.789	6.917	7.034	7.143	7.244	7.338	7.426	7.508
7	3.344	4.165	4.681	5.060	5.359	5.606	5.815	5.998	6.158	6.302	6.431	6.550	6.658	6.759	6.852	6.939	7.020	7.097
8	3.261	4.041	4.529	4.886	5.167	5.399	5.597	5.767	5.918	6.054	6.175	6.287	6.389	6.483	6.571	6.653	6.729	6.802
9	3.199	3.949	4.415	4.756	5.024	5.244	5.432	5.595	5.739	5.867	5.983	6.089	6.186	6.276	6.359	6.437	6.510	6.579
10	3.151	3.877	4.327	4.654	4.912	5.124	5.305	5.461	5.599	5.722	5.833	5.935	6.028	6.114	6.194	6.269	6.339	6.405
11	3.113	3.820	4.256	4.574	4.823	5.028	5.202	5.353	5.487	5.605	5.713	5.811	5.901	5.984	6.062	6.134	6.202	6.265
12	3.082	3.773	4.199	4.508	4.751	4.950	5.119	5.265	5.395	5.511	5.615	5.710	5.798	5.878	5.953	6.023	6.089	6.151
13	3.055	3.735	4.151	4.453	4.690	4.885	5.049	5.192	5.318	5.431	5.533	5.625	5.711	5.789	5.862	5.931	5.995	6.055
14	3.033	3.702	4.111	4.407	4.639	4.829	4.990	5.131	5.254	5.364	5.463	5.554	5.637	5.714	5.786	5.852	5.915	5.974
15	3.014	3.674	4.076	4.367	4.595	4.782	4.940	5.077	5.198	5.306	5.404	5.493	5.574	5.649	5.720	5.785	5.846	5.904
16	2.998	3.649	4.046	4.333	4.557	4.741	4.897	5.031	5.150	5.256	5.352	5.439	5.520	5.593	5.662	5.727	5.786	5.843
17	2.984	3.628	4.020	4.303	4.524	4.705	4.858	4.991	5.108	5.212	5.307	5.392	5.471	5.544	5.612	5.675	5.734	5.790
18	2.971	3.609	3.997	4.277	4.495	4.673	4.824	4.956	5.071	5.174	5.267	5.352	5.429	5.501	5.568	5.630	5.688	5.743
19	2.960	3.593	3.977	4.253	4.469	4.465	4.794	4.924	5.038	5.140	5.231	5.315	5.391	5.462	5.528	5.589	5.647	5.701
20	2.950	3.578	3.958	4.232	4.445	4.620	4.768	4.896	5.008	5.108	5.199	5.282	5.357	5.427	5.493	5.553	5.610	5.663
24	2.919	3.532	3.901	4.166	4.373	4.541	4.684	4.807	4.915	5.012	5.099	5.179	5.251	5.319	5.381	5.439	5.494	5.545
30	2.888	3.486	3.845	4.102	4.302	4.464	4.602	4.720	4.824	4.917	5.001	5.077	5.147	5.211	5.271	5.327	5.379	5.429
40	2.858	3.442	3.791	4.039	4.232	4.389	4.521	4.635	4.735	4.824	4.904	4.977	5.044	5.106	5.163	5.216	5.266	5.313
60	2.829	3.399	3.737	3.977	4.163	4.314	4.441	4.550	4.646	4.732	4.808	4.878	4.942	5.001	5.056	5.107	5.154	5.199
120	2.800	3.356	3.685	3.917	4.096	4.241	4.363	4.468	4.560	4.641	4.714	4.781	4.842	4.898	4.950	4.998	5.044	5.086
∞	2.772	3.314	3.633	3.858	4.030	4.170	4.286	4.387	4.474	4.552	4.622	4.685	4.743	4.796	4.845	4.891	4.934	4.974

TABLE A14 Percentage Points of the Studentized Range

$\alpha = 0.05$

ν \ k	20	22	24	26	28	30	32	34	36	38	40	50	60	70	80	90	100
1	59.56	60.91	62.12	68.22	64.23	65.15	66.01	66.81	67.56	68.26	68.92	71.73	73.97	75.82	77.40	78.77	79.98
2	16.77	17.13	17.45	17.75	18.02	18.27	18.50	18.72	18.92	19.11	19.28	20.05	20.66	21.16	21.59	21.96	22.29
3	11.24	11.47	11.68	11.87	12.05	12.21	12.36	12.50	12.63	12.75	12.87	13.36	13.76	14.08	14.36	14.61	14.82
4	9.233	9.418	9.584	9.736	9.875	10.00	10.12	10.23	10.34	10.44	10.53	10.93	11.24	11.51	11.73	11.92	12.09
5	8.208	8.368	8.512	8.643	8.764	8.875	8.979	9.075	9.165	9.250	9.330	9.674	9.949	10.18	10.38	10.54	10.69
6	7.587	7.730	7.861	7.979	8.088	8.189	8.288	8.370	8.452	8.529	8.601	8.913	9.163	9.370	9.548	9.702	9.839
7	7.170	7.303	7.423	7.533	7.634	7.728	7.814	7.895	7.972	8.043	8.110	8.400	8.632	8.824	8.989	9.133	9.261
8	6.870	6.995	7.109	7.212	7.307	7.395	7.477	7.554	7.625	7.693	7.756	8.029	8.248	8.430	8.586	8.722	8.843
9	6.644	6.763	6.871	6.970	7.061	7.145	7.222	7.295	7.363	7.428	7.488	7.749	7.958	8.132	8.281	8.410	8.526
10	6.467	6.582	6.686	6.781	6.868	6.948	7.023	7.093	7.159	7.220	7.279	7.529	7.730	7.897	8.041	8.166	8.276
11	6.326	6.436	6.536	6.628	6.712	6.790	6.863	6.930	6.994	7.053	7.110	7.352	7.546	7.708	7.847	7.968	8.075
12	6.209	6.317	6.414	6.503	6.585	6.660	6.731	6.796	6.858	6.916	6.970	7.205	7.394	7.552	7.687	7.804	7.909
13	6.112	6.217	6.312	6.398	6.478	6.551	6.620	6.684	6.744	6.800	6.854	7.083	7.267	7.421	7.552	7.667	7.769
14	6.029	6.132	6.224	6.309	6.387	6.459	6.526	6.588	6.647	6.702	6.754	6.979	7.159	7.309	7.438	7.550	7.650
15	5.958	6.059	6.149	6.233	6.309	6.379	6.445	6.506	6.564	6.618	6.669	6.888	7.065	7.212	7.339	7.449	7.546
16	5.897	5.995	6.084	6.166	6.241	6.310	6.374	6.434	6.491	6.544	6.594	6.810	6.984	7.128	7.252	7.360	7.457
17	5.842	5.940	6.027	6.107	6.181	6.249	6.313	6.372	6.427	6.479	6.529	6.741	6.912	7.054	7.176	7.283	7.377
18	5.794	5.890	5.977	6.055	6.128	6.195	6.258	6.316	6.371	6.422	6.471	6.680	6.848	6.989	7.109	7.213	7.307
19	5.752	5.846	5.932	6.009	6.081	6.147	6.209	6.267	6.321	6.371	6.419	6.626	6.792	6.930	7.048	7.152	7.244
20	5.714	5.807	5.891	5.968	6.039	6.104	6.165	6.222	6.275	6.325	6.373	6.576	6.740	6.877	6.994	7.097	7.187
24	5.594	5.683	5.764	5.838	5.906	5.968	6.027	6.081	6.132	6.181	6.226	6.421	6.579	6.710	6.822	6.920	7.008
30	5.475	5.561	5.638	5.709	5.774	5.833	5.889	5.941	5.990	6.037	6.080	6.267	6.417	6.543	6.650	6.744	6.827
40	5.358	5.439	5.513	5.581	5.642	5.700	5.753	5.803	5.849	5.893	5.934	6.112	6.255	6.375	6.477	6.566	6.645
60	5.241	5.319	5.389	5.453	5.512	5.566	5.617	5.664	5.708	5.750	5.789	5.958	6.093	6.206	6.308	6.387	6.462
120	5.126	5.200	5.266	5.327	5.382	5.434	5.481	5.526	5.568	5.607	5.644	5.802	5.929	6.035	6.126	6.206	6.275
∞	5.012	5.081	5.144	5.201	5.253	5.301	5.346	5.388	5.427	5.463	5.498	5.646	5.764	5.863	5.947	6.020	6.085

TABLE A14 Percentage Points of the Studentized Range

$\alpha = 0.01$

ν \ k	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	90.03	135.0	164.3	185.6	202.2	215.8	227.2	237.0	245.6	253.2	260.0	266.2	271.8	277.0	281.8	286.3	290.4	294.3
2	14.04	19.02	22.29	24.72	26.63	28.20	29.53	30.68	31.69	32.59	33.40	34.13	34.81	35.43	36.00	36.53	37.03	37.50
3	8.261	10.62	12.17	13.33	14.24	15.00	15.64	16.20	16.69	17.13	17.53	17.89	18.22	18.52	18.81	19.07	19.32	19.55
4	6.512	8.120	9.173	9.958	10.58	11.10	11.55	11.98	12.27	12.57	12.84	13.09	13.32	13.53	13.73	13.91	14.08	14.24
5	5.702	6.976	7.804	8.421	8.913	9.321	9.669	9.972	10.24	10.48	10.70	10.89	11.08	11.24	11.40	11.55	11.68	11.81
6	5.243	6.331	7.033	7.556	7.973	8.318	8.613	8.869	9.097	9.301	9.485	9.653	9.808	9.951	10.08	10.21	10.32	10.43
7	4.949	5.919	6.543	7.005	7.373	7.679	7.939	8.166	8.368	8.548	8.711	8.860	8.997	9.124	9.242	9.353	9.456	9.554
8	4.746	5.635	6.204	6.625	6.960	7.287	7.474	7.681	7.863	8.027	8.176	8.312	8.436	8.552	8.659	8.760	8.854	8.943
9	4.596	5.428	5.957	6.348	6.658	6.915	7.134	7.325	7.495	7.647	7.784	7.910	8.025	8.132	8.232	8.325	8.412	8.496
10	4.482	5.270	5.769	6.136	6.428	6.669	6.875	7.055	7.213	7.356	7.485	7.603	7.712	7.812	7.906	7.993	8.076	8.153
11	4.392	5.146	5.621	5.970	6.247	6.476	6.672	6.842	6.992	7.128	7.250	7.362	7.465	7.560	7.649	7.732	7.809	7.883
12	4.320	5.046	5.502	5.836	6.101	6.321	6.507	6.670	6.814	6.943	7.060	7.167	7.265	7.356	7.441	7.520	7.594	7.665
13	4.260	4.964	5.404	5.727	5.981	6.192	6.372	6.528	6.667	6.791	6.903	7.006	7.101	7.188	7.269	7.345	7.417	7.485
14	4.210	4.895	5.322	5.634	5.881	6.085	6.258	6.409	6.543	6.664	6.772	6.871	6.962	7.047	7.126	7.199	7.268	7.333
15	4.168	4.836	5.252	5.556	5.796	5.994	6.162	6.309	6.489	6.555	6.660	6.757	6.845	6.927	7.003	7.074	7.142	7.204
16	4.131	4.786	5.192	5.489	5.722	5.915	6.079	6.222	6.349	6.462	6.564	6.658	6.744	6.823	6.898	6.967	7.032	7.093
17	4.099	4.742	5.140	5.430	5.659	5.847	6.007	6.147	6.270	6.381	6.480	6.572	6.656	6.734	6.806	6.873	6.937	6.997
18	4.071	4.708	5.094	5.379	5.608	5.788	5.944	6.081	6.201	6.310	6.407	6.497	6.579	6.655	6.725	6.792	6.854	6.912
19	4.046	4.670	5.054	5.334	5.554	5.735	5.889	6.022	6.141	6.247	6.342	6.430	6.510	6.585	6.654	6.719	6.780	6.837
20	4.024	4.639	5.018	5.294	5.510	5.688	5.839	5.970	6.087	6.191	6.285	6.371	6.450	6.523	6.591	6.654	6.714	6.771
24	3.956	4.546	4.907	5.168	5.374	5.542	5.685	5.809	5.919	6.017	6.106	6.186	6.261	6.330	6.394	6.453	6.510	6.563
30	3.889	4.455	4.799	5.048	5.242	5.401	5.586	5.653	5.756	5.849	5.932	6.008	6.078	6.143	6.203	6.259	6.311	6.361
40	3.825	4.367	4.696	4.981	5.114	5.265	5.392	5.502	5.599	5.686	5.764	5.835	5.900	5.961	6.017	6.069	6.119	6.165
60	3.762	4.282	4.595	4.818	4.991	5.133	5.253	5.356	5.447	5.528	5.601	5.667	5.728	5.785	5.837	5.886	5.981	5.974
120	3.702	4.200	4.497	4.709	4.872	5.005	5.118	5.214	5.299	5.375	5.443	5.505	5.562	5.614	5.662	5.708	5.750	5.790
∞	3.643	4.120	4.408	4.603	4.757	4.882	4.987	5.078	5.157	5.227	5.290	5.348	5.400	5.448	5.493	5.535	5.574	5.611

TABLE A14 Percentage Points of the Studentized Range

$\alpha = 0.01$

ν \ k	20	22	24	26	28	30	32	34	36	38	40	50	60	70	80	90	100
1	298.0	304.7	310.8	316.3	321.3	326.0	330.3	334.3	338.0	341.5	344.8	358.9	370.1	379.4	387.3	394.1	400.1
2	37.95	38.76	39.49	40.15	40.76	41.32	41.84	42.33	42.78	43.21	43.61	45.33	46.70	47.83	48.80	49.64	50.38
3	19.77	20.17	20.53	20.86	21.16	21.44	21.70	21.95	22.17	22.39	22.59	23.45	24.13	24.71	25.19	25.62	25.99
4	14.40	14.68	14.93	15.16	15.37	15.57	15.75	15.92	16.08	16.23	16.37	16.98	17.46	17.86	18.20	18.50	18.77
5	11.93	12.16	12.36	12.54	12.71	12.87	13.02	13.15	13.28	13.40	13.52	14.00	14.39	14.72	14.99	15.23	15.45
6	10.54	10.73	10.91	11.06	11.21	11.34	11.47	11.58	11.69	11.80	11.90	12.31	12.65	12.92	13.16	13.37	13.55
7	9.646	9.815	9.970	10.11	10.24	10.36	10.47	10.58	10.67	10.77	10.85	11.23	11.52	11.77	11.99	12.17	12.34
8	9.027	9.182	9.322	9.450	9.569	9.678	9.779	9.874	9.964	10.05	10.13	10.47	10.75	10.97	11.17	11.34	11.49
9	8.573	8.717	8.847	8.966	9.075	9.177	9.271	9.360	9.443	9.521	9.594	9.912	10.17	10.38	10.57	10.73	10.87
10	8.226	8.361	8.483	8.595	8.698	8.794	8.883	8.966	9.044	9.117	9.187	9.486	9.726	9.927	10.10	10.25	10.39
11	7.952	8.080	8.196	8.303	8.400	8.491	8.575	8.654	8.728	8.798	8.864	9.148	9.377	9.568	9.732	9.875	10.00
12	7.731	7.853	7.964	8.066	8.159	8.246	8.327	8.402	8.473	8.539	8.603	8.875	9.094	9.277	9.434	9.571	9.693
13	7.548	7.665	7.772	7.870	7.960	8.043	8.121	8.193	8.262	8.326	8.387	8.648	8.859	9.035	9.187	9.318	9.436
14	7.395	7.508	7.611	7.705	7.792	7.873	7.948	8.018	8.084	8.146	8.204	8.457	8.661	8.832	8.978	9.106	9.219
15	7.264	7.374	7.474	7.566	7.650	7.728	7.800	7.869	7.932	7.992	8.049	8.295	8.492	8.658	8.800	8.924	9.035
16	7.152	7.258	7.356	7.445	7.527	7.602	7.673	7.739	7.802	7.860	7.916	8.154	8.347	8.507	8.646	8.767	8.874
17	7.053	7.158	7.253	7.340	7.420	7.493	7.563	7.627	7.687	7.745	7.799	8.031	8.219	8.377	8.511	8.630	8.735
18	6.968	7.070	7.163	7.247	7.325	7.398	7.465	7.528	7.587	7.643	7.696	7.924	8.107	8.261	8.393	8.508	8.611
19	6.891	6.992	7.082	7.166	7.242	7.313	7.379	7.440	7.498	7.553	7.605	7.828	8.008	8.159	8.288	8.401	8.502
20	6.823	6.922	7.011	7.092	7.168	7.237	7.302	7.362	7.419	7.473	7.523	7.742	7.919	8.067	8.194	8.305	8.404
24	6.612	6.705	6.789	6.865	6.936	7.001	7.062	7.119	7.173	7.223	7.270	7.476	7.642	7.780	7.900	8.004	8.097
30	6.407	6.494	6.572	6.644	6.710	6.772	6.828	6.881	6.932	6.978	7.023	7.215	7.370	7.500	7.611	7.709	7.796
40	6.209	6.289	6.362	6.429	6.490	6.547	6.600	6.650	6.697	6.740	6.782	6.960	7.104	7.225	7.328	7.419	7.500
60	6.015	6.090	6.158	6.220	6.277	6.330	6.378	6.424	6.467	6.507	6.546	6.710	6.843	6.954	7.050	7.133	7.207
120	5.827	5.897	5.969	6.016	6.069	6.117	6.162	6.204	6.244	6.281	6.316	6.467	6.588	6.689	6.776	6.852	6.919
∞	5.645	5.709	5.766	5.818	5.866	5.911	5.952	5.990	6.026	6.060	6.092	6.228	6.338	6.429	6.507	6.575	6.636

Source: Adapted from Harter, H. L. (1960). "Tables of Range and Studentized Range," *Annals of Mathematical Statistics*, **31**, 1122–1147. Used by permission of the Institute of Mathematical Statistics.

Index

Added factors, 260
Adjacent value, 70
Adjusted factor-level average, 362
  analysis of covariance, 551, 555, 556
Alias, 230, 319, 324
All-possible subset comparisons, 662
Alternative hypothesis, 52, 55
Analysis of covariance (ANACOVA), 535, 543
Analysis of marginal means, 276
Analysis of variance (ANOVA), 481, 503
  model, 173
  table, 176, 481, 504, 546
Assignable causes, 173
Assumptions
  analysis of covariance model, 544, 554
  fixed effects analysis of variance model, 173
  linear regression analysis, 463, 481
  random effects analysis of variance model, 349
Average, 33

Backward elimination, 668
Balance, 145, 248, 252
Balanced incomplete block design (BIB), 325, 357
Balancing, 316
Bartlett's test, 98
Bias measurement, 402
Block, 110, 316
Block design, 311, 317, 318, 400, 552
  complete, 317, 356
  incomplete, 318, 357
Bonferroni comparisons, 211
Box–Behnken design, 248, 585
Box–Cox procedure, 642
Boxplot comparisons, 70

C_m (C_p)statistic, 661
Canonical analysis, 593, 604
Capability study, 404
Carryover effects, 331
Categorical variable, 536, 539
Central composite design, 248, 582
Central limit property, 47
Chi-square probability distribution, 46, 62, 82
Coding, 541, 578
Coefficient of determination, 482, 504, 661
  adjusted, 505, 661
Collinear predictors, 518
Collinearity detection, 672
Combined array design, 594
Complete block design, 317, 356
Completely randomized design, 141, 170, 229, 542
  three-level, 247
  two-level, 239
Comprehensive regression analysis, 470
Computer-generated design, 580
Confidence coefficient, 75
Confidence interval, 49
  equivalence to hypothesis test, 78
  for analysis of variance model, 186
  for factor level means, 275
  for normal distribution parameters, 49, 76, 83, 92, 94
  for ratio of expected mean squares, 355

Confidence interval (*continued*)
for regression model parameters, 484, 510
for regression model response mean, 485, 524
interpretation, 50, 75
simultaneous, 524
Confidence level, 55
Confounding, 112, 318
effects, 229
partial, 320
pattern, 233, 281, 287
Constrained factor space, 588
Contrast, 161, 197, 230
orthogonal, 198
sum of squares, 200
Correlation coefficient, 468
Covariate, 110, 542
Criteria for comparing fitted models, 661
Critical value, 55
Crossed array design, 594
Crossover design, 331
Curvature, 413

Data, 4
collection, 4
Defining contrast, 239, 319, 324
Defining equation, 239, 321
Degrees of freedom, 46, 73, 90, 180
Density, 20, 42
Design problems
erroneous efficiency, 119
error variation, 115
masked factors, 115
one-factor-at-a-time, 121
uncontrolled factors, 117
Design resolution, 237
Design selection criteria
efficiency, 127
factor effects, 126
objective, 125
precision, 127
randomization, 128
Deviations, 35
DFBETAS, 625
DFFITS, 625
Discrimination, 402
Distribution, 19
frequency, 21
normal, 20
sampling, 21, 45
Dot notation, 153
Dual response model, 598

Effects, 110
calculation of, 152, 156, 160
coding of factor levels, 154
confounded, 230
fixed, 171
graphical assessment, 158
interaction, 153, 175
joint factor, 145, 178
linear, 290
main, 153
parameters, 277
plot, 280
polynomial, 190
quadratic, 290
random, 171, 347, 424
representation, 153
Effects sum of squares, *see* Sum of squares
Error mean square (MSE), 185, 273, 482, 504
Error rates
comparisonwise, 201
experimentwise, 201
type I, 201
type II, 202
Error standard deviation, 184
Estimate, 44
Estimated error standard deviation, 73, 157, 276, 482, 504
Estimated experimental error, 185
Evolutionary operation (EVOP), 130
Expected mean square, 350, 366, 425
Experimental error, 411
Experimental layout, 112
Experimental region, 110, 575
Experimental studies, 4
Experimental unit, 110

F probability distribution, 46, 62
F ratio, 46, 93, 195, 661
F statistic, 93
Face-centered cube design, 583
Factor, 12
Factor effects, *see* Effects, 171
Factor levels, 110, 189
quantitative, 189
random, 347
space, 111

Factorial experiments, 141, 228, 580
Factors, 12
  balanced, 382
  control, 408, 598
  crossed, 379
  environmental, 408
  hard to vary, 391
  nested, 379
  noise, 598
  uncontrolled, 117
F-max test, 97
Fold-over designs, 260
Forward selection, 666
Fractional factorial, 228, 321
Fractional factorial experiment, 144, 228, 278, 287, 290, 580
  analysis of, 278, 287, 290
French curve, 516

Gage R&R studies, 401, 433
Graeco-Latin-Square design, 330
Grouping, 316
Grubbs test, 617

Hierarchical model, *see* Model
Hierarchically nested designs, 381, 423
Histogram
  relative-frequency, 22
Hybrid design, 588
Hypothesis tests
  analysis of covariance model parameters, 544
  analysis of variance model parameters, 194, 275
  decision rules, 77
  for factor effects, 195
  for normal distribution parameters, 52, 78, 85, 92, 95
  lack of fit, 506
  p-value calculations, 77
  regression model parameters, 484, 508
Hypothesis types, 52

Incomplete block design, 318, 357
  balanced, 325
  three-level factorial, 323
  two-level factorial, 318
Independence, 44
Indicator variables, 461, 536
Inferences
  on means, 72, 86, 89
  on standard deviations two samples, 81, 93
  on regression models, 481, 503
Influential observations, 624
Inner array, 408
Integrated approach, 410
Integrated design model, 598
Interaction, 110, 146, 511, 541
Interaction plot, 216

Lack-of-fit
  error, 482, 506
  test, 506
Latin-square design, 328, 364
Least significant
  difference, 210
  interval, 218
Least squares estimation, 475, 478, 480, 497
  interpretation, 478, 499
Least squares fit, 476
Least squares means, 278
Leverage values, 628
Local control, 316
Loess smoothing, 474

Masking, 115, 618
Mean, 33
Mean square, 181
  Expected, 350, 366, 425
Measurement process, 401
Measurement variation, 6, 402
Median, 34
Mixed-levels designs, 254
Mixture design, 588
Model
  analysis of covariance, 543, 553
  analysis of variance, 173
  assumptions, 630
  extrapolation, 463, 518
  first-order, 513, 515
  fixed effects, 173, 348, 429
  hierarchical, 176, 272
  integrated design, 587
  linear, 462, 497, 536
  mathematical, 25
  no-intercept regression, 480, 485
  nonlinear, 640
  one-way classification, 184
  order, 513, 515
  polynomial, 514, 588

Model (*continued*)
random-effect, 349
regression, 462, 497, 536
respecification, 639
response surface, 588
saturated, 185
second-order, 513, 515
specification, 462, 472, 497, 634
statistical, 25
sum of squares, *see* Sum of squares
Multi-panel conditioning, 214
Multiple comparison procedures, 196

Nested design, 378, 423
Nested factors, *see* Factor
Noncentral composite design, 588
Nonlinear
relationship, 635, 640
response function, 640
Nonorthogonal designs, 252
Normal density function, 43
Normal equations, 499
Normal probability distribution, 20, 43, 59
Null hypothesis, 52, 55

Observation, 12
Observational
data, 587
studies, 4
Observed value, 11
One-Factor-at-a-Time (OFAT) Testing, 121
Operating characteristic curve, 59, 80
Optimum response, 573, 576
Orthogonal arrays, 252, 407, 414
Orthogonal contrast, *see* Contrast
Orthogonal polynomials, 207
Outer array, 408
Outliers, 70
accommodation, 615
detection, 614
in predictor variables, 626
in response variables, 619

Parameter, 19
analysis of variance model, 174
constraints, 174
estimation, 186
interaction, 174
main-effect, 174
Parsimony, 516
Partial regression coefficient estimate, 499
Pearson's r, 468
Pick the winner, 437, 457
PISEAS, 470
Plackett–Burman design, 256
analysis, 293
Plots
boxplot, 70
contour, 122, 571
cube, 213
factor effects, 158
interaction, 150, 216
labeled scatterplot, 148
least significant interval, 218
normal quantile-quantile, 159, 630
overlaying, 575
partial-regression, 637
partial-residual, 635
point, 39
residual, 634
scatter, 6, 148
studentized deleted residual, 644
trellis, 214
Pooled standard deviation estimate, 90
Population, 10
Power, 56
Precision, 127
Prediction
equation, 475, 498
interval, 485
Probability concepts, 42
Product array design, 594
Process, 10
capability, 404
Pure error, 506
p-value, 55, 77

Quadratic model, 515, 588
Quality control procedures, 128
Quality loss function, 407
Quantile, 159
Quartile, 37

Random sample, 14
Randomization, 128, 142, 391
restricted, 391
Randomized complete block (RCB) design, 317, 552
Range, 35

Reduction in error sum of squares, 273
Reexpression, 642, 647
Regression analysis
  assumptions, 463, 481
  common uses and misuses, 462
  analysis linear, 470
  local fit, 473
  strategy, 470
  sum of squares, 481, 503
Regression coefficient, 462, 497
  beta-weight, 502, 519
  standardized, 502, 519
Regression fallacy, 480
Repeat test, 110, 144, 312
Repeatability, 315, 402, 434
Replication, 110, 312
Reproducibility, 315, 402, 434
Residuals, 475, 498
  partial, 635
  studentized deleted, 623
Response
  dispersion, 130
  location, 130
  predicted, 475, 498
  variable, 120
Response surface designs, 129, 410, 568, 580
  Box–Behnken, 413, 585
  central composite, 413, 582
Rising ridge, 572
Robust design, 401
  parameter design, 587, 594
Robustness, 401
Rotatable design, 581
Ruggedness tests, 267, 293

Saddle, 572
Sample correlation coefficient, 468
Sample, 13
  mean 34
  sampling distribution of, 45
  median, 34
  size, 56, 79
  standard deviation, 36, 73
  types of, 13
  variance, 46
Sampling distribution, 21, 45, 73
Saturated designs, 253
Scatterplot 6; smoothing, 473
Screening design, 144, 256
  analysis, 293
Screening experiments, 129, 158, 256
Semi-interquartile range, 37, 70
Sequential experimentation, 255, 331
Shapiro–Wilk test for normality, 633
Signal-to-noise ratio, 436
Significance level, 55
Significance probability, 55
Simplex design, 588
Small composite design, 588
Smoothing, 473
Span, 473
Split plot, 384, 388
Split-plot design, 384, 388, 428
Stable process, 404
Staggered nested design, 382, 427
Standard deviation, 36, 81
Standard error, 45
Standard normal variate, 44
Standard process, 404
Standardized predictor variable, 519, 591
Stationary point, 593
Stationary ridge, 571
Statistic, 4, 19
Steepest ascent method, 602
Stepwise variable selection techniques, 665
  collinear effects, 672
Student t distribution, 46, 60
  approximate, 91
Sum of squares
  contrast, 200
  error, 180, 481, 498
  interaction effect, 178
  main effect, 178
  model, 177
  partitioned, 177
  regression, 481, 503
  total, 177, 481
Supersaturated design analysis, 297

Taguchi approach, 129, 406
Taguchi design, 406, 436
t-distribution, 46
Test run, 110
Tolerance intervals, 50
Transformation
  Box–Cox, 642
  logarithmic, 640
  power-family, 642
Transmitted variation, 411
t-statistic, 73, 209

Tukey's significant difference, 212
Type I error, 55, 202
Type II error, 55, 202

Unbalanced design
  analysis, 272

Variable 11
  categorical, 536
  collinear, 518
  continuous, 42
  indicator, 461, 536
  predictor, 461, 462, 497
  response, 12, 110
  selection techniques, 659
  standardized, 579
Variance component, 350, 434
Variance inflation factors, 673
Variation
  assignable cause, 173
  measurement, 402
  random, 173
  sources of, 173
  transmitted, 411

Whole plots, 388

Data Index

Chapter	Title/Description	Page	Other Chapter	Data Location*
1	Chemical analysis variation study	7	no	T1.1
1	Employee identification numbers	16	no	T1.3
1	Skin elasticity measurements	18	no	T1.4
2	Diesel fuel economy, 4 vehicles, 8 fuels	38	no	T2.1
2	Wire tensile strength	40	2,3	T2.2
3	Percent solids in two pipeline locations	86	no	T3.5
4	Cylinder pressure measurements	116	no	T4.4
4	Skin color measurements	116	12	T4.5
5	Cutting tool life	147	no	T5.4
6	Chemical yield from pilot plant	172	5,7,8,9	T6.1
6	Filtration flow rates	184	no	T6.5
6	Copper plate warping	191	no	T6.7
8	Chemical yield from pilot plant (unbalanced)	274	5,6,7	T8.1
8	Product color in a chemical process	280	no	T8.4
8	Density of bricks	288	no	T8.9
8	Diesel fuel NO_x experiment	292	no	T8.14
8	Solder-coverage ruggedness testing data	294	no	T8.17
8	Ruggedness test viscosity data	296	no	T8.20
8	Fatigue life of weld-repaired castings	298	no	T8.23
10	Asphalt pavement rating study	358	no	T10.4
10	Tire brands road test	365	no	T10.9
11	NO_x emissions for 2 vehicles from 3 models	385	no	T11.2
11	NO_x emissions for 6 vehicles and 3 fuels	386	no	T11.3
12	Gage R&R study on innerspring length	405	9	F12.2

Chapter	Title/Description	Page	Other Chapter	Data Location
13	Polyethylene strength	426	no	T13.2
13	Lawnmower cutoff times	431	11	T13.5
13	Steering column torque measurements	438	no	T13.9
14	Tire tread lifetimes	467	no	T14.3
14	Acid content and acid number	479	no	T14.5
15	Tire treadwear rates for tire brand C	501	16	T15.1
15	Synthetic-rubber process data	516	no	T15.7
16	Tire treadwear rates for tire brands A and B	537	no	T16.1
16	Reactor residence times	548	no	T16.5
17	Polymer density	590	no	T17.7
17	Steering column torque	595	12	T17.10
18	Chemical process yield data	620	no	T18.1
18	Torque data	631	no	T18.5
19	Cryogenic flow meter	674	no	T19.A

*T refers to the Table Number, and F refers to the Figure Number where the data are located.