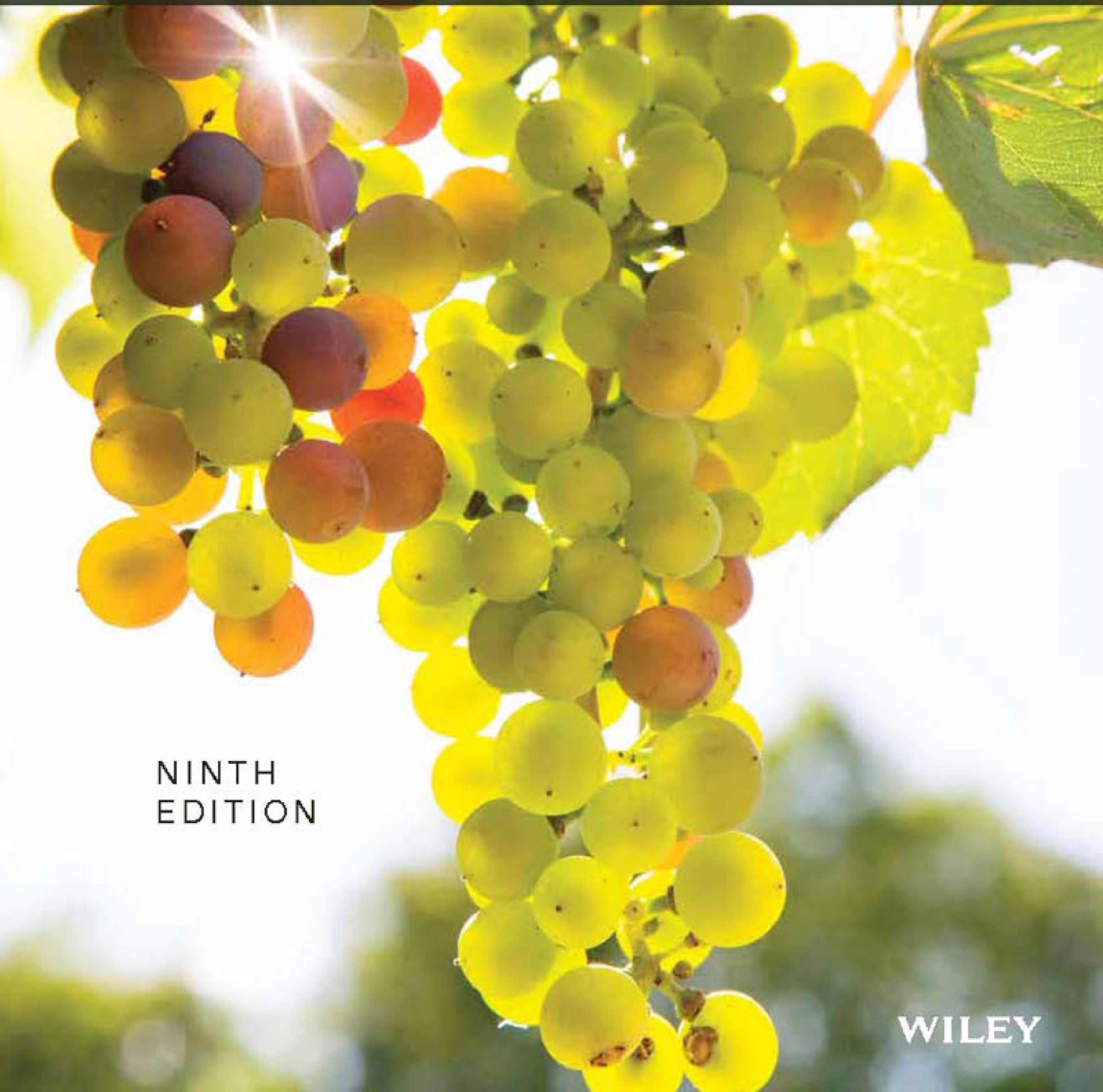
DESIGN AND ANALYSIS OF EXPERIMENTS

DOUGLAS C. MONTGOMERY



NINTH EDITION

WILEY

Design and Analysis of Experiments

Ninth Edition

DOUGLAS C. MONTGOMERY

Arizona State University

WILEY

VP AND EDITORIAL DIRECTOR	Laurie Rosatone
SENIOR DIRECTOR	Don Fowley
ACQUISITIONS EDITOR	Linda Ratts
DEVELOPMENT EDITOR	Chris Nelson
EDITORIAL MANAGER	Gladys Soto
CONTENT MANAGEMENT DIRECTOR	Lisa Wojcik
CONTENT MANAGER	Nichole Urban
SENIOR CONTENT SPECIALIST	Nicole Repasky
PRODUCTION EDITOR	Linda Christina E
COVER PHOTO CREDIT	© Echo / Getty Images

This book was set in TimesLTStd by SPi Global and printed and bound by Lightning Source, Inc.

ISBN: 9781119113478 (PBK)
ISBN: 9781119299455 (EVALC)

Library of Congress Cataloging-in-Publication Data:

Names: Montgomery, Douglas C., author.
Title: Design and analysis of experiments / Douglas C. Montgomery, Arizona
State University.
Description: Ninth edition. | Hoboken, NJ : John Wiley & Sons, Inc., [2017] |
Includes bibliographical references and index.
Identifiers: LCCN 2017002355 (print) | LCCN 2017002997 (ebook) | ISBN
9781119113478 (pbk.) | ISBN 9781119299363 (pdf) | ISBN 9781119320937 (epub)
Subjects: LCSH: Experimental design.
Classification: LCC QA279 .M66 2017 (print) | LCC QA279 (ebook) | DDC
519.5/7—dc23
LC record available at https://lccn.loc.gov/2017002355

Preface

Audience

This is an introductory textbook dealing with the design and analysis of experiments. It is based on college-level courses in design of experiments that I have taught for over 40 years at Arizona State University, the University of Washington, and the Georgia Institute of Technology. It also reflects the methods that I have found useful in my own professional practice as an engineering and statistical consultant in many areas of science and engineering, including the research and development activities required for successful technology commercialization and product realization.

The book is intended for students who have completed a first course in statistical methods. This background course should include at least some techniques of descriptive statistics, the standard sampling distributions, and an introduction to basic concepts of confidence intervals and hypothesis testing for means and variances. Chapters 10, 11, and 12 require some familiarity with matrix algebra.

Because the prerequisites are relatively modest, this book can be used in a second course on statistics focusing on statistical design of experiments for undergraduate students in engineering, the physical and chemical sciences, statistics, mathematics, and other fields of science. For many years I have taught a course from the book at the first-year graduate level in engineering. Students in this course come from all of the fields of engineering, materials science, physics, chemistry, mathematics, operations research life sciences, and statistics. I have also used this book as the basis of an industrial short course on design of experiments for practicing technical professionals with a wide variety of backgrounds. There are numerous examples illustrating all of the design and analysis techniques. These examples are based on real-world applications of experimental design and are drawn from many different fields of engineering and the sciences. This adds a strong applications flavor to an academic course for engineers and scientists and makes the book useful as a reference tool for experimenters in a variety of disciplines.

About the Book

The ninth edition is a significant revision of the book. I have tried to maintain the balance between design and analysis topics of previous editions; however, there are many new topics and examples, and I have reorganized some of the material. There continues to be a lot of emphasis on the computer in this edition.

Design-Expert, JMP, and Minitab Software

During the last few years a number of excellent software products to assist experimenters in both the design and analysis phases of this subject have appeared. I have included output from three of these products, Design-Expert, JMP, and Minitab at many points in the text. Minitab and JMP are widely available general-purpose statistical software packages that have good data analysis capabilities and that handles the analysis of experiments with both fixed and random factors (including the mixed model). Design-Expert is a package focused exclusively on experimental design. All three of these packages have many capabilities for construction and evaluation of designs and extensive analysis features. I urge all instructors who use this book to incorporate computer software into your course. (In my course, I bring a laptop computer, and every design or analysis topic discussed in class is illustrated with the computer.)

Empirical Model

I have continued to focus on the connection between the experiment and the model that the experimenter can develop from the results of the experiment. Engineers (and physical, chemical and life scientists to a large extent) learn about physical mechanisms and their underlying mechanistic models early in their academic training, and throughout much of their professional careers they are involved with manipulation of these models. Statistically designed experiments offer the engineer a valid basis for developing an *empirical* model of the system being investigated. This empirical model can then be manipulated (perhaps through a response surface or contour plot, or perhaps mathematically) just as any other engineering model. I have discovered through many years of teaching that this viewpoint is very effective in creating enthusiasm in the engineering community for statistically designed experiments. Therefore, the notion of an underlying empirical model for the experiment and response surfaces appears early in the book and continues to receive emphasis.

Factorial Designs

I have expanded the material on factorial and fractional factorial designs (Chapters 5–9) in an effort to make the material flow more effectively from both the reader's and the instructor's viewpoint and to place more emphasis on the empirical model. There is new material on a number of important topics, including follow-up experimentation following a fractional factorial, nonregular and nonorthogonal designs, and small, efficient resolution IV and V designs. Nonregular fractions as alternatives to traditional minimum aberration fractions in 16 runs and analysis methods for these design are discussed and illustrated.

Additional Important Changes

I have added material on optimal designs and their application. The chapter on response surfaces (Chapter 11) has several new topics and problems. I have expanded Chapter 12 on robust parameter design and process robustness experiments. Chapters 13 and 14 discuss experiments involving random effects and some applications of these concepts to nested and split-plot designs. The residual maximum likelihood method is now widely available in software and I have emphasized this technique throughout the book. Because there is expanding industrial interest in nested and split-plot designs, Chapters 13 and 14 have several new topics. Chapter 15 is an overview of important design and analysis topics: nonnormality of the response, the Box–Cox method for selecting the form of a transformation, and other alternatives; unbalanced factorial experiments; the analysis of covariance, including covariates in a factorial design, and repeated measures. I have also added new examples and problems from various fields, including biochemistry and biotechnology.

Experimental Design

Throughout the book I have stressed the importance of experimental design as a tool for engineers and scientists to use for product design and development as well as process development and improvement. The use of experimental design in developing products that are robust to environmental factors and other sources of variability is illustrated. I believe that the use of experimental design early in the product cycle can substantially reduce development lead time and cost, leading to processes and products that perform better in the field and have higher reliability than those developed using other approaches.

The book contains more material than can be covered comfortably in one course, and I hope that instructors will be able to either vary the content of each course offering or discuss some topics in greater depth, depending on class interest. There are problem sets at the end of each chapter. These problems vary in scope from computational exercises, designed to reinforce the fundamentals, to extensions or elaboration of basic principles.

Course Suggestions

My own course focuses extensively on factorial and fractional factorial designs. Consequently, I usually cover Chapter 1, Chapter 2 (very quickly), most of Chapter 3, Chapter 4 (excluding the material on incomplete blocks and only mentioning Latin squares briefly), and I discuss Chapters 5 through 8 on factorials and two-level factorial and fractional factorial designs in detail. To conclude the course, I introduce response surface methodology (Chapter 11) and give an overview of random effects models (Chapter 13) and nested and split-plot designs (Chapter 14). I always require the students to complete a term project that involves designing, conducting, and presenting the results of a statistically designed experiment. I require them to do this in teams because this is the way that much industrial experimentation is conducted. They must present the results of this project, both orally and in written form.

The Supplemental Text Material

For this edition I have provided supplemental text material for each chapter of the book. Often, this supplemental material elaborates on topics that could not be discussed in greater detail in the book. I have also presented some subjects that do not appear directly in the book, but an introduction to them could prove useful to some students and professional practitioners. Some of this material is at a higher mathematical level than the text. I realize that instructors use this book with a wide array of audiences, and some more advanced design courses could possibly benefit from including several of the supplemental text material topics. This material is in electronic form on the World Wide Website for this book, located at www.wiley.com/college/montgomery.

Website

Current supporting material for instructors and students is available at the website www.wiley.com/college/montgomery. This site will be used to communicate information about innovations and recommendations for effectively using this text. The supplemental text material described above is available at the site, along with electronic versions of data sets used for examples and homework problems, a course syllabus, and some representative student term projects from the course at Arizona State University.

Student Companion Site

The student's section of the textbook website contains the following:

1. The supplemental text material described above
2. Data sets from the book examples and homework problems, in electronic form
3. Sample Student Projects

Instructor Companion Site

The instructor's section of the textbook website contains the following:

1. Solutions to the text problems
2. The supplemental text material described above
3. PowerPoint lecture slides
4. Figures from the text in electronic format, for easy inclusion in lecture slides
5. Data sets from the book examples and homework problems, in electronic form
6. Sample Syllabus
7. Sample Student Projects

The instructor's section is for instructor use only, and is password-protected. Visit the Instructor Companion Site portion of the website, located at www.wiley.com/college/montgomery, to register for a password.

Student Solutions Manual

The purpose of the Student Solutions Manual is to provide the student with an in-depth understanding of how to apply the concepts presented in the textbook. Along with detailed instructions on how to solve the selected chapter exercises, insights from practical applications are also shared.

Solutions have been provided for problems selected by the author of the text. Occasionally a group of "continued exercises" is presented and provides the student with a full solution for a specific data set. Problems that are included in the Student Solutions Manual are indicated by an icon appearing in the text margin next to the problem statement.

This is an excellent study aid that many text users will find extremely helpful. The Student Solutions Manual may be ordered in a set with the text, or purchased separately. Contact your local Wiley representative to request the set for your bookstore, or purchase the Student Solutions Manual from the Wiley website.

Acknowledgments

I express my appreciation to the many students, instructors, and colleagues who have used the eight earlier editions of this book and who have made helpful suggestions for its revision. The contributions of Dr. Raymond H. Myers, Dr. G. Geoffrey Vining, Dr. Brad Jones, Dr. Christine Anderson-Cook, Dr. Connie M. Borror, Dr. Scott Kowalski, Dr. Rachel Silvestrini, Dr. Megan Olson Hunt, Dr. Dennis Lin, Dr. John Ramberg, Dr. Joseph Pignatiello, Dr. Lloyd S. Nelson, Dr. Andre Khuri, Dr. Peter Nelson, Dr. John A. Cornell, Dr. Saeed Maghsoodloo, Dr. Don Holcomb, Dr. George C. Runger, Dr. Bert Keats, Dr. Dwayne Rollier, Dr. Norma Hubele, Dr. Murat Kulahci, Dr. Cynthia Lowry, Dr. Russell G. Heikes, Dr. Harrison M. Wadsworth, Dr. William W. Hines, Dr. Arvind Shah, Dr. Jane Ammons, Dr. Diane Schaub, Mr. Mark Anderson, Mr. Pat Whitcomb, Dr. Pat Spagon, and Dr. William DuMouche were particularly valuable. My current and former School Director and Department Chair, Dr. Ron Askin and Dr. Gary Hogg, have provided an intellectually stimulating environment in which to work.

The contributions of the professional practitioners with whom I have worked have been invaluable. It is impossible to mention everyone, but some of the major contributors include Dr. Dan McCarville, Dr. Lisa Custer, Dr. Richard Post, Mr. Tom Bingham, Mr. Dick Vaughn, Dr. Julian Anderson, Mr. Richard Alkire, and Mr. Chase Neilson of the Boeing Company; Mr. Mike Goza, Mr. Don Walton, Ms. Karen Madison, Mr. Jeff Stevens, and Mr. Bob Kohm of Alcoa; Dr. Jay Gardiner, Mr. John Butora, Mr. Dana Lesher, Mr. Lolly Marwah, Mr. Leon Mason of IBM; Dr. Paul Tobias of IBM and Sematech; Ms. Elizabeth A. Peck of The Coca-Cola Company; Dr. Sadri Khalessi and Mr. Franz Wagner of Signetics; Mr. Robert V. Baxley of Monsanto Chemicals; Mr. Harry Peterson-Nedry and Dr. Russell Boyles of Precision Castparts Corporation; Mr. Bill New and Mr. Randy Schmid of Allied-Signal Aerospace; Mr. John M. Fluke, Jr. of the John Fluke Manufacturing Company; Mr. Larry Newton and Mr. Kip Howlett of Georgia-Pacific; and Dr. Ernesto Ramos of BBN Software Products Corporation.

I am indebted to Professor E. S. Pearson and the *Biometrika* Trustees, John Wiley & Sons, Prentice Hall, The American Statistical Association, The Institute of Mathematical Statistics, and the editors of *Biometrics* for permission to use copyrighted material. Dr. Lisa Custer and Dr. Dan McCorville did an excellent job of preparing the solutions that appear in the Instructor's Solutions Manual, and Dr. Cheryl Jennings provided effective and very helpful proofreading assistance. I am grateful to NASA, the Office of Naval Research, the Department of Defense, the National Science Foundation, the member companies of the NSF/Industry/University Cooperative Research Center in Quality and Reliability Engineering at Arizona State University, and the IBM Corporation for supporting much of my research in engineering statistics and experimental design over many years.

DOUGLAS C. MONTGOMERY
TEMPE, ARIZONA

Contents

8
Two-Level Fractional Factorial Designs 328

9
Additional Design and Analysis Topics for Factorial and Fractional Factorial Designs 405

10
Fitting Regression Models (online at www.wiley.com/college/montgomery) 460

11
Response Surface Methods and Designs 489

CHAPTER 1

Introduction

CHAPTER OUTLINE

The supplemental material is on the textbook website www.wiley.com/college/montgomery.

CHAPTER LEARNING OBJECTIVES

1. Learn about the objectives of experimental design and the role it plays in the knowledge discovery process.
2. Learn about different strategies of experimentation.
3. Understand the role that statistical methods play in designing and analyzing experiments.
4. Understand the concepts of main effects of factors and interaction between factors.
5. Know about factorial experiments.
6. Know the practical guidelines for designing and conducting experiments.

1.1 Strategy of Experimentation

Observing a system or process while it is in operation is an important part of the learning process and is an integral part of understanding and learning about how systems and processes work. The great New York Yankees catcher Yogi Berra said that "... you can observe a lot just by watching." However, to understand what happens to a process when you change certain input factors, you have to do more than just watch—you actually have to change the factors. This means that to really understand cause-and-effect relationships in a system you must deliberately change the input variables to the system and observe the changes in the system output that these changes to the inputs produce. In other words, you need to conduct **experiments** on the system. Observations on a system or process can lead to theories or hypotheses about what makes the system work, but experiments of the type described above are required to demonstrate that these theories are correct.

Investigators perform experiments in virtually all fields of inquiry, usually to discover something about a particular process or system or to confirm previous experience or theory. Each experimental **run** is a **test**. More formally, we can define an **experiment** as a test or series of runs in which purposeful changes are made to the input variables of a process or system so that we may observe and identify the reasons for changes that may be observed in the output response. We may want to determine which input variables are responsible for the observed changes in the response, develop a model relating the response to the important input variables, and use this model for process or system improvement or other decision-making.

This book is about planning and conducting experiments and about analyzing the resulting data so that valid and objective conclusions are obtained. Our focus is on experiments in engineering and science. Experimentation plays an important role in **technology commercialization** and **product realization** activities, which consist of new product design and formulation, manufacturing process development, and process improvement. The objective in many cases may be to develop a **robust** process, that is, a process affected minimally by external sources of variability. There are also many applications of designed experiments in a nonmanufacturing or non-product-development setting, such as marketing, service operations, and general business operations. Designed experiments are a key technology for **innovation**. Both **break through innovation** and **incremental innovation** activities can benefit from the effective use of designed experiments.

As an example of an experiment, suppose that a metallurgical engineer is interested in studying the effect of two different hardening processes, oil quenching and saltwater quenching, on an aluminum alloy. Here the objective of the **experimenter** (the engineer) is to determine which quenching solution produces the maximum hardness for this particular alloy. The engineer decides to subject a number of alloy specimens or test coupons to each quenching medium and measure the hardness of the specimens after quenching. The average hardness of the specimens treated in each quenching solution will be used to determine which solution is best.

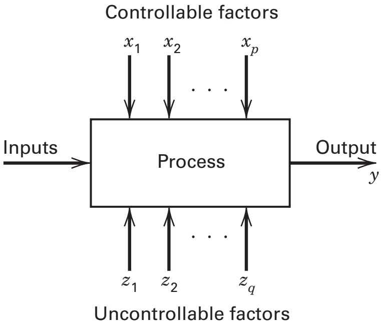
As we consider this simple experiment, a number of important questions come to mind:

1. Are these two solutions the only quenching media of potential interest?
2. Are there any other factors that might affect hardness that should be investigated or controlled in this experiment (such as the temperature of the quenching media)?
3. How many coupons of alloy should be tested in each quenching solution?
4. How should the test coupons be assigned to the quenching solutions, and in what order should the data be collected?
5. What method of data analysis should be used?
6. What difference in average observed hardness between the two quenching media will be considered important?

All of these questions, and perhaps many others, will have to be answered satisfactorily before the experiment is performed.

Experimentation is a vital part of the **scientific** (or **engineering**) **method**. Now there are certainly situations where the scientific phenomena are so well understood that useful results including mathematical models can be developed directly by applying these well-understood principles. The models of such phenomena that follow directly from the physical mechanism are usually called **mechanistic models**. A simple example is the familiar equation for current flow in an electrical circuit, Ohm's law, $E = IR$. However, most problems in science and engineering require **observation** of the system at work and **experimentation** to elucidate information about why and how it works. Well-designed experiments can often lead to a model of system performance; such experimentally determined models are called **empirical models**. Throughout this book, we will present techniques for turning the results of a designed experiment into an empirical model of the system under study. These empirical models can be manipulated by a scientist or an engineer just as a mechanistic model can.

A well-designed experiment is important because the results and conclusions that can be drawn from the experiment depend to a large extent on the manner in which the data were collected. To illustrate this point, suppose that the metallurgical engineer in the above experiment used specimens from one heat in the oil quench and specimens from a second heat in the saltwater quench. Now, when the mean hardness is compared, the engineer is unable to say how much of the observed difference is the result of the quenching media and how much is the result of inherent differences



■ **FIGURE 1.1 General model of a process or system**

between the heats.[1] Thus, the method of data collection has adversely affected the conclusions that can be drawn from the experiment.

In general, experiments are used to study the performance of processes and systems. The process or system can be represented by the model shown in Figure 1.1. We can usually visualize the process as a combination of operations, machines, methods, people, and other resources that transforms some input (often a material) into an output that has one or more observable **response** variables. Some of the process variables and material properties $x_1, x_2, \ldots, x_p$ are **controllable**, whereas other variables such as environmental factors or some material properties $z_1, z_2, \ldots, z_q$ are **uncontrollable** (although they may be controllable for purposes of a test). The objectives of the experiment may include the following:

1. Determining which variables are most influential on the response y
2. Determining where to set the influential x's so that y is almost always near the desired nominal value
3. Determining where to set the influential x's so that variability in y is small
4. Determining where to set the influential x's so that the effects of the uncontrollable variables $z_1, z_2, \ldots, z_q$ are minimized.

As you can see from the foregoing discussion, experiments often involve several factors. Usually, an objective of the **experimenter** is to determine the influence that these factors have on the output response of the system. The general approach to planning and conducting the experiment is called the **strategy of experimentation**. An experimenter can use several strategies. We will illustrate some of these with a very simple example.

I really like to play golf. Unfortunately, I do not enjoy practicing, so I am always looking for a simpler solution to lowering my score. Some of the factors that I think may be important, or that may influence my golf score, are as follows:

1. The type of driver used (oversized or regular sized)
2. The type of ball used (balata or three piece)
3. Walking and carrying the golf clubs or riding in a golf cart
4. Drinking water or drinking "something else" while playing
5. Playing in the morning or playing in the afternoon
6. Playing when it is cool or playing when it is hot
7. The type of golf shoe spike worn (metal or soft)
8. Playing on a windy day or playing on a calm day.

Obviously, many other factors could be considered, but let's assume that these are the ones of primary interest. Furthermore, based on long experience with the game, I decide that factors 5 through 8 can be ignored; that is, these

[1] A specialist in experimental design would say that the effects of quenching media and heat were *confounded*; that is, the effects of these two factors cannot be separated.

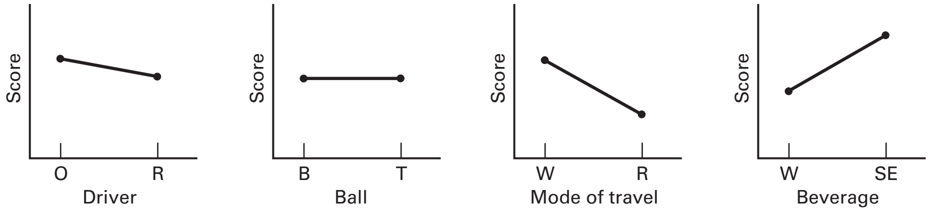
factors are not important because their effects are so small that they have no practical value. Engineers, scientists, and business analysts often must make these types of decisions about some of the factors they are considering in real experiments.

Now, let's consider how factors 1 through 4 could be experimentally tested to determine their effect on my golf score. Suppose that a maximum of eight rounds of golf can be played over the course of the experiment. One approach would be to select an arbitrary combination of these factors, test them, and see what happens. For example, suppose the oversized driver, balata ball, golf cart, and water combination is selected, and the resulting score is 87. During the round, however, I noticed several wayward shots with the big driver (long is not always good in golf), and, as a result, I decide to play another round with the regular-sized driver, holding the other factors at the same levels used previously. This approach could be continued almost indefinitely, switching the levels of one or two (or perhaps several) factors for the next test, based on the outcome of the current test. This strategy of experimentation, which we call the **best-guess approach**, is frequently used in practice by engineers and scientists. It often works reasonably well, too, because the experimenters often have a great deal of technical or theoretical knowledge of the system they are studying, as well as considerable practical experience. The best-guess approach has at least two disadvantages. First, suppose the initial best-guess does not produce the desired results. Now the experimenter has to take another guess at the correct combination of factor levels. This could continue for a long time, without any guarantee of success. Second, suppose the initial best-guess produces an acceptable result. Now the experimenter is tempted to stop testing, although there is no guarantee that the *best* solution has been found.

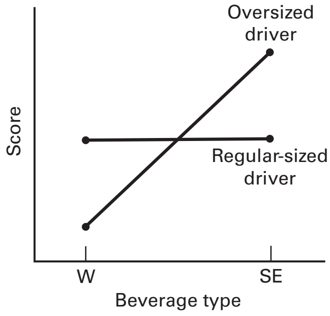
Another strategy of experimentation that is used extensively in practice is the **one-factor-at-a-time** (**OFAT**) approach. The OFAT method consists of selecting a starting point, or **baseline** set of levels, for each factor, and then successively varying each factor over its range with the other factors held constant at the baseline level. After all tests are performed, a series of graphs are usually constructed showing how the response variable is affected by varying each factor with all other factors held constant. Figure 1.2 shows a set of these graphs for the golf experiment, using the oversized driver, balata ball, walking, and drinking water levels of the four factors as the baseline. The interpretation of these graphs is straightforward; for example, because the slope of the mode of travel curve is negative, we would conclude that riding improves the score. Using these one-factor-at-a-time graphs, we would select the optimal combination to be the regular-sized driver, riding, and drinking water. The type of golf ball seems unimportant.

The major disadvantage of the OFAT strategy is that it fails to consider any possible **interaction** between the factors. An interaction is the failure of one factor to produce the same effect on the response at different levels of another factor. Figure 1.3 shows an interaction between the type of driver and the beverage factors for the golf experiment. Notice that if I use the regular-sized driver, the type of beverage consumed has virtually no effect on the score, but if I use the oversized driver, much better results are obtained by drinking water instead of "something else." Interactions between factors are very common, and if they occur, the one-factor-at-a-time strategy will usually produce poor results. Many people do not recognize this, and, consequently, OFAT experiments are run frequently in practice. (Some individuals actually think that this strategy is related to the scientific method or that it is a "sound" engineering principle.) One-factor-at-a-time experiments are always less efficient than other methods based on a statistical approach to design. We will discuss this in more detail in Chapter 5.

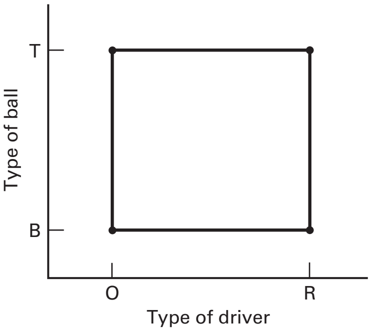
The correct approach to dealing with several factors is to conduct a **factorial** experiment. This is an experimental strategy in which factors are varied *together*, instead of one at a time. The factorial experimental design concept is



■ **FIGURE 1.2 Results of the one-factor-at-a-time strategy for the golf experiment**



■ **FIGURE 1.3** **Interaction between type of driver and type of beverage for the golf experiment**



■ **FIGURE 1.4** **A two-factor factorial experiment involving type of driver and type of ball**

extremely important, and several chapters in this book are devoted to presenting basic factorial experiments and a number of useful variations and special cases.

To illustrate how a factorial experiment is conducted, consider the golf experiment and suppose that only two factors, type of driver and type of ball, are of interest. Figure 1.4 shows a two-factor factorial experiment for studying the joint effects of these two factors on my golf score. Notice that this factorial experiment has both factors at two levels and that all possible combinations of the two factors across their levels are used in the design. Geometrically, the four runs form the corners of a square. This particular type of factorial experiment is called a **2^2 factorial design** (two factors, each at two levels). Because I can reasonably expect to play eight rounds of golf to investigate these factors, a reasonable plan would be to play two rounds of golf at each combination of factor levels shown in Figure 1.4. An experimental designer would say that we have **replicated** the design twice. This experimental design would enable the experimenter to investigate the individual effects of each factor (or the **main** effects) and to determine whether the factors interact.

Figure 1.5*a* shows the results of performing the factorial experiment in Figure 1.4. The scores from each round of golf played at the four test combinations are shown at the corners of the square. Notice that there are four rounds of golf that provide information about using the regular-sized driver and four rounds that provide information about using the oversized driver. By finding the average difference in the scores on the right- and left-hand sides of the square (as in Figure 1.5*b*), we have a measure of the effect of switching from the oversized driver to the regular-sized driver, or

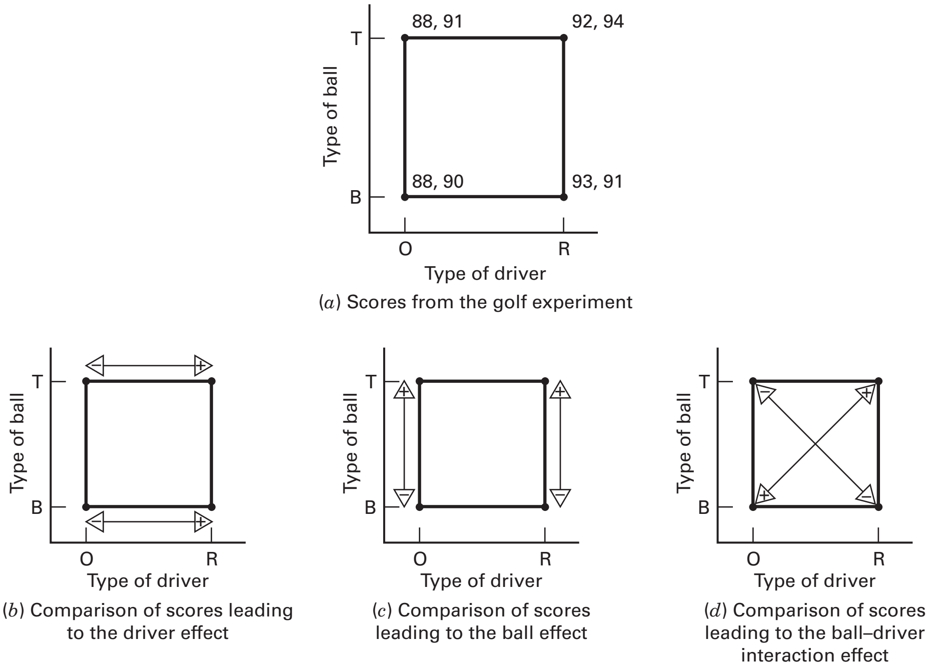
$$\begin{aligned}\text{Driver effect} &= \frac{92+94+93+91}{4} - \frac{88+91+88+90}{4} \\ &= 3.25\end{aligned}$$

That is, on average, switching from the oversized to the regular-sized driver increases the score by 3.25 strokes per round. Similarly, the average difference in the four scores at the top of the square and the four scores at the bottom measures the effect of the type of ball used (see Figure 1.5*c*):

$$\begin{aligned}\text{Ball effect} &= \frac{88+91+92+94}{4} - \frac{88+90+93+91}{4} \\ &= 0.75\end{aligned}$$

Finally, a measure of the interaction effect between the type of ball and the type of driver can be obtained by subtracting the average scores on the left-to-right diagonal in the square from the average scores on the right-to-left diagonal (see Figure 1.5*d*), resulting in

$$\begin{aligned}\text{Ball–driver interaction effect} &= \frac{92+94+88+90}{4} - \frac{88+91+93+91}{4} \\ &= 0.25\end{aligned}$$


88, 91
92, 94
88, 90
93, 91
T
B
Type of ball
O
R
Type of driver
(a) Scores from the golf experiment
(b) Comparison of scores leading to the driver effect
(c) Comparison of scores leading to the ball effect
(d) Comparison of scores leading to the ball–driver interaction effect
+
−

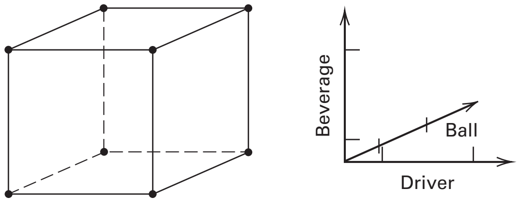
■ **FIGURE 1.5** **Scores from the golf experiment in Figure 1.4 and calculation of the factor effects**

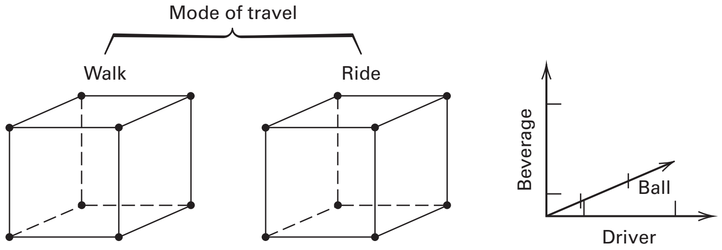
The results of this factorial experiment indicate that driver effect is larger than either the ball effect or the interaction. Statistical testing could be used to determine whether any of these effects differ from zero. In fact, it turns out that there is reasonably strong statistical evidence that the driver effect differs from zero and the other two effects do not. Therefore, this experiment indicates that I should always play with the oversized driver.

One very important feature of the factorial experiment is evident from this simple example; namely, factorials make the most efficient use of the experimental data. Notice that this experiment included eight observations, and all eight observations are used to calculate the driver, ball, and interaction effects. No other strategy of experimentation makes such an efficient use of the data. This is an important and useful feature of factorials.

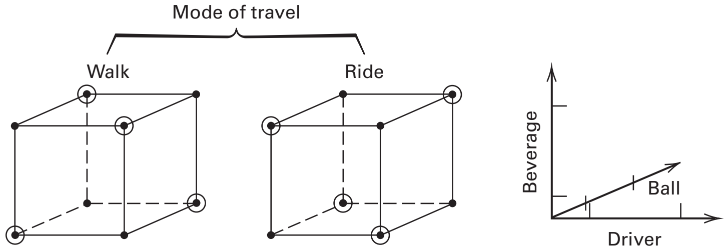
We can extend the factorial experiment concept to three factors. Suppose that I wish to study the effects of type of driver, type of ball, and the type of beverage consumed on my golf score. Assuming that all three factors have two levels, a factorial design can be set up as shown in Figure 1.6. Notice that there are eight test combinations of these three factors across the two levels of each and that these eight trials can be represented geometrically as the corners of a cube. This is an example of a **2^3 factorial design**. Because I only want to play eight rounds of golf, this experiment would require that one round be played at each combination of factors represented by the eight corners of the cube in Figure 1.6. However, if we compare this to the two-factor factorial in Figure 1.4, the 2^3 factorial design would provide the same information about the factor effects. For example, there are four tests in both designs that provide information about the regular-sized driver and four tests that provide information about the oversized driver, assuming that each run in the two-factor design in Figure 1.4 is replicated twice.

■ **FIGURE 1.6** **A three-factor factorial experiment involving type of driver, type of ball, and type of beverage**





■ **FIGURE 1.7 A four-factor factorial experiment involving type of driver, type of ball, type of beverage, and mode of travel**



■ **FIGURE 1.8 A four-factor fractional factorial experiment involving type of driver, type of ball, type of beverage, and mode of travel**

Figure 1.7 illustrates how all four factors—driver, ball, beverage, and mode of travel (walking or riding)—could be investigated in a **2^4 factorial design**. As in any factorial design, all possible combinations of the levels of the factors are used. Because all four factors are at two levels, this experimental design can still be represented geometrically as a cube (actually a hypercube).

Generally, if there are k factors, each at two levels, the factorial design would require 2^k runs. For example, the experiment in Figure 1.7 requires 16 runs. Clearly, as the number of factors of interest increases, the number of runs required increases rapidly; for instance, a 10-factor experiment with all factors at two levels would require 1024 runs. This quickly becomes infeasible from a time and resource viewpoint. In the golf experiment, I can only play eight rounds of golf, so even the experiment in Figure 1.7 is too large.

Fortunately, if there are four to five or more factors, it is usually unnecessary to run all possible combinations of factor levels. A **fractional factorial experiment** is a variation of the basic factorial design in which only a subset of the runs is used. Figure 1.8 shows a fractional factorial design for the four-factor version of the golf experiment. This design requires only 8 runs instead of the original 16 and would be called a **one-half fraction**. If I can play only eight rounds of golf, this is an excellent design in which to study all four factors. It will provide good information about the main effects of the four factors as well as some information about how these factors interact.

Fractional factorial designs are used extensively in industrial research and development, and for process improvement. These designs will be discussed in Chapters 8 and 9.

1.2 Some Typical Applications of Experimental Design

Experimental design methods have found broad application in many disciplines. As noted previously, we may view experimentation as part of the scientific process and as one of the ways by which we learn about how systems or processes work. Generally, we learn through a series of activities in which we make conjectures about a process, perform experiments to generate data from the process, and then use the information from the experiment to establish new conjectures, which lead to new experiments, and so on.

Experimental design is a critically important tool in the scientific and engineering world for driving innovation in the product realization process. Critical components of these activities are in new manufacturing process design and development and process management. The application of experimental design techniques early in process development can result in

1. Improved process yields
2. Reduced variability and closer conformance to nominal or target requirements
3. Reduced development time
4. Reduced overall costs.

Experimental design methods are also of fundamental importance in **engineering design** activities, where new products are developed and existing ones improved. Some applications of experimental design in engineering design include

1. Evaluation and comparison of basic design configurations
2. Evaluation of material alternatives
3. Selection of design parameters so that the product will work well under a wide variety of field conditions, that is, so that the product is **robust**
4. Determination of key product design parameters that impact product performance
5. Formulation of new products.

The use of experimental design in product realization can result in products that are easier to manufacture and that have enhanced field performance and reliability, lower product cost, and shorter product design and development time. Designed experiments also have extensive applications in marketing, market research, transactional and service operations, and general business operations. We now present several examples that illustrate some of these ideas.

EXAMPLE 1.1 Characterizing a Process

A flow solder machine is used in the manufacturing process for printed circuit boards. The machine cleans the boards in a flux, preheats the boards, and then moves them along a conveyor through a wave of molten solder. This solder process makes the electrical and mechanical connections for the leaded components on the board.

The process currently operates around the 1 percent defective level. That is, about 1 percent of the solder joints on a board are defective and require manual retouching. However, because the average printed circuit board contains over 2000 solder joints, even a 1 percent defective level results in far too many solder joints requiring rework. The process engineer responsible for this area would like to use a designed experiment to determine which machine parameters are influential in the occurrence of solder defects and which adjustments should be made to those variables to reduce solder defects.

The flow solder machine has several variables that can be controlled. They include

1. Solder temperature
2. Preheat temperature
3. Conveyor speed
4. Flux type
5. Flux specific gravity
6. Solder wave depth
7. Conveyor angle.

In addition to these controllable factors, several other factors cannot be easily controlled during routine manufacturing, although they could be controlled for the purposes of a test. They are

1. Thickness of the printed circuit board
2. Types of components used on the board
3. Layout of the components on the board
4. Operator
5. Production rate.

In this situation, engineers are interested in **characterizing** the flow solder machine; that is, they want to determine which factors (both controllable and uncontrollable) affect the occurrence of defects on the printed circuit boards. To accomplish this, they can design an experiment that will enable them to estimate the magnitude and direction of the factor effects; that is, how much does the response variable (defects per unit) change when each factor is changed, and does changing the factors *together* produce different results than are obtained from individual factor adjustments—that is, do the factors interact? Sometimes we call an experiment such as this a **screening experiment**.

Typically, screening or characterization experiments involve using fractional factorial designs, such as in the golf example in Figure 1.8.

The information from this screening or characterization experiment will be used to identify the critical process factors and to determine the direction of adjustment for these factors to reduce further the number of defects per unit. The experiment may also provide information about which factors should be more carefully controlled during routine manufacturing to prevent high defect levels and erratic process performance. Thus, one result of the experiment could be the application of techniques such as control charts to one or more **process variables** (such as solder temperature), in addition to control charts on process output. Over time, if the process is improved enough, it may be possible to base most of the process control plan on controlling process input variables instead of control charting the output.

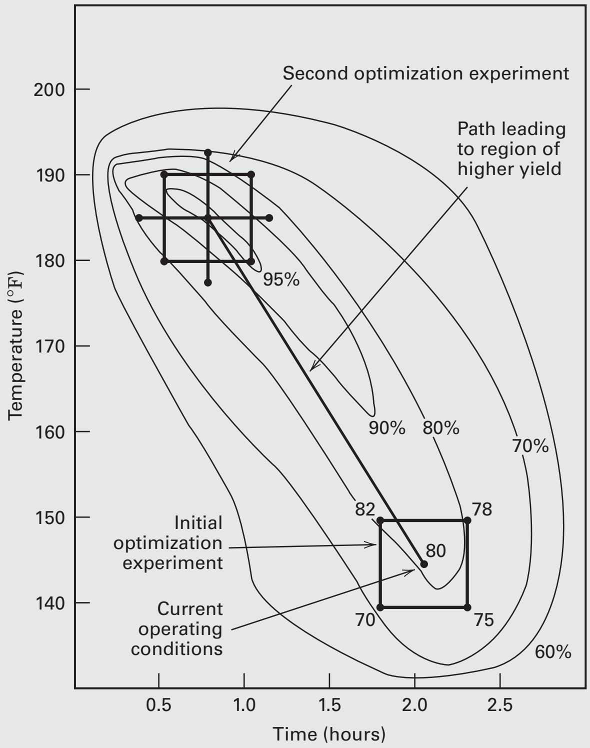
EXAMPLE 1.2 Optimizing a Process

In a characterization experiment, we are usually interested in determining which process variables affect the response. A logical next step is to optimize, that is, to determine the region in the important factors that leads to the best possible response. For example, if the response is yield, we would look for a region of maximum yield, whereas if the response is variability in a critical product dimension, we would seek a region of minimum variability.

Suppose that we are interested in improving the yield of a chemical process. We know from the results of a characterization experiment that the two most important process variables that influence the yield are operating temperature and reaction time. The process currently runs at 145°F and 2.1 hours of reaction time, producing yields of around 80 percent. Figure 1.9 shows a view of the time–temperature region from above. In this graph, the lines of constant yield are connected to form response **contours**, and we have shown the contour lines for yields of 60, 70, 80, 90, and 95 percent. These contours are projections on the time–temperature region of cross sections of the yield surface corresponding to the aforementioned percent yields. This surface is sometimes called a **response surface**. The true response surface in Figure 1.9 is unknown to the process personnel, so experimental methods will be required to optimize the yield with respect to time and temperature.

To locate the optimum, it is necessary to perform an experiment that varies both time and temperature together, that is, a factorial experiment. The results of an initial factorial experiment with both time and temperature run at two levels is shown in Figure 1.9. The responses observed at the four corners of the square indicate that we should move in the general direction of increased temperature and decreased reaction time to increase yield. A few additional runs would be performed in this direction, and this additional experimentation would lead us to the region of maximum yield.

Once we have found the region of the optimum, a second experiment would typically be performed. The objective of this second experiment is to develop an empirical model of the process and to obtain a more precise estimate of the optimum operating conditions for time and temperature. This approach to process optimization is called **response surface methodology**, and it is explored in detail in Chapter 11. The second design illustrated in Figure 1.9 is a **central composite design**, one of the most important experimental designs used in process optimization studies.



■ **FIGURE 1.9 Contour plot of yield as a function of reaction time and reaction temperature, illustrating experimentation to optimize a process**

EXAMPLE 1.3 Designing a Product—I

A biomedical engineer is designing a new pump for the intravenous delivery of a drug. The pump should deliver a constant quantity or dose of the drug over a specified period of time. She must specify a number of variables or design parameters. Among these are the diameter and length of the cylinder, the fit between the cylinder and the plunger, the plunger length, the diameter and wall thickness of the tube connecting the pump and the needle inserted into the patient's vein, the material to use for fabricating both the cylinder and the tube, and the nominal pressure at which the system must operate. The impact of some of these parameters on the design can be evaluated by building prototypes in which these factors can be varied over appropriate ranges. Experiments can then be designed and the prototypes tested to investigate which design parameters are most influential on pump performance. Analysis of this information will assist the engineer in arriving at a design that provides reliable and consistent drug delivery.

EXAMPLE 1.4 Designing a Product—II

An engineer is designing an aircraft engine. The engine is a commercial turbofan, intended to operate in the cruise configuration at 40,000 ft and 0.8 Mach. The design parameters include inlet flow, fan pressure ratio, overall pressure, stator outlet temperature, and many other factors. The output response variables in this system are specific fuel consumption and engine thrust. In designing this system, it would be prohibitive to build prototypes or actual test articles early in the design process, so the engineers use a **computer model** of the system that allows them to focus on the key design parameters of the engine and to vary them in an effort to optimize the performance of the engine. Designed experiments can be employed with the computer model of the engine to determine the most important design parameters and their optimal settings.

Designers frequently use computer models to assist them in carrying out their activities. Examples include finite element models for many aspects of structural and mechanical design, electrical circuit simulators for integrated circuit design, factory or enterprise-level models for scheduling and capacity planning or supply chain management, and computer models of complex chemical processes. Statistically designed experiments can be applied to these models just as easily and successfully as they can to actual physical systems and will result in reduced development lead time and better designs.

EXAMPLE 1.5 Formulating a Product

A biochemist is formulating a diagnostic product to detect the presence of a certain disease. The product is a mixture of biological materials, chemical reagents, and other materials that when combined with human blood react to provide a diagnostic indication. The type of experiment used here is a **mixture experiment**, because various ingredients that are combined to form the diagnostic make up 100 percent of the mixture composition (on a volume, weight, or mole ratio basis), and the response is a function of the mixture proportions that are present in the product. Mixture experiments are a special type of response surface experiment that we will study in Chapter 11. They are very useful in designing biotechnology products, pharmaceuticals, foods and beverages, paints and coatings, consumer products such as detergents, soaps, and other personal care products, and a wide variety of other products.

EXAMPLE 1.6 Designing a Web Page

A lot of business today is conducted via the World Wide Web. Consequently, the design of a business' web page has potentially important economic impact. Suppose that the website has the following components: (1) a photoflash image, (2) a main headline, (3) a subheadline, (4) a main text copy, (5) a main image on the right side, (6) a background design, and (7) a footer. We are interested in finding the factors that influence the click-through rate; that is, the number of visitors who click through into the site divided by the total number of visitors to the site. Proper selection of the important factors can lead to an optimal web page design. Suppose that there are four choices for the photoflash image, eight choices for the main headline, six choices for the subheadline, five choices for the main text copy, four choices for the main image, three choices for the background design, and seven choices for the footer. If we use a factorial design, web pages for all possible combinations of these factor levels must be constructed and tested. This is a total of $4 \times 8 \times 6 \times 5 \times 4 \times 3 \times 7 = 80{,}640$ web pages. Obviously, it is not feasible to design and test this many combinations of web pages, so a complete factorial experiment cannot be considered. However, a fractional factorial experiment that uses a small number of the possible web page designs would likely be successful. This experiment would require a fractional factorial where the factors have different numbers of levels. We will discuss how to construct these designs in Chapter 9.

1.3 Basic Principles

If an experiment such as the ones described in Examples 1.1 through 1.6 is to be performed most efficiently, a scientific approach to planning the experiment must be employed. **Statistical design of experiments** refers to the process of planning the experiment so that appropriate data will be collected and analyzed by statistical methods, resulting in valid and objective conclusions. The statistical approach to experimental design is necessary if we wish to draw meaningful conclusions from the data. When the problem involves data that are subject to experimental errors, statistical methods are the only **objective** approach to analysis. Thus, there are two aspects to any experimental problem: the design of the experiment and the statistical analysis of the data. These two subjects are closely related because the method of analysis depends directly on the design employed. Both topics will be addressed in this book.

The three basic principles of experimental design are **randomization, replication**, and **blocking**. Sometimes we add the **factorial principle** to these three. Randomization is the cornerstone underlying the use of statistical methods in experimental design. By randomization we mean that both the allocation of the experimental material and the order in which the individual runs of the experiment are to be performed are randomly determined. Statistical methods require that the observations (or errors) be independently distributed random variables. Randomization usually makes this assumption valid. By properly randomizing the experiment, we also assist in "averaging out" the effects of extraneous factors that may be present. For example, suppose that the specimens in the hardness experiment are of slightly different thicknesses and that the effectiveness of the quenching medium may be affected by specimen thickness. If all the specimens subjected to the oil quench are thicker than those subjected to the saltwater quench, we may be introducing systematic bias into the experimental results. This bias handicaps one of the quenching media and consequently invalidates our results. Randomly assigning the specimens to the quenching media alleviates this problem.

Computer software programs are widely used to assist experimenters in selecting and constructing experimental designs. These programs often present the runs in the experimental design in random order. This random order is created by using a random number generator. Even with such a computer program, it is still often necessary to assign units of experimental material (such as the specimens in the hardness example mentioned above), operators, gauges or measurement devices, and so forth for use in the experiment.

Sometimes experimenters encounter situations where randomization of some aspect of the experiment is difficult. For example, in a chemical process, temperature may be a very hard-to-change variable as we may want to change it less often than we change the levels of other factors. In an experiment of this type, **complete randomization** would be difficult because it would add time and cost. There are statistical design methods for dealing with restrictions on randomization. Some of these approaches will be discussed in subsequent chapters (see in particular Chapter 14).

By **replication** we mean an independent **repeat run** of each factor combination. In the metallurgical experiment discussed in Section 1.1, replication would consist of treating a specimen by oil quenching and treating a specimen by saltwater quenching. Thus, if five specimens are treated in each quenching medium, we say that five **replicates** have been obtained. Each of the 10 observations should be run in random order. Replication has two important properties. First, it allows the experimenter to obtain an estimate of the experimental error. This estimate of error becomes a basic unit of measurement for determining whether observed differences in the data are really *statistically* different. Second, if the sample mean ($\overline{y}$) is used to estimate the true mean response for one of the factor levels in the experiment, replication permits the experimenter to obtain a more precise estimate of this parameter. For example, if σ^2 is the variance of an individual observation and there are n replicates, the variance of the sample mean is

$$\sigma_{\overline{y}}^2 = \frac{\sigma^2}{n}$$

The practical implication of this is that if we had $n = 1$ replicates and observed $y_1 = 145$ (oil quench) and $y_2 = 147$ (saltwater quench), we would probably be unable to make satisfactory inferences about the effect of the quenching medium—that is, the observed difference could be the result of experimental error. The point is that without replication we have no way of knowing why the two observations are different. On the other hand, if n was reasonably large and the experimental error was sufficiently small and if we observed sample averages $\overline{y}_1 < \overline{y}_2$, we would be reasonably safe in concluding that saltwater quenching produces a higher hardness in this particular aluminum alloy than does oil quenching.

Often when the runs in an experiment are randomized, two (or more) consecutive runs will have exactly the same levels for some of the factors. For example, suppose we have three factors in an experiment: pressure, temperature, and time. When the experimental runs are randomized, we find the following:

Run number	**Pressure (psi)**	**Temperature (°C)**	**Time (min)**
i	30	100	30
$i+1$	30	125	45
$i+2$	40	125	45

Notice that between runs i and $i + 1$, the levels of pressure are identical and between runs $i + 1$ and $i + 2$, the levels of both temperature and time are identical. To obtain a true replicate, the experimenter needs to "twist the pressure knob" to an intermediate setting between runs i and $i + 1$, and reset pressure to 30 psi for run $i + 1$. Similarly, temperature and time should be reset to intermediate levels between runs $i + 1$ and $i + 2$ before being set to their design levels for run $i + 2$. Part of the experimental error is the variability associated with hitting and holding factor levels.

There is an important distinction between **replication** and **repeated measurements**. For example, suppose that a silicon wafer is etched in a single-wafer plasma etching process, and a critical dimension (CD) on this wafer is measured three times. These measurements are not replicates; they are a form of repeated measurements, and in this case the observed variability in the three repeated measurements is a direct reflection of the inherent variability in the measurement system or gauge and possibly the variability in this CD at different locations on the wafer where the measurements were taken. As another illustration, suppose that as part of an experiment in semiconductor manufacturing four wafers are processed simultaneously in an oxidation furnace at a particular gas flow rate and time and then a measurement is taken on the oxide thickness of each wafer. Once again, the measurements on the four wafers are not replicates but repeated measurements. In this case, they reflect differences among the wafers and other sources of variability within that particular furnace run. Replication reflects sources of variability both **between** runs and (potentially) **within** runs.

Blocking is a design technique used to improve the precision with which comparisons among the factors of interest are made. Often blocking is used to reduce or eliminate the variability transmitted from **nuisance factors**—that is, factors that may influence the experimental response but in which we are not directly interested. For example, an experiment in a chemical process may require two batches of raw material to make all the required runs.

However, there could be differences between the batches due to supplier-to-supplier variability, and if we are not specifically interested in this effect, we would think of the batches of raw material as a nuisance factor. Generally, a block is a set of relatively homogeneous experimental conditions. In the chemical process example, each batch of raw material would form a block, because the variability within a batch would be expected to be smaller than the variability between batches. Typically, as in this example, each level of the nuisance factor becomes a block. Then the experimenter divides the observations from the statistical design into groups that are run in each block. We study blocking in detail in several places in the text, including Chapters 4, 5, 7, 8, 9, 11, and 13. A simple example illustrating the blocking principal is given in Section 2.5.1.

The three basic principles of experimental design, randomization, replication, and blocking are part of every experiment. We will illustrate and emphasize them repeatedly throughout this book.

1.4 Guidelines for Designing Experiments

To use the statistical approach in designing and analyzing an experiment, it is necessary for everyone involved in the experiment to have a clear idea in advance of exactly what is to be studied, how the data are to be collected, and at least a qualitative understanding of how these data are to be analyzed. An outline of the recommended procedure is shown in Table 1.1. We now give a brief discussion of this outline and elaborate on some of the key points. For more details, see Coleman and Montgomery (1993), and the references therein. The **supplemental text material** for this chapter is also useful.

1. ***Recognition of and statement of the problem.*** This may seem to be a rather obvious point, but in practice often neither is it simple to realize that a problem requiring experimentation exists, nor is it simple to develop a clear and generally accepted statement of this problem. It is necessary to develop all ideas about the objectives of the experiment. Usually, it is important to solicit input from all concerned parties: engineering, quality assurance, manufacturing, marketing, management, customer, and operating personnel (who usually have much insight and who are too often ignored). For this reason, a **team approach** to designing experiments is recommended.

   It is usually helpful to prepare a list of specific problems or questions that are to be addressed by the experiment. A clear statement of the problem often contributes substantially to better understanding of the phenomenon being studied and the final solution of the problem.

   It is also important to keep the overall objectives of the experiment in mind. There are several broad reasons for running experiments and each type of experiment will generate its own list of specific questions that need to be addressed. Some (but by no means all) of the reasons for running experiments include:

   a. ***Factor screening or characterization.*** When a system or process is new, it is usually important to learn which factors have the most influence on the response(s) of interest. Often there are a lot of factors. This usually indicates that the experimenters do not know much about the system

■ **TABLE 1.1**
Guidelines for Designing an Experiment

1. Recognition of and statement of the problem	Pre-experimental
2. Selection of the response variable[a]	Planning
3. Choice of factors, levels, and ranges[a]	
4. Choice of experimental design	
5. Performing the experiment	
6. Statistical analysis of the data	
7. Conclusions and recommendations	

[a] In practice, steps 2 and 3 are often done simultaneously or in reverse order.

so screening is essential if we are to efficiently get the desired performance from the system. Screening experiments are extremely important when working with new systems or technologies so that valuable resources will not be wasted using best guess and OFAT approaches.

b. ***Optimization.*** After the system has been characterized and we are reasonably certain that the important factors have been identified, the next objective is usually optimization, that is, find the settings or levels of the important factors that result in desirable values of the response. For example, if a screening experiment on a chemical process results in the identification of time and temperature as the two most important factors, the optimization experiment may have as its objective finding the levels of time and temperature that maximize yield, or perhaps maximize yield while keeping some product property that is critical to the customer within specifications. An optimization experiment is usually a follow-up to a screening experiment. It would be very unusual for a screening experiment to produce the optimal settings of the important factors.

c. ***Confirmation.*** In a confirmation experiment, the experimenter is usually trying to verify that the system operates or behaves in a manner that is consistent with some theory or past experience. For example, if theory or experience indicates that a particular new material is equivalent to the one currently in use and the new material is desirable (perhaps less expensive, or easier to work with in some way), then a confirmation experiment would be conducted to verify that substituting the new material results in no change in product characteristics that impact its use. Moving a new manufacturing process to full-scale production based on results found during experimentation at a pilot plant or development site is another situation that often results in confirmation experiments—that is, are the same factors and settings that were determined during development work appropriate for the full-scale process?

d. ***Discovery.*** In discovery experiments, the experimenters are usually trying to determine what happens when we explore new materials, or new factors, or new ranges for factors. Discovery experiments often involve screening of several (perhaps many) factors. In the pharmaceutical industry, scientists are constantly conducting discovery experiments to find new materials or combinations of materials that will be effective in treating disease.

e. ***Robustness.*** These experiments often address questions such as under what conditions do the response variables of interest seriously degrade? Or what conditions would lead to unacceptable variability in the response variables? A variation of this is determining how we can set the factors in the system that we can control to minimize the variability transmitted into the response from factors that we cannot control very well. We will discuss some experiments of this type in Chapter 12.

Obviously, the specific questions to be addressed in the experiment relate directly to the overall objectives. An important aspect of problem formulation is the recognition that one large comprehensive experiment is unlikely to answer the key questions satisfactorily. A single comprehensive experiment requires the experimenters to know the answers to a lot of questions, and if they are wrong, the results will be disappointing. This leads to wasting time, materials, and other resources and may result in never answering the original research questions satisfactorily. A **sequential** approach employing a series of smaller experiments, each with a specific objective, such as factor screening, is a better strategy.

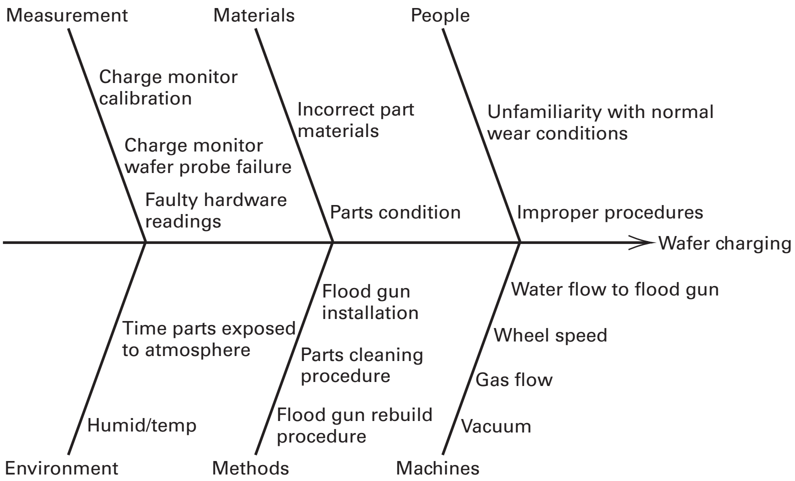
2. ***Selection of the response variable.*** In selecting the response variable, the experimenter should be certain that this variable really provides useful information about the process under study. Most often, the average or standard deviation (or both) of the measured characteristic will be the response variable. Multiple responses are not unusual. The experimenters must decide how each response will be measured, and address issues such as how will any measurement system be calibrated and how this calibration will be maintained during the experiment. The gauge or measurement system capability (or measurement error) is also an important factor. If gauge capability is inadequate, only relatively large factor effects will be detected by the experiment or perhaps additional replication will be required. In some situations where gauge capability is poor, the experimenter may decide to measure each experimental unit several times and use the average of the repeated measurements as the observed response. It is usually critically important to identify issues related to defining the responses of interest and how they are to be measured *before* conducting the experiment. Sometimes designed experiments are employed to study and improve the performance of measurement systems. For an example, see Chapter 13.

3. ***Choice of factors, levels, and range.*** (As noted in Table 1.1, steps 2 and 3 are often done simultaneously or in the reverse order.) When considering the factors that may influence the performance of a process or system, the experimenter usually discovers that these factors can be classified as either **potential design factors** or nuisance factors. The potential design factors are those factors that the experimenter may wish to vary in the experiment. Often we find that there are a lot of potential design factors, and some further classification of them is helpful. Some useful classifications are **design factors, held-constant factors**, and **allowed-to-vary** factors. The design factors are the factors actually selected for study in the experiment. Held-constant factors are variables that may exert some effect on the response, but for purposes of the present experiment these factors are not of interest, so they will be held at a specific level. For example, in an etching experiment in the semiconductor industry, there may be an effect that is unique to the specific plasma etch tool used in the experiment. However, this factor would be very difficult to vary in an experiment, so the experimenter may decide to perform all experimental runs on one particular (ideally "typical") etcher. Thus, this factor has been held constant. As an example of allowed-to-vary factors, the experimental units or the "materials" to which the design factors are applied are usually nonhomogeneous, yet we often ignore this unit-to-unit variability and rely on randomization to balance out any material or experimental unit effect. We often assume that the effects of held-constant factors and allowed-to-vary factors are relatively small.

   Nuisance factors, on the other hand, may have large effects that must be accounted for, yet we may not be interested in them in the context of the present experiment. Nuisance factors are often classified as **controllable, uncontrollable**, or **noise factors**. A controllable nuisance factor is one whose levels may be set by the experimenter. For example, the experimenter can select different batches of raw material or different days of the week when conducting the experiment. The blocking principle, discussed in the previous section, is often useful in dealing with controllable nuisance factors. If a nuisance factor is uncontrollable in the experiment, but it can be measured, an analysis procedure called the **analysis of covariance** can often be used to compensate for its effect. For example, the relative humidity in the process environment may affect process performance, and if the humidity cannot be controlled, it probably can be measured and treated as a covariate. When a factor that varies naturally and uncontrollably in the process can be controlled for purposes of an experiment, we often call it a noise factor. In such situations, our objective is usually to find the settings of the controllable design factors that minimize the variability transmitted from the noise factors. This is sometimes called a process robustness study or a robust design problem. Blocking, analysis of covariance, and process robustness studies are discussed later in the text.

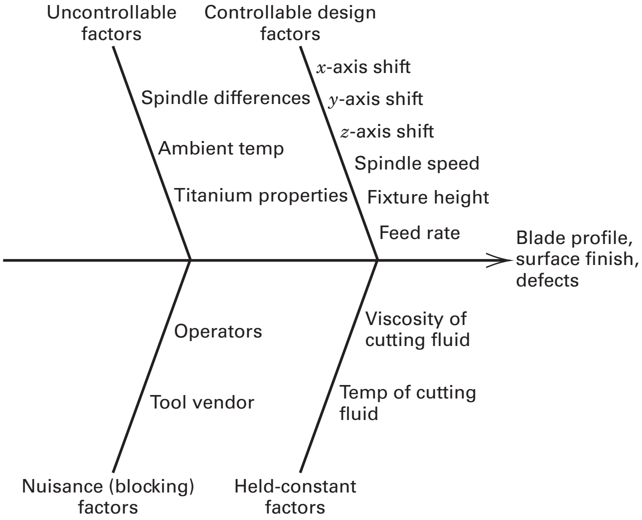
   Once the experimenter has selected the design factors, he or she must choose the ranges over which these factors will be varied and the specific levels at which runs will be made. Thought must also be given to how these factors are to be controlled at the desired values and how they are to be measured. For instance, in the flow solder experiment, the engineer has defined 12 variables that may affect the occurrence of solder defects. The experimenter will also have to decide on a region of interest for each variable (that is, the range over which each factor will be varied) and on how many levels of each variable to use. **Process knowledge** is required to do this. This process knowledge is usually a combination of practical experience and theoretical understanding. It is important to investigate all factors that may be of importance and to be not overly influenced by past experience, particularly when we are in the early stages of experimentation or when the process is not very mature.

   When the objective of the experiment is **factor screening** or **process characterization**, it is usually best to keep the number of factor levels low. Generally, two levels work very well in factor screening studies. Choosing the region of interest is also important. In factor screening, the region of interest should be relatively large—that is, the range over which the factors are varied should be broad. As we learn more about which variables are important and which levels produce the best results, the region of interest in subsequent experiments will usually become narrower.



■ **FIGURE 1.10 A cause-and-effect diagram for the etching process experiment**

The **cause-and-effect diagram** can be a useful technique for organizing some of the information generated in pre-experimental planning. Figure 1.10 is the cause-and-effect diagram constructed while planning an experiment to resolve problems with wafer charging (a charge accumulation on the wafers) encountered in an etching tool used in semiconductor manufacturing. The cause-and-effect diagram is also known as a **fishbone diagram** because the “effect” of interest or the response variable is drawn along the spine of the diagram and the potential causes or design factors are organized in a series of ribs. The cause-and-effect diagram uses the traditional causes of measurement, materials, people, environment, methods, and machines to organize the information and potential design factors. Notice that some of the individual causes will probably lead directly to a design factor that will be included in the experiment (such as wheel speed, gas flow, and vacuum), while others represent potential areas that will need further study to turn them into design factors (such as operators following improper procedures), and still others will probably lead to either factors that will be held constant during the experiment or blocked (such as temperature and relative humidity). Figure 1.11 is a cause-and-effect diagram for an experiment to study the effect of several factors on the turbine blades produced on a computer-numerical-controlled (CNC) machine. This experiment has three response



■ **FIGURE 1.11 A cause-and-effect diagram for the CNC machine experiment** variables: blade profile, blade surface finish, and surface finish defects in the finished blade. The causes are organized into groups of controllable factors from which the design factors for the experiment may be selected, uncontrollable factors whose effects will probably be balanced out by randomization, nuisance factors that may be blocked, and factors that may be held constant when the experiment is conducted. It is not unusual for experimenters to construct several different cause-and-effect diagrams to assist and guide them during pre-experimental planning. For more information on the CNC machine experiment and further discussion of graphical methods that are useful in pre-experimental planning, see the supplemental text material for this chapter.

We reiterate how crucial it is to bring out all points of view and process information in steps 1 through 3. We refer to this as **pre-experimental planning**. Coleman and Montgomery (1993) provide worksheets that can be useful in pre-experimental planning. Also see the **supplemental text material** for more details and an example of using these worksheets. It is unlikely that one person has all the knowledge required to do this adequately in many situations. Therefore, we strongly argue for a team effort in planning the experiment. Most of your success will hinge on how well the pre-experimental planning is done.

4. ***Choice of experimental design.*** If the above pre-experimental planning activities are done correctly, this step is relatively easy. Choice of design involves consideration of sample size (number of replicates), selection of a suitable run order for the experimental trials, and determination of whether or not blocking or other randomization restrictions are involved. This book discusses some of the more important types of experimental designs, and it can ultimately be used as a guide for selecting an appropriate experimental design for a wide variety of problems.

There are also several interactive statistical software packages that support this phase of experimental design. The experimenter can enter information about the number of factors, levels, and ranges, and these programs will either present a selection of designs for consideration or recommend a particular design. (We usually prefer to see several alternatives instead of relying entirely on a computer recommendation in most cases.) Most software packages also provide some diagnostic information about how each design will perform. This is useful in evaluation of different design alternatives for the experiment. These programs will usually also provide a worksheet (with the order of the runs randomized) for use in conducting the experiment.

Design selection also involves thinking about and selecting a tentative **empirical model** to describe the results. The model is just a quantitative relationship (equation) between the response and the important design factors. In many cases, a low-order polynomial model will be appropriate. A **first-order** model in two variables is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where y is the response, the x's are the design factors, the β's are unknown parameters that will be estimated from the data in the experiment, and ε is a random error term that accounts for the experimental error in the system that is being studied. The first-order model is also sometimes called a **main effects** model. First-order models are used extensively in screening or characterization experiments. A common extension of the first-order model is to add an **interaction** term, say

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

where the cross-product term $x_1 x_2$ represents the two-factor interaction between the design factors. Because interactions between factors is relatively common, the first-order model with interaction is widely used. Higher-order interactions can also be included in experiments with more than two factors if necessary. Another widely used model is the **second-order** model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_{11}^2 + \beta_{22} x_2^2 + \varepsilon$$

Second-order models are often used in optimization experiments.

In selecting the design, it is important to keep the experimental objectives in mind. In many engineering experiments, we already know at the outset that some of the factor levels will result in different values for the response. Consequently, we are interested in identifying *which* factors cause this difference and in estimating the *magnitude* of the response change. In other situations, we may be more interested in verifying uniformity. For example, two production conditions A and B may be compared, A being the standard and B being a more cost-effective alternative. The experimenter will then be interested in demonstrating that, say, there is no difference in yield between the two conditions.

5. ***Performing the experiment.*** When running the experiment, it is vital to monitor the process carefully to ensure that everything is being done according to plan. Errors in experimental procedure at this stage will usually destroy experimental validity. One of the most common mistakes that I have encountered is that the people conducting the experiment failed to set the variables to the proper levels on some runs. Someone should be assigned to check factor settings before each run. Up-front planning to prevent mistakes like this is crucial to success. It is easy to underestimate the logistical and planning aspects of running a designed experiment in a complex manufacturing or research and development environment.

   Coleman and Montgomery (1993) suggest that prior to conducting the experiment a few trial runs or pilot runs are often helpful. These runs provide information about consistency of experimental material, a check on the measurement system, a rough idea of experimental error, and a chance to practice the overall experimental technique. This also provides an opportunity to revisit the decisions made in steps 1–4, if necessary.

6. ***Statistical analysis of the data.*** Statistical methods should be used to analyze the data so that results and conclusions are **objective** rather than judgmental in nature. If the experiment has been designed correctly and performed according to the design, the statistical methods required are not elaborate. There are many excellent software packages designed to assist in data analysis, and many of the programs used in step 4 to select the design provide a seamless, direct interface to the statistical analysis. Often we find that simple **graphical methods** play an important role in data analysis and interpretation. Because many of the questions that the experimenter wants to answer can be cast into an hypothesis-testing framework, hypothesis testing and confidence interval estimation procedures are very useful in analyzing data from a designed experiment. It is also usually very helpful to present the results of many experiments in terms of an **empirical model**, that is, an equation derived from the data that express the relationship between the response and the important design factors. Residual analysis and model adequacy checking are also important analysis techniques. We will discuss these issues in detail later.

   Remember that statistical methods cannot prove that a factor (or factors) has a particular effect. They only provide guidelines as to the reliability and validity of results. When properly applied, statistical methods do not allow anything to be proved experimentally, but they do allow us to measure the likely error in a conclusion or to attach a level of confidence to a statement. The primary advantage of statistical methods is that they add **objectivity** to the decision-making process. Statistical techniques coupled with good engineering or process knowledge and common sense will usually lead to sound conclusions.

7. ***Conclusions and recommendations.*** Once the data have been analyzed, the experimenter must draw *practical* conclusions about the results and recommend a course of action. Graphical methods are often useful in this stage, particularly in presenting the results to others. **Follow-up runs** and **confirmation testing** should also be performed to validate the conclusions from the experiment.

   Throughout this entire process, it is important to keep in mind that experimentation is an important part of the learning process, where we tentatively formulate hypotheses about a system, perform experiments to investigate these hypotheses, and on the basis of the results formulate new hypotheses, and so on. This suggests that experimentation is **iterative**. It is usually a major mistake to design a single, large, comprehensive experiment at the start of a study. A successful experiment requires knowledge of the important factors, the ranges over which these factors should be varied, the appropriate number of levels to use, and the proper units of measurement for these variables. Generally, we do not perfectly know the answers to these questions, but we learn about them as we go along. As an experimental program progresses, we often drop some input variables, add others, change the region of exploration for some factors, or add new response variables. Consequently, we usually experiment **sequentially**, and as a general rule, no more than about 25 percent of the available resources should be invested in the first experiment. This will ensure that sufficient resources are available to perform confirmation runs and ultimately accomplish the final objective of the experiment.

Finally, it is important to recognize that **all** experiments are designed experiments. The important issue is whether they are well designed or not. Good pre-experimental planning will usually lead to a good, successful experiment. Failure to do such planning usually leads to wasted time, money, and other resources and often poor or disappointing results.

1.5 A Brief History of Statistical Design

Experimentation is an important part of the knowledge discovery process. An early record of a designed experiment in the medical field is the study of scurvy by James Lind on board the Royal Navy ship *Salisbury* in 1747. Lind conducted a study to determine the effect of diet on scurvy and discovered the importance of fruit as a preventative measure. Today we would call the type of experiment he conducted as a completely randomized single-factor design. Experiments of this type are discussed in Chapters 2 and 3. Between 1843 and 1846 several agricultural field trials were begun at the Rothamsted Agricultural Research Station outside of London. These experiments were not carried out using modern techniques but they laid the foundation for the pioneering work of Sir Ronald A. Fisher starting about 1920. This led to the first of the four eras in the modern development of experimental design, the agricultural era.

Fisher was responsible for statistics and data analysis at Rothamsted. Fisher recognized that flaws in the way the experiment that generated the data had been performed often hampered the analysis of data from systems (in this case, agricultural systems). By interacting with scientists and researchers in many fields, he developed the insights that led to the three basic principles of experimental design that we discussed in Section 1.3: randomization, replication, and blocking. Fisher systematically introduced statistical thinking and principles into designing experimental investigations, including the factorial design concept and the analysis of variance. His two books [the most recent editions are Fisher (1958, 1966)] had profound influence on the use of statistics, particularly in agricultural and related life sciences. For an excellent biography of Fisher, see Box (1978).

Although applications of statistical design in industrial settings certainly began in the 1930s, the second, or industrial, era was catalyzed by the development of response surface methodology (RSM) by Box and Wilson (1951). They recognized and exploited the fact that many industrial experiments are fundamentally different from their agricultural counterparts in two ways: (1) the response variable can usually be observed (nearly) immediately, and (2) the experimenter can quickly learn crucial information from a small group of runs that can be used to plan the next experiment. Box (1999) calls these two features of industrial experiments **immediacy** and **sequentiality**. Over the next 30 years, RSM and other design techniques spread throughout the chemical and the process industries, mostly in research and development work. George Box was the intellectual leader of this movement. However, the application of statistical design at the plant or manufacturing process level was still not extremely widespread. Some of the reasons for this include an inadequate training in basic statistical concepts and methods for engineers and other process specialists and the lack of computing resources and user-friendly statistical software to support the application of statistically designed experiments.

It was during this second or industrial era that work on **optimal** design of experiments began. Kiefer (1959, 1961) and Kiefer and Wolfowitz (1959) proposed a formal approach to selecting a design based on specific objective optimality criteria. Their initial approach was to select a design that would result in the model parameters being estimated with the best possible precision. This approach did not find much application because of the lack of computer tools for its implementation. However, there have been great advances in both algorithms for generating optimal designs and computing capability over the last 25 years. Optimal designs have great application and are discussed at several places in the book.

The increasing interest of Western industry in quality improvement that began in the late 1970s ushered in the third era of statistical design. The work of Genichi Taguchi [Taguchi and Wu (1980), Kackar (1985), and Taguchi

(1987, 1991)] had a significant impact on expanding the interest in and use of designed experiments. Taguchi advocated using designed experiments for what he termed **robust parameter design**, or

1. Making processes insensitive to environmental factors or other factors that are difficult to control
2. Making products insensitive to variation transmitted from components
3. Finding levels of the process variables that force the mean to a desired value while simultaneously reducing variability around this value.

Taguchi suggested highly fractionated factorial designs and other orthogonal arrays along with some novel statistical methods to solve these problems. The resulting methodology generated much discussion and controversy. Part of the controversy arose because Taguchi's methodology was advocated in the West initially (and primarily) by entrepreneurs, and the underlying statistical science had not been adequately peer reviewed. By the late 1980s, the results of peer review indicated that although Taguchi's engineering concepts and objectives were well founded, there were substantial problems with his experimental strategy and methods of data analysis. For specific details of these issues, see Box (1988), Box, Bisgaard, and Fung (1988), Hunter (1985, 1989), Myers, Montgomery, and Anderson-Cook (2016), and Pignatiello and Ramberg (1992). Many of these concerns were also summarized in the extensive panel discussion in the May 1992 issue of *Technometrics* [see Nair et al. (1992)].

There were several positive outcomes of the Taguchi controversy. First, designed experiments became more widely used in the discrete parts industries, including automotive and aerospace manufacturing, electronics and semiconductors, and many other industries that had previously made little use of the technique. Second, the fourth era of statistical design began. This era has included a renewed general interest in statistical design by both researchers and practitioners and the development of many new and useful approaches to experimental problems in the industrial world, including alternatives to Taguchi's technical methods that allow his engineering concepts to be carried into practice efficiently and effectively. Some of these alternatives will be discussed and illustrated in subsequent chapters, particularly in Chapter 12. Third, computer software for construction and evaluation of designs has improved greatly with many new features and capability. Forth, formal education in statistical experimental design is becoming part of many engineering programs in universities, at both undergraduate and graduate levels. The successful integration of good experimental design practice into engineering and science is a key factor in future industrial competitiveness.

Applications of designed experiments have grown far beyond the agricultural origins. There is not a single area of science and engineering that has not successfully employed statistically designed experiments. In recent years, there has been a considerable utilization of designed experiments in many other areas, including the service sector of business, financial services, government operations, and many nonprofit business sectors. An article appeared in *Forbes* magazine on March 11, 1996, entitled "The New Mantra: MVT," where MVT stands for "multivariable testing," a term some authors use to describe factorial designs. The article notes the many successes that a diverse group of companies have had through their use of statistically designed experiments. Today e-commerce companies routinely conduct on-line experiments when users access their websites and email marketing services conduct on-line experiments for their clients.

1.6 Summary: Using Statistical Techniques in Experimentation

Much of the research in engineering, science, and industry is empirical and makes extensive use of experimentation. Statistical methods can greatly increase the efficiency of these experiments and often strengthen the conclusions so obtained. The proper use of statistical techniques in experimentation requires that the experimenter keep the following points in mind:

1. ***Use your nonstatistical knowledge of the problem.*** Experimenters are usually highly knowledgeable in their fields. For example, a civil engineer working on a problem in hydrology typically has considerable practical experience and formal academic training in this area. In some fields, there is a large body of physical theory on which to draw in explaining relationships between factors and responses. This type of nonstatistical

knowledge is invaluable in choosing factors, determining factor levels, deciding how many replicates to run, interpreting the results of the analysis, and so forth. Using a designed experiment is no substitute for thinking about the problem.

2. ***Keep the design and analysis as simple as possible.*** Don't be overzealous in the use of complex, sophisticated statistical techniques. Relatively simple design and analysis methods are almost always best. This is a good place to reemphasize steps 1–3 of the procedure recommended in Section 1.4. If you do the pre-experimental planning carefully and select a reasonable design, the analysis will almost always be relatively straightforward. In fact, a well-designed experiment will sometimes almost analyze itself! However, if you botch the pre-experimental planning and execute the experimental design badly, it is unlikely that even the most complex and elegant statistics can save the situation.
3. ***Recognize the difference between practical and statistical significance.*** Just because two experimental conditions produce mean responses that are statistically different, there is no assurance that this difference is large enough to have any practical value. For example, an engineer may determine that a modification to an automobile fuel injection system may produce a true mean improvement in gasoline mileage of 0.1 mi/gal and be able to determine that this is a statistically significant result. However, if the cost of the modification is $1000, the 0.1 mi/gal difference is probably too small to be of any practical value.
4. ***Experiments are usually iterative.*** Remember that in most situations it is unwise to design too comprehensive an experiment at the start of a study. Successful design requires the knowledge of important factors, the ranges over which these factors are varied, the appropriate number of levels for each factor, and the proper methods and units of measurement for each factor and response. Generally, we are not well equipped to answer these questions at the beginning of the experiment, but we learn the answers as we go along. This argues in favor of the **iterative**, or **sequential**, approach discussed previously. Of course, there are situations where comprehensive experiments are entirely appropriate, but as a general rule most experiments should be iterative. Consequently, we usually should not invest more than about 25 percent of the resources of experimentation (runs, budget, time, and so forth) in the initial experiment. Often these first efforts are just learning experiences, and some resources must be available to accomplish the final objectives of the experiment.

1.7 Problems

1.1 Suppose that you want to design an experiment to study the proportion of unpopped kernels of popcorn. Complete steps 1–3 of the guidelines for designing experiments in Section 1.4. Are there any major sources of variation that would be difficult to control?

1.2 Suppose that you want to investigate the factors that potentially affect cooking rice.

(a) What would you use as a response variable in this experiment? How would you measure the response?

(b) List all of the potential sources of variability that could impact the response.

(c) Complete the first three steps of the guidelines for designing experiments in Section 1.4.

1.3 Suppose that you want to compare the growth of garden flowers with different conditions of sunlight, water, fertilizer, and soil conditions. Complete steps 1–3 of the guidelines for designing experiments in Section 1.4.

1.4 Select an experiment of interest to you. Complete steps 1–3 of the guidelines for designing experiments in Section 1.4.

1.5 Search the World Wide Web for information about Sir Ronald A. Fisher and his work on experimental design in agricultural science at the Rothamsted Experimental Station.

1.6 Find a website for a business that you are interested in. Develop a list of factors that you would use in an experiment to improve the effectiveness of this website.

1.7 Almost everyone is concerned about the price of gasoline. Construct a cause-and-effect diagram identifying the factors that potentially influence the gasoline mileage that you get in your car. How would you go about conducting an experiment to determine any of these factors actually affect your gasoline mileage?

1.8 What is replication? Why do we need replication in an experiment? Present an example that illustrates the difference between replication and repeated measurements.

1.9 Why is randomization important in an experiment?

1.10 What are the potential risks of a single, large, comprehensive experiment in contrast to a sequential approach?

1.11 Have you received an offer to obtain a credit card in the mail? What "factors" were associated with the offer, such as an introductory interest rate? Do you think the credit card company is conducting experiments to investigate which factors produce the highest positive response rate to their offer? What potential factors in this experiment can you identify?

1.12 What factors do you think an e-commerce company could use in an experiment involving their web page to encourage more people to "click-through" into their site?

CHAPTER 2

Simple Comparative Experiments

CHAPTER OUTLINE

SUPPLEMENTAL MATERIAL FOR CHAPTER 2

The supplemental material is on the textbook website www.wiley.com/college/montgomery.

CHAPTER LEARNING OBJECTIVES

1. Know the importance of obtaining a random sample.
2. Be familiar with the standard sampling distributions: normal, *t*, chi-square, and *F*.
3. Know how to interpret the *P*-value for a statistical test.
4. Know how to use the *Z* test and *t*-test to compare means.
5. Know how to construct and interpret confidence intervals involving means.
6. Know how the paired *t*-test incorporates the blocking principle.

In this chapter, we consider experiments to compare two **conditions** (sometimes called **treatments**). These are often called **simple comparative experiments**. We begin with an example of an experiment performed to determine whether two different formulations of a product give equivalent results. The discussion leads to a review of several basic statistical concepts, such as random variables, probability distributions, random samples, sampling distributions, and tests of hypotheses.

2.1 Introduction

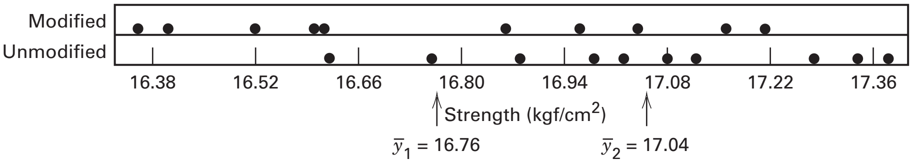
An engineer is studying the formulation of a Portland cement mortar. He has added a polymer latex emulsion during mixing to determine if this impacts the curing time and tension bond strength of the mortar. The experimenter prepared 10 samples of the original formulation and 10 samples of the modified formulation. We will refer to the two different formulations as two **treatments** or as two **levels** of the **factor** formulations. When the cure process was completed, the experimenter did find a very large reduction in the cure time for the modified mortar formulation. Then he began to address the tension bond strength of the mortar. If the new mortar formulation has an adverse effect on bond strength, this could impact its usefulness.

The tension bond strength data from this experiment are shown in Table 2.1 and plotted in Figure 2.1. The graph is called a **dot diagram**. Visual examination of these data gives the impression that the strength of the unmodified mortar may be greater than the strength of the modified mortar. This impression is supported by comparing the *average* tension bond strengths $\overline{y}_1 = 16.76\,\mathrm{kgf/cm^2}$ for the modified mortar and $\overline{y}_2 = 17.04\,\mathrm{kgf/cm^2}$ for the unmodified mortar. The average tension bond strengths in these two samples differ by what seems to be a modest amount. However, it is not obvious that this difference is large enough to imply that the two formulations really *are* different. Perhaps this observed difference in average strengths is the result of sampling fluctuation and the two formulations are really identical. Possibly another two samples would give opposite results, with the strength of the modified mortar exceeding that of the unmodified formulation.

A technique of statistical inference called **hypothesis testing** can be used to assist the experimenter in comparing these two formulations. Hypothesis testing allows the comparison of the two formulations to be made on **objective** terms, with knowledge of the risks associated with reaching the wrong conclusion. Before presenting procedures for hypothesis testing in simple comparative experiments, we will briefly summarize some elementary statistical concepts.

■ **TABLE 2.1**
Tension Bond Strength Data for the Portland Cement Formulation Experiment

j	**Modified Mortar** y_{1j}	**Unmodified Mortar** y_{2j}
1	16.85	16.62
2	16.40	16.75
3	17.21	17.37
4	16.35	17.12
5	16.52	16.98
6	17.04	16.87
7	16.96	17.34
8	17.15	17.02
9	16.59	17.08
10	16.57	17.27



■ **FIGURE 2.1** **Dot diagram for the tension bond strength data in Table 2.1**

2.2 Basic Statistical Concepts

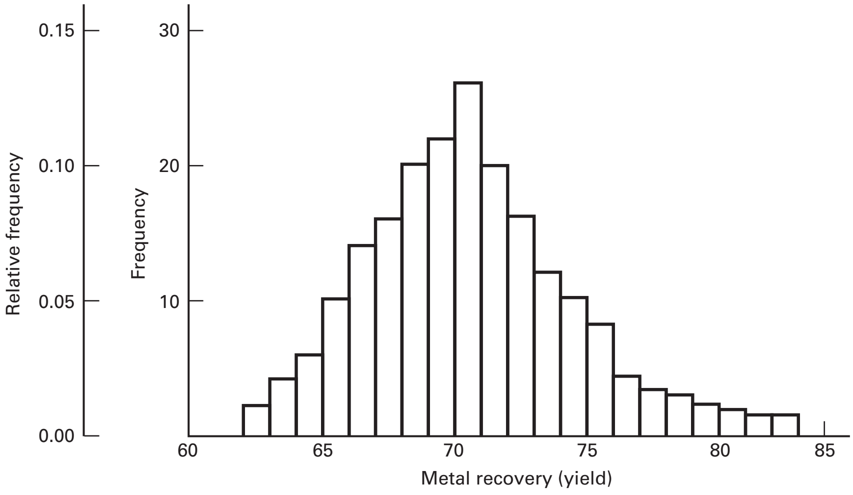
Each of the observations in the Portland cement experiment described above would be called a **run**. Notice that the individual runs differ, so there is fluctuation, or **noise**, in the observed bond strengths. This noise is usually called **experimental error** or simply **error**. It is a **statistical error**, meaning that it arises from variation that is uncontrolled and generally unavoidable. The presence of error or noise implies that the response variable, tension bond strength, is a **random variable**. A random variable may be either **discrete** or **continuous**. If the set of all possible values of the random variable is either finite or countably infinite, then the random variable is discrete, whereas if the set of all possible values of the random variable is an interval, then the random variable is continuous.

Graphical Description of Variability. We often use simple graphical methods to assist in analyzing the data from an experiment. The **dot diagram**, illustrated in Figure 2.1, is a very useful device for displaying a small body of data (say up to about 20 observations). The dot diagram enables the experimenter to see quickly the general **location** or **central tendency** of the observations and their **spread** or **variability**. For example, in the Portland cement tension bond experiment, the dot diagram reveals that the two formulations may differ in mean strength but that both formulations produce about the same variability in strength.

If the data are fairly numerous, the dots in a dot diagram become difficult to distinguish and a **histogram** may be preferable. Figure 2.2 presents a histogram for 200 observations on the metal recovery, or yield, from a smelting process. The histogram shows the central tendency, spread, and general shape of the distribution of the data. Recall that a histogram is constructed by dividing the horizontal axis into bins (usually of equal length) and drawing a rectangle over the *j*th bin with the area of the rectangle proportional to n_j, the number of observations that fall in that bin. The histogram is a large-sample tool. When the sample size is small, the shape of the histogram can be very sensitive to the number of bins, the width of the bins, and the starting value for the first bin. Histograms should not be used with fewer than 75–100 observations.

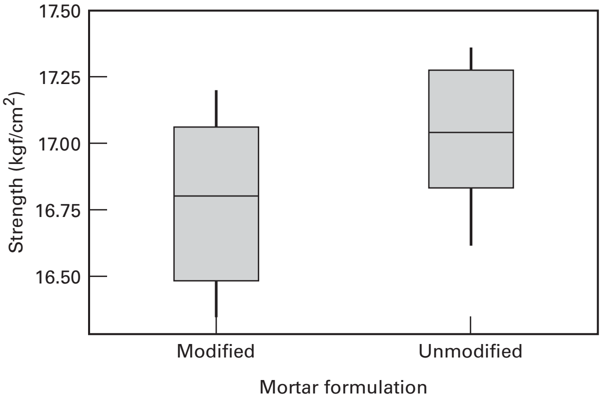
The **box plot** (or **box-and-whisker plot**) is a very useful way to display data. A box plot displays the minimum, the maximum, the lower and upper quartiles (the 25th percentile and the 75th percentile, respectively), and the median (the 50th percentile) on a rectangular box aligned either horizontally or vertically. The box extends from the lower quartile to the upper quartile, and a line is drawn through the box at the median. Lines (or whiskers) extend from the ends of the box to (typically) the minimum and maximum values. [There are several variations of box plots that have different rules for denoting the extreme sample points. See Montgomery and Runger (2011) for more details.]

Figure 2.3 presents the box plots for the two samples of tension bond strength in the Portland cement mortar experiment. This display indicates some difference in mean strength between the two formulations. It also indicates that both formulations produce reasonably symmetric distributions of strength with similar variability or spread.



■ **FIGURE 2.2 Histogram for 200 observations on metal recovery (yield) from a smelting process**

■ **FIGURE 2.3 Box plots for the Portland cement tension bond strength experiment**

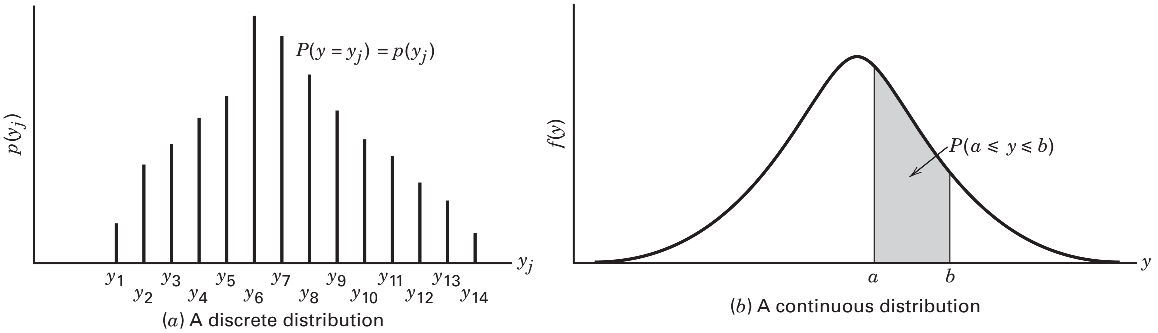


Dot diagrams, histograms, and box plots are useful for summarizing the information in a **sample** of data. To describe the observations that might occur in a sample more completely, we use the concept of the probability distribution.

Probability Distributions. The probability structure of a random variable, say y, is described by its **probability distribution**. If y is discrete, we often call the probability distribution of y, say $p(y)$, the probability mass function of y. If y is continuous, the probability distribution of y, say $f(y)$, is often called the probability density function for y.

Figure 2.4 illustrates hypothetical discrete and continuous probability distributions. Notice that in the discrete probability distribution Figure 2.4a, it is the height of the function $p(y_j)$ that represents probability, whereas in the continuous case Figure 2.4b, it is the area under the curve $f(y)$ associated with a given interval that represents probability. The properties of probability distributions may be summarized quantitatively as follows:

$$
\begin{array}{lll}
y \text{ discrete:} & 0 \le p(y_j) \le 1 & \text{all values of } y_j \\
 & P(y = y_j) = p(y_j) & \text{all values of } y_j \\
 & \displaystyle\sum_{\substack{\text{all values}\\ \text{of } y_j}} p(y_j) = 1 & \\
y \text{ continuous:} & 0 \le f(y) & \\
 & P(a \le y \le b) = \displaystyle\int_a^b f(y)\,dy & \\
 & \displaystyle\int_{-\infty}^{\infty} f(y)\,dy = 1 &
\end{array}
$$



(a) A discrete distribution

(b) A continuous distribution

■ **FIGURE 2.4 Discrete and continuous probability distributions**

Mean, Variance, and Expected Values. The **mean**, μ, of a probability distribution is a measure of its central tendency or location. Mathematically, we define the mean as

$$\mu = \begin{cases} \displaystyle\int_{-\infty}^{\infty} yf(y)\,dy & y \text{ continuous} \\ \displaystyle\sum_{\text{all } y} yp(y_j) & y \text{ discrete} \end{cases} \tag{2.1}$$

We may also express the mean in terms of the **expected value** or the long-run average value of the random variable y as

$$\mu = E(y) = \begin{cases} \displaystyle\int_{-\infty}^{\infty} yf(y)\,dy & y \text{ continuous} \\ \displaystyle\sum_{\text{all } y} yp(y_j) & y \text{ discrete} \end{cases} \tag{2.2}$$

where E denotes the **expected value operator**.

The variability or dispersion of a probability distribution can be measured by the **variance**, defined as

$$\sigma^2 = \begin{cases} \displaystyle\int_{-\infty}^{\infty} (y-\mu)^2 f(y)\,dy & y \text{ continuous} \\ \displaystyle\sum_{\text{all } y} (y-\mu)^2 p(y_j) & y \text{ discrete} \end{cases} \tag{2.3}$$

Note that the variance can be expressed entirely in terms of expectation because

$$\sigma^2 = E[(y-\mu)^2] \tag{2.4}$$

Finally, the variance is used so extensively that it is convenient to define a **variance operator** V such that

$$V(y) = E[(y-\mu)^2] = \sigma^2 \tag{2.5}$$

The concepts of expected value and variance are used extensively throughout this book, and it may be helpful to review several elementary results concerning these operators. If y is a random variable with mean μ and variance σ^2 and c is a constant, then

1. $E(c) = c$
2. $E(y) = \mu$
3. $E(cy) = cE(y) = c\mu$
4. $V(c) = 0$
5. $V(y) = \sigma^2$
6. $V(cy) = c^2V(y) = c^2\sigma^2$

If there are two random variables, say, y_1 with $E(y_1) = \mu_1$ and $V(y_1) = \sigma_1^2$ and y_2 with $E(y_2) = \mu_2$ and $V(y_2) = \sigma_2^2$, we have

7. $E(y_1 + y_2) = E(y_1) + E(y_2) = \mu_1 + \mu_2$

It is possible to show that

8. $V(y_1 + y_2) = V(y_1) + V(y_2) + 2\,\text{Cov}(y_1, y_2)$

where

$$\text{Cov}(y_1, y_2) = E[(y_1 - \mu_1)(y_2 - \mu_2)] \tag{2.6}$$

is the **covariance** of the random variables y_1 and y_2. The covariance is a measure of the linear association between y_1 and y_2. More specifically, we may show that if y_1 and y_2 are independent,[1] then $\text{Cov}(y_1, y_2) = 0$. We may also show that

9. $V(y_1 - y_2) = V(y_1) + V(y_2) - 2\,\text{Cov}(y_1, y_2)$

If y_1 and y_2 are **independent**, we have

10. $V(y_1 \pm y_2) = V(y_1) + V(y_2) = \sigma_1^2 + \sigma_2^2$

and

11. $E(y_1 \cdot y_2) = E(y_1) \cdot E(y_2) = \mu_1 \cdot \mu_2$

However, note that, in general

12. $E\left(\dfrac{y_1}{y_2}\right) \neq \dfrac{E(y_1)}{E(y_2)}$

regardless of whether or not y_1 and y_2 are independent.

2.3 Sampling and Sampling Distributions

Random Samples, Sample Mean, and Sample Variance. The objective of statistical inference is to draw conclusions about a population using a sample from that population. Most of the methods that we will study assume that **random samples** are used. A **random sample** is a sample that has been selected from the population in such a way that every possible sample has an equal probability of being selected. In practice, it is sometimes difficult to obtain random samples, and random numbers generated by a computer program may be helpful.

Statistical inference makes considerable use of quantities computed from the observations in the sample. We define a **statistic** as any function of the observations in a sample that does not contain unknown parameters. For example, suppose that $y_1, y_2, \ldots, y_n$ represents a sample. Then the **sample mean**

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{2.7}$$

and the **sample variance**

$$S^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1} \tag{2.8}$$

are both statistics. These quantities are measures of the central tendency and dispersion of the sample, respectively. Sometimes $S = \sqrt{S^2}$, called the **sample standard deviation**, is used as a measure of dispersion. Experimenters often prefer to use the standard deviation to measure dispersion because its units are the same as those for the variable of interest y.

Properties of the Sample Mean and Variance. The sample mean $\overline{y}$ is a point estimator of the population mean μ, and the sample variance S^2 is a point estimator of the population variance σ^2. In general, an **estimator** of an

[1] Note that the converse of this is not necessarily so; that is, we may have $\text{Cov}(y_1, y_2) = 0$ and yet this does not imply independence. For an example, see Hines et al. (2003).

unknown parameter is a statistic that corresponds to that parameter. Note that a point estimator is a random variable. A particular numerical value of an estimator, computed from sample data, is called an **estimate**. For example, suppose we wish to estimate the mean and variance of the suspended solid material in the water of a lake. A random sample of $n = 25$ observations is tested, and the mg/l of suspended solid material is measured and recorded for each. The sample mean and variance are computed according to Equations 2.7 and 2.8, respectively, and are $\overline{y} = 18.6$ and $S^2 = 1.20$. Therefore, the estimate of μ is $\overline{y} = 18.6$ mg/l, and the estimate of σ^2 is $S^2 = 1.20$ (mg/l)^2.

Several properties are required of good point estimators. Two of the most important are the following:

1. The point estimator should be **unbiased**. That is, the long-run average or expected value of the point estimator should be equal to the parameter that is being estimated. Although unbiasedness is desirable, this property alone does not always make an estimator a good one.
2. An unbiased estimator should have **minimum variance**. This property states that the minimum variance point estimator has a variance that is smaller than the variance of any other estimator of that parameter.

We may easily show that $\overline{y}$ and S^2 are unbiased estimators of μ and σ^2, respectively. First consider $\overline{y}$. Using the properties of expectation, we have

$$
\begin{aligned}
E(\overline{y}) &= E\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \mu
\end{aligned}
$$

because the expected value of each observation y_i is μ. Thus, $\overline{y}$ is an unbiased estimator of μ.

Now consider the sample variance S^2. We have

$$
\begin{aligned}
E(S^2) &= E\left[\frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}\right] \\
&= \frac{1}{n-1} E\left[\sum_{i=1}^{n} (y_i - \overline{y})^2\right] \\
&= \frac{1}{n-1} E(SS)
\end{aligned}
$$

where $SS = \sum_{i=1}^{n} (y_i - \overline{y})^2$ is the **corrected sum of squares** of the observations y_i. Now

$$
\begin{aligned}
E(SS) &= E\left[\sum_{i=1}^{n} (y_i - \overline{y})^2\right] && \textbf{(2.9)} \\
&= E\left[\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right] \\
&= \sum_{i=1}^{n} (\mu^2 + \sigma^2) - n(\mu^2 + \sigma^2/n) \\
&= (n-1)\sigma^2 && \textbf{(2.10)}
\end{aligned}
$$

Therefore,

$$E(S^2) = \frac{1}{n-1} E(SS) = \sigma^2$$

Therefore S^2 is an unbiased estimator of σ^2.

Degrees of Freedom. The quantity $n - 1$ in Equation 2.10 is called the **number of degrees of freedom** of the sum of squares SS. This is a very general result; that is, if y is a random variable with variance σ^2 and $SS = \sum (y_i - \overline{y})^2$ has v degrees of freedom, then

$$E\left(\frac{SS}{v}\right) = \sigma^2 \qquad \textbf{(2.11)}$$

The number of degrees of freedom of a sum of squares is equal to the number of independent elements in that sum of squares. For example, $SS = \sum_{i=1}^{n} (y_i - \overline{y})^2$ in Equation 2.9 consists of the sum of squares of the n elements $y_1 - \overline{y}, y_2 - \overline{y}, \ldots, y_n - \overline{y}$. These elements are not all independent because $\sum_{i=1}^{n}(y_i - \overline{y}) = 0$; in fact, only $n - 1$ of them are independent, implying that SS has $n - 1$ degrees of freedom.

The Normal and Other Sampling Distributions. Often we are able to determine the probability distribution of a particular statistic if we know the probability distribution of the population from which the sample was drawn. The probability distribution of a statistic is called a **sampling distribution**. We will now briefly discuss several useful sampling distributions.

One of the most important sampling distributions is the **normal distribution**. If y is a normal random variable, the probability distribution of y is

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)[(y-\mu)/\sigma]^2} \qquad -\infty < y < \infty \qquad \textbf{(2.12)}$$

where $-\infty < \mu < \infty$ is the mean of the distribution and $\sigma^2 > 0$ is the variance. The normal distribution is shown in Figure 2.5.

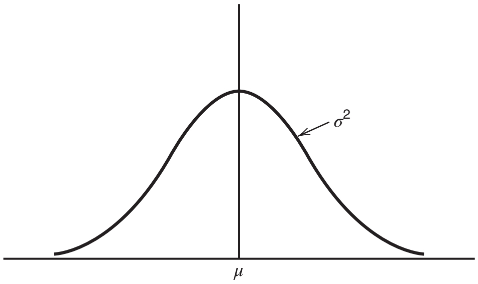
Because sample observations that differ as a result of experimental error often are well described by the normal distribution, the normal plays a central role in the analysis of data from designed experiments. Many important sampling distributions may also be defined in terms of normal random variables. We often use the notation $y \sim N(\mu, \sigma^2)$ to denote that y is distributed normally with mean μ and variance σ^2.

An important special case of the normal distribution is the **standard normal distribution**; that is, $\mu = 0$ and $\sigma^2 = 1$. We see that if $y \sim N(\mu, \sigma^2)$, the random variable

$$z = \frac{y - \mu}{\sigma} \qquad \textbf{(2.13)}$$

follows the standard normal distribution, denoted $z \sim N(0, 1)$. The operation demonstrated in Equation 2.13 is often called **standardizing** the normal random variable y. The cumulative standard normal distribution is given in Table I of the Appendix.

■ **FIGURE 2.5** **The normal distribution**



Many statistical techniques assume that the random variable is normally distributed. The central limit theorem is often a justification of approximate normality.

THEOREM 2-1
The Central Limit Theorem

If $y_1, y_2, \ldots, y_n$ is a sequence of n independent and identically distributed random variables with $E(y_i) = \mu$ and $V(y_i) = \sigma^2$ (both finite) and $x = y_1 + y_2 + \cdots + y_n$, then the limiting form of the distribution of

$$z_n = \frac{x - n\mu}{\sqrt{n\sigma^2}}$$

as $n \to \infty$, is the standard normal distribution.

This result states essentially that the sum of n independent and identically distributed random variables is approximately normally distributed. In many cases, this approximation is good for very small n, say $n < 10$, whereas in other cases large n is required, say $n > 100$. Frequently, we think of the error in an experiment as arising in an additive manner from several independent sources; consequently, the normal distribution becomes a plausible model for the combined experimental error.

An important sampling distribution that can be defined in terms of normal random variables is the **chi-square** or **χ^2 distribution**. If $z_1, z_2, \ldots, z_k$ are normally and independently distributed random variables with mean 0 and variance 1, abbreviated NID(0, 1), then the random variable

$$x = z_1^2 + z_2^2 + \cdots + z_k^2$$

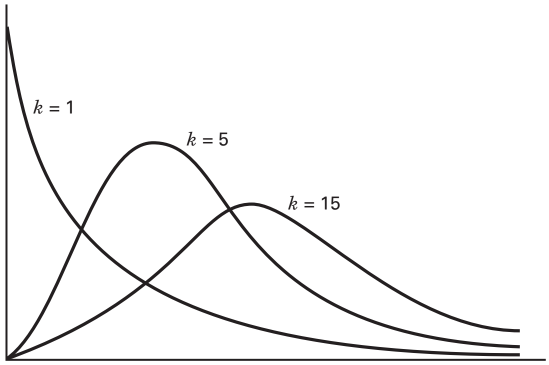
follows the **chi-square distribution with *k* degrees of freedom**. The density function of chi-square is

$$f(x) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} x^{(k/2)-1} e^{-x/2} \qquad x > 0 \tag{2.14}$$

Several chi-square distributions are shown in Figure 2.6. The distribution is asymmetric, or **skewed**, with mean and variance

$$\mu = k$$
$$\sigma^2 = 2k$$

respectively. Percentage points of the chi-square distribution are given in Table III of the Appendix.



■ **FIGURE 2.6** **Several chi-square distributions**

As an example of a random variable that follows the chi-square distribution, suppose that $y_1, y_2, \ldots, y_n$ is a random sample from an $N(\mu, \sigma^2)$ distribution. Then

$$\frac{SS}{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sigma^2} \sim \chi^2_{n-1} \tag{2.15}$$

That is, SS/σ^2 is distributed as chi-square with $n-1$ degrees of freedom.

Many of the techniques used in this book involve the computation and manipulation of sums of squares. The result given in Equation 2.15 is extremely important and occurs repeatedly; a sum of squares in normal random variables when divided by σ^2 follows the chi-square distribution.

Examining Equation 2.8, note the sample variance can be written as

$$S^2 = \frac{SS}{n-1} \tag{2.16}$$

If the observations in the sample are NID(μ, σ^2), then the distribution of S^2 is $[\sigma^2/(n-1)]\chi^2_{n-1}$. Thus, the sampling distribution of the sample variance is a constant times the chi-square distribution if the population is normally distributed.

If z and χ^2_k are independent standard normal and chi-square random variables, respectively, the random variable

$$t_k = \frac{z}{\sqrt{\chi^2_k/k}} \tag{2.17}$$

follows the ***t* distribution with *k* degrees of freedom**, denoted t_k. The density function of t is

$$f(t) = \frac{\Gamma[(k+1)/2]}{\sqrt{k\pi}\Gamma(k/2)} \frac{1}{[(t^2/k)+1]^{(k+1)/2}} \qquad -\infty < t < \infty \tag{2.18}$$

and the mean and variance of t are $\mu = 0$ and $\sigma^2 = k/(k-2)$ for $k > 2$, respectively. Several t distributions are shown in Figure 2.7. Note that if $k = \infty$, the t distribution becomes the standard normal distribution. The percentage points of the t distribution are given in Table II of the Appendix. If $y_1, y_2, \ldots, y_n$ is a random sample from the $N(\mu, \sigma^2)$ distribution, then the quantity

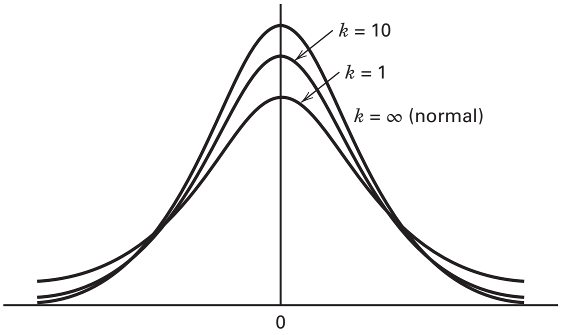
$$t = \frac{\bar{y} - \mu}{S/\sqrt{n}} \tag{2.19}$$

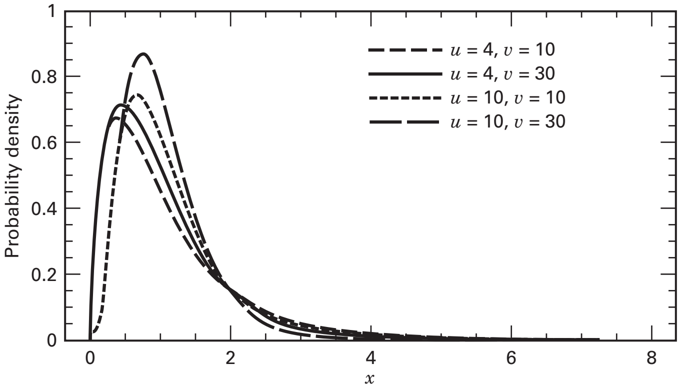
is distributed as t with $n-1$ degrees of freedom.

The final sampling distribution that we will consider is the ***F* distribution**. If χ^2_u and χ^2_v are two independent chi-square random variables with u and v degrees of freedom, respectively, then the ratio

$$F_{u,v} = \frac{\chi^2_u/u}{\chi^2_v/v} \tag{2.20}$$

■ **FIGURE 2.7 Several *t* distributions**





■ **FIGURE 2.8** **Several *F* distributions**

follows the ***F* distribution with *u numerator* degrees of freedom and *v denominator* degrees of freedom**. If x is an F random variable with u numerator and v denominator degrees of freedom, then the probability distribution of x is

$$h(x) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{u/2} x^{(u/2)-1}}{\Gamma\left(\frac{u}{x}\right)\Gamma\left(\frac{v}{2}\right)\left[\left(\frac{u}{v}\right)x+1\right]^{(u+v)/2}} \qquad 0 < x < \infty \tag{2.21}$$

Several F distributions are shown in Figure 2.8. This distribution is very important in the statistical analysis of designed experiments. Percentage points of the F distribution are given in Table IV of the Appendix.

As an example of a statistic that is distributed as F, suppose we have two independent normal populations with common variance σ^2. If $y_{11}, y_{12}, \ldots, y_{1n_1}$ is a random sample of n_1 observations from the first population, and if $y_{21}, y_{22}, \ldots, y_{2n_2}$ is a random sample of n_2 observations from the second, then

$$\frac{S_1^2}{S_2^2} \sim F_{n_1-1,\, n_2-1} \tag{2.22}$$

where S_1^2 and S_2^2 are the two sample variances. This result follows directly from Equations 2.15 and 2.20.

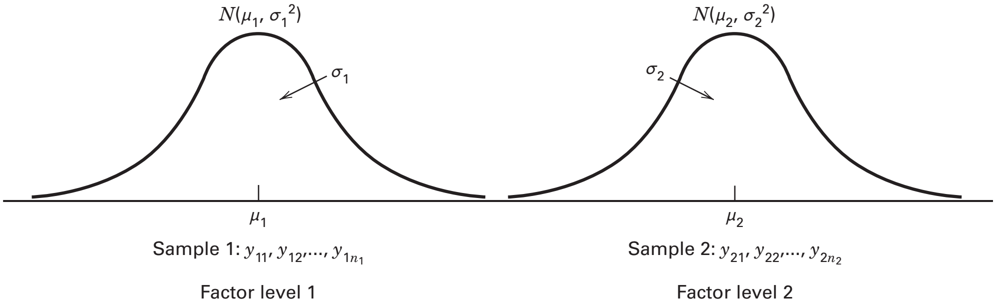
2.4 Inferences About the Differences in Means, Randomized Designs

We are now ready to return to the Portland cement mortar problem posed in Section 2.1. Recall that two different formulations of mortar were being investigated to determine if they differ in tension bond strength. In this section, we discuss how the data from this simple comparative experiment can be analyzed using **hypothesis testing** and **confidence interval** procedures for comparing two treatment means.

Throughout this section, we assume that a **completely randomized experimental design** is used. In such a design, the data are viewed as a random sample from a normal distribution. The random sample assumption is very important.

2.4.1 Hypothesis Testing

We now reconsider the Portland cement experiment introduced in Section 2.1. Recall that we are interested in comparing the strength of two different formulations: an unmodified mortar and a modified mortar. In general, we can think of these two formulations as two **levels of the factor** "formulations." Let $y_{11}, y_{12}, \ldots, y_{1n_1}$ represent the n_1 observations from the first factor level and $y_{21}, y_{22}, \ldots, y_{2n_2}$ represent the n_2 observations from the second factor level. We assume that the samples are drawn at random from two independent normal populations. Figure 2.9 illustrates the situation.



■ **FIGURE 2.9** **The sampling situation for the two-sample *t*-test**

A Model for the Data. We often describe the results of an experiment with a **model**. A simple statistical model that describes the data from an experiment such as we have just described is

$$y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, 2 \\ j = 1, 2, \ldots, n_i \end{cases} \tag{2.23}$$

where y_{ij} is the jth observation from factor level i, μ_i is the mean of the response at the ith factor level, and ϵ_{ij} is a normal random variable associated with the ijth observation. We assume that ϵ_{ij} are NID$(0, \sigma_i^2)$, $i = 1, 2$. It is customary to refer to ϵ_{ij} as the **random error** component of the model. Because the means μ_1 and μ_2 are constants, we see directly from the model that y_{ij} are NID(μ_i, σ_i^2), $i = 1, 2$, just as we previously assumed. For more information about models for the data, refer to the supplemental text material.

Statistical Hypotheses. A **statistical hypothesis** is a statement either about the parameters of a probability distribution or the parameters of a model. The hypothesis reflects some **conjecture** about the problem situation. For example, in the Portland cement experiment, we may think that the mean tension bond strengths of the two mortar formulations are equal. This may be stated formally as

$$H_0{:}\mu_1 = \mu_2$$
$$H_1{:}\mu_1 \neq \mu_2$$

where μ_1 is the mean tension bond strength of the modified mortar and μ_2 is the mean tension bond strength of the unmodified mortar. The statement $H_0{:}\mu_1 = \mu_2$ is called the **null hypothesis** and $H_1{:}\mu_1 \neq \mu_2$ is called the **alternative hypothesis**. The alternative hypothesis specified here is called a **two-sided alternative hypothesis** because it would be true if $\mu_1 < \mu_2$ or if $\mu_1 > \mu_2$.

To test a hypothesis, we devise a procedure for taking a random sample, computing an appropriate **test statistic**, and then rejecting or failing to reject the null hypothesis H_0 based on the computed value of the test statistic. Part of this procedure is specifying the set of values for the test statistic that leads to rejection of H_0. This set of values is called the **critical region** or **rejection region** for the test.

Two kinds of errors may be committed when testing hypotheses. If the null hypothesis is rejected when it is true, a type I error has occurred. If the null hypothesis is *not* rejected when it is false, a type II error has been made. The probabilities of these two errors are given special symbols

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 | H_0 \text{ is true})$$
$$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 | H_0 \text{ is false})$$

Sometimes it is more convenient to work with the **power** of the test, where

$$\text{Power} = 1 - \beta = P(\text{reject } H_0 | H_0 \text{ is false})$$

The general procedure in hypothesis testing is to specify a value of the probability of type I error α, often called the **significance level** of the test, and then design the test procedure so that the probability of type II error β has a suitably small value.

The Two-Sample t-Test. Suppose that we could assume that the variances of tension bond strengths were identical for both mortar formulations. Then the appropriate test statistic to use for comparing two treatment means in the completely randomized design is

$$t_0 = \frac{\overline{y}_1 - \overline{y}_2}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \tag{2.24}$$

where $\overline{y}_1$ and $\overline{y}_2$ are the sample means, n_1 and n_2 are the sample sizes, S_p^2 is an estimate of the common variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$ computed from

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{2.25}$$

and S_1^2 and S_2^2 are the two individual sample variances. The quantity $S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$ in the denominator of Equation 2.24 is often called the **standard error** of the difference in means in the numerator, abbreviated $se(\overline{y}_1 - \overline{y}_2)$. To determine whether to reject $H_0 : \mu_1 = \mu_2$, we would compare t_0 to the t distribution with $n_1 + n_2 - 2$ degrees of freedom. If $|t_0| > t_{\alpha/2,n_1+n_2-2}$, where $t_{\alpha/2,n_1+n_2-2}$ is the upper $\alpha/2$ percentage point of the t distribution with $n_1 + n_2 - 2$ degrees of freedom, we would *reject* H_0 and conclude that the mean strengths of the two formulations of Portland cement mortar differ. This test procedure is usually called the **two-sample *t*-test**.

This procedure may be justified as follows. If we are sampling from independent normal distributions, then the distribution of $\overline{y}_1 - \overline{y}_2$ is $N[\mu_1 - \mu_2, \sigma^2(1/n_1 + 1/n_2)]$. Thus, if σ^2 were known, and if $H_0 : \mu_1 = \mu_2$ were true, the distribution of

$$Z_0 = \frac{\overline{y}_1 - \overline{y}_2}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \tag{2.26}$$

would be $N(0, 1)$. However, in replacing σ in Equation 2.26 by S_p, the distribution of Z_0 changes from standard normal to t with $n_1 + n_2 - 2$ degrees of freedom. Now if H_0 is true, t_0 in Equation 2.24 is distributed as $t_{n_1+n_2-2}$ and, consequently, we would expect $100(1 - \alpha)$ percent of the values of t_0 to fall between $-t_{\alpha/2,n_1+n_2-2}$ and $t_{\alpha/2,n_1+n_2-2}$. A sample producing a value of t_0 outside these limits would be unusual if the null hypothesis were true and is evidence that H_0 should be rejected. Thus the t distribution with $n_1 + n_2 - 2$ degrees of freedom is the appropriate **reference distribution** for the test statistic t_0. That is, it describes the behavior of t_0 when the null hypothesis is true. Note that α is the probability of type I error for the test. Sometimes α is called the **significance level** of the test.

In some problems, one may wish to reject H_0 only if one mean is larger than the other. Thus, one would specify a **one-sided alternative hypothesis** $H_1 : \mu_1 > \mu_2$ and would reject H_0 only if $t_0 > t_{\alpha,n_1+n_2-2}$. If one wants to reject H_0 only if μ_1 is less than μ_2, then the alternative hypothesis is $H_1 : \mu_1 < \mu_2$, and one would reject H_0 if $t_0 < -t_{\alpha,n_1+n_2-2}$.

To illustrate the procedure, consider the Portland cement data in Table 2.1. For these data, we find that

Modified Mortar	**Unmodified Mortar**
$\overline{y}_1 = 16.76\ \text{kgf/cm}^2$	$\overline{y}_2 = 17.04\ \text{kgf/cm}^2$
$S_1^2 = 0.100$	$S_2^2 = 0.061$
$S_1 = 0.316$	$S_2 = 0.248$
$n_1 = 10$	$n_2 = 10$

Because the sample standard deviations are reasonably similar, it is not unreasonable to conclude that the population standard deviations (or variances) are equal. Therefore, we can use Equation 2.24 to test the hypotheses

$$H_0\colon \mu_1 = \mu_2$$
$$H_1\colon \mu_1 \neq \mu_2$$

Furthermore, $n_1 + n_2 - 2 = 10 + 10 - 2 = 18$, and if we choose $\alpha = 0.05$, then we would reject $H_0\colon \mu_1 = \mu_2$ if the numerical value of the test statistic $t_0 > t_{0.025,18} = 2.101$, or if $t_0 < -t_{0.025,18} = -2.101$. These boundaries of the critical region are shown on the reference distribution (t with 18 degrees of freedom) in Figure 2.10.

Using Equation 2.25 we find that

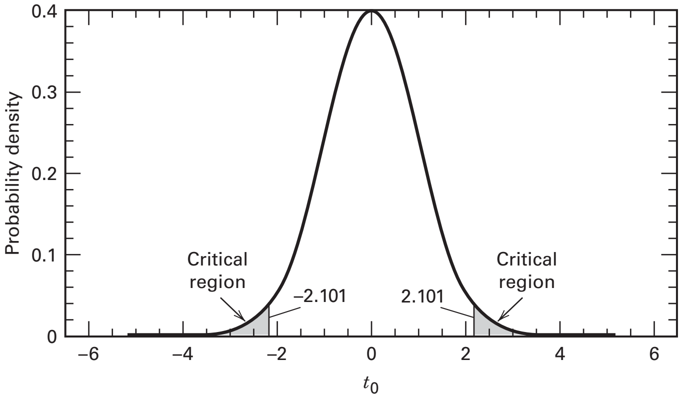
$$\begin{aligned} S_p^2 &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \\ &= \frac{9(0.100) + 9(0.061)}{10 + 10 - 2} = 0.081 \\ S_p &= 0.284 \end{aligned}$$

and the test statistic is

$$\begin{aligned} t_0 &= \frac{\bar{y}_1 - \bar{y}_2}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = \frac{16.76 - 17.04}{0.284\sqrt{\dfrac{1}{10} + \dfrac{1}{10}}} \\ &= \frac{-0.28}{0.127} = -2.20 \end{aligned}$$

Because $t_0 = -2.20 < -t_{0.025,18} = -2.101$, we would reject H_0 and conclude that the mean tension bond strengths of the two formulations of Portland cement mortar are different. This is a potentially important engineering finding. The change in mortar formulation had the desired effect of reducing the cure time, but there is evidence that the change also affected the tension bond strength. One can conclude that the modified formulation reduces the bond strength (just because we conducted a two-sided test, this does not preclude drawing a one-sided conclusion when the null hypothesis is rejected). If the reduction in mean bond strength is of practical importance (or has engineering significance in addition to statistical significance), then more development work and further experimentation will likely be required.

The Use of P-Values in Hypothesis Testing. One way to report the results of a hypothesis test is to state that the null hypothesis was or was not rejected at a specified α-value or **level of significance**. This is often called **fixed significance level testing**. For example, in the Portland cement mortar formulation above, we can say that $H_0\colon \mu_1 = \mu_2$



■ **FIGURE 2.10 The *t* distribution with 18 degrees of freedom with the critical region $\pm t_{0.025,18} = \pm 2.101$**

was rejected at the 0.05 level of significance. This statement of conclusions is often inadequate because it gives the decision maker no idea about whether the computed value of the test statistic was just barely in the rejection region or whether it was very far into this region. Furthermore, stating the results this way imposes the predefined level of significance on other users of the information. This approach may be unsatisfactory because some decision makers might be uncomfortable with the risks implied by $\alpha = 0.05$.

To avoid these difficulties, the ***P*-value approach** has been adopted widely in practice. The P-value is the probability that the test statistic will take on a value that is at least as extreme as the observed value of the statistic when the null hypothesis H_0 is true. Thus, a P-value conveys much information about the weight of evidence against H_0, and so a decision maker can draw a conclusion at *any* specified level of significance. More formally, we define the ***P*-value** as the smallest level of significance that would lead to rejection of the null hypothesis H_0.

It is customary to call the test statistic (and the data) significant when the null hypothesis H_0 is rejected; therefore, we may think of the P-value as the smallest level α at which the data are significant. Once the P-value is known, the decision maker can determine how significant the data are without the data analyst formally imposing a preselected level of significance.

It is not always easy to compute the exact P-value for a test. However, most modern computer programs for statistical analysis report P-values, and they can be obtained on some handheld calculators. We will show how to approximate the P-value for the Portland cement mortar experiment. Because $|t_0| = 2.20 > t_{0.025,18} = 2.101$, we know that the P-value is less than 0.05. From Appendix Table II, for a t distribution with 18 degrees of freedom, and tail area probability 0.01 we find $t_{0.01,18} = 2.552$. Now $|t_0| = 2.20 < 2.552$, so because the alternative hypothesis is two sided, we know that the P-value must be between 0.05 and $2(0.01) = 0.02$. Some handheld calculators have the capability to calculate P-values. One such calculator is the HP-48. From this calculator, we obtain the P-value for the value $t_0 = -2.20$ in the Portland cement mortar formulation experiment as $P = 0.0411$. Thus, the null hypothesis $H_0{:}\,\mu_1 = \mu_2$ would be rejected at any level of significance $\alpha > 0.0411$.

Computer Solution. Many statistical software packages have capability for statistical hypothesis testing. The output from both the Minitab and the JMP two-sample t-test procedure applied to the Portland cement mortar formulation experiment is shown in Table 2.2. Notice that the output includes some summary statistics about the two samples (the abbreviation "SE mean" in the Minitab section of the table refers to the standard error of the mean, $s/\sqrt{n}$) as well as some information about confidence intervals on the difference in the two means (which we will discuss in the next section). The programs also test the hypothesis of interest, allowing the analyst to specify the nature of the alternative hypothesis ("not =" in the Minitab output implies $H_1{:}\,\mu_1 \neq \mu_2$).

The output includes the computed value of t_0, the value of the test statistic t_0 (JMP reports a positive value of t_0 because of how the sample means are subtracted in the numerator of the test statistic), and the P-value. Notice that the computed value of the t statistic differs slightly from our manually calculated value and that the P-value is reported to be $P = 0.042$. JMP also reports the P-values for the one-sided alternative hypothesis. Many software packages will not report an actual P-value less than some predetermined value such as 0.0001 and instead will return a "default" value such as "< 0.001" or, in some cases, zero.

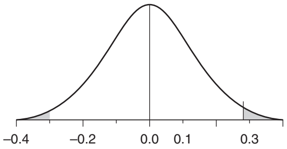
Checking Assumptions in the t-Test. In using the t-test procedure we make the assumptions that both samples are random samples that are drawn from independent populations that can be described by a normal distribution and that the standard deviation or variances of both populations are equal. The assumption of independence is critical, and if the run order is randomized (and, if appropriate, other experimental units and materials are selected at random), this assumption will usually be satisfied. The equal variance and normality assumptions are easy to check using a **normal probability plot**.

Generally, probability plotting is a graphical technique for determining whether sample data conform to a hypothesized distribution based on a subjective visual examination of the data. The general procedure is very simple and can be performed quickly with most statistics software packages. The **supplemental text material** discusses manual construction of normal probability plots.

■ **TABLE 2.2**
Computer Output for the Two-Sample *t*-Test

```
Minitab
Two-sample T for Modified vs Unmodified
                   N         Mean        Std. Dev.       SE Mean
Modified          10       16.764            0.316          0.10
Unmodified        10       17.042            0.248         0.078

Difference = mu (Modified) - mu (Unmodified)
Estimate for difference: -0.278000
95% CI for difference: (-0.545073, -0.010927)
T-Test of difference = 0 (vs not = ): T-Value = -2.19
P-Value = 0.042 DF = 18
Both use Pooled Std. Dev. = 0.2843
```

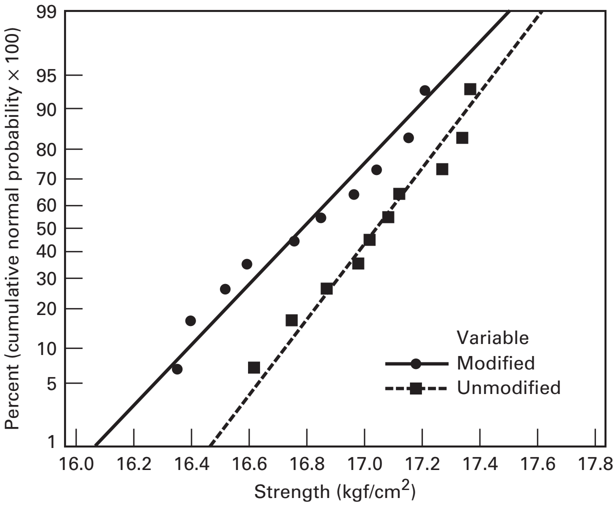
```
JMP t-test

Unmodified-Modified

Assuming equal variances

Difference     0.278000 t  Ratio   2.186876
Std Err Dif    0.127122 DF               18
Upper CL Dif   0.545073 Prob>|t|     0.0422
Lower CL Dif   0.010927 Prob>t       0.0211
Confidence         0.95 Prob<t       0.9789
```



To construct a probability plot, the observations in the sample are first ranked from smallest to largest. That is, the sample $y_1, y_2, \ldots, y_n$ is arranged as $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$, where $y_{(1)}$ is the smallest observation, $y_{(2)}$ is the second smallest observation, and so forth, with $y_{(n)}$ being the largest. The ordered observations $y_{(j)}$ are then plotted against their observed cumulative frequency $(j - 0.5)/n$. The cumulative frequency scale has been arranged so that if the hypothesized distribution adequately describes the data, the plotted points will fall approximately along a straight line; if the plotted points deviate significantly from a straight line, the hypothesized model is not appropriate. Usually, the determination of whether or not the data plot as a straight line is subjective.

To illustrate the procedure, suppose that we wish to check the assumption that tension bond strength in the Portland cement mortar formulation experiment is normally distributed. We initially consider only the observations from the unmodified mortar formulation. A computer-generated normal probability plot is shown in Figure 2.11. Most normal probability plots present $100(j - 0.5)/n$ on the left vertical scale (and occasionally $100[1 - (j - 0.5)/n]$ is plotted on the right vertical scale), with the variable value plotted on the horizontal scale. Some computer-generated normal probability plots convert the cumulative frequency to a standard normal z score. A straight line, chosen subjectively, has been drawn through the plotted points. In drawing the straight line, you should be influenced more by the points near the middle of the plot than by the extreme points. A good rule of thumb is to draw the line approximately between the 25th and 75th percentile points. This is how the lines in Figure 2.11 for each sample were determined. In assessing the "closeness" of the points to the straight line, imagine a fat pencil lying along the line. If all the points are covered by this imaginary pencil, a normal distribution adequately describes the data. Because the points for each sample in Figure 2.11 would pass the fat pencil test, we conclude that the normal distribution is an appropriate model for tension bond strength for both the modified and the unmodified mortar.

We can obtain an estimate of the mean and standard deviation directly from the normal probability plot. The mean is estimated as the 50th percentile on the probability plot, and the standard deviation is estimated as the difference



■ **FIGURE 2.11 Normal probability plots of tension bond strength in the Portland cement experiment**

between the 84th and 50th percentiles. This means that we can verify the assumption of equal population variances in the Portland cement experiment by simply comparing the slopes of the two straight lines in Figure 2.11. Both lines have very similar slopes, and so the assumption of equal variances is a reasonable one. If this assumption is violated, you should use the version of the t-test described in Section 2.4.4. The supplemental text material has more information about checking assumptions on the t-test.

When assumptions are badly violated, the performance of the t-test will be affected. Generally, small to moderate violations of assumptions are not a major concern, but *any* failure of the independence assumption and strong indications of nonnormality should not be ignored. Both the significance level of the test and the ability to detect differences between the means will be adversely affected by departures from assumptions. **Transformations** are one approach to dealing with this problem. We will discuss this in more detail in Chapter 3. Nonparametric hypothesis testing procedures can also be used if the observations come from nonnormal populations. Refer to Montgomery and Runger (2011) for more details.

An Alternate Justification to the t-Test. The two-sample t-test we have just presented depends in theory on the underlying assumption that the two populations from which the samples were randomly selected are normal. Although the normality assumption is required to develop the test procedure formally, as we discussed above, moderate departures from normality will not seriously affect the results. It can be argued that the use of a randomized design enables one to test hypotheses without *any* assumptions regarding the form of the distribution. Briefly, the reasoning is as follows. If the treatments have no effect, all $[20!/(10!10!)] = 184{,}756$ possible ways that the 20 observations could occur are equally likely. Corresponding to each of these 184,756 possible arrangements is a value of t_0. If the value of t_0 actually obtained from the data is unusually large or unusually small with reference to the set of 184,756 possible values, it is an indication that $\mu_1 \neq \mu_2$.

This type of procedure is called a **randomization test**. It can be shown that the t-test is a good approximation of the randomization test. Thus, we will use t-tests (and other procedures that can be regarded as approximations of randomization tests) without extensive concern about the assumption of normality. This is one reason a simple procedure such as normal probability plotting is adequate to check the assumption of normality.

2.4.2 Confidence Intervals

Although hypothesis testing is a useful procedure, it sometimes does not tell the entire story. It is often preferable to provide an interval within which the value of the parameter or parameters in question would be expected to lie. These interval statements are called **confidence intervals**. In many engineering and industrial experiments, the experimenter already knows that the means μ_1 and μ_2 differ; consequently, hypothesis testing on $\mu_1 = \mu_2$ is of little interest. The experimenter would usually be more interested in knowing how much the means differ. A confidence interval on the difference in means $\mu_1 - \mu_2$ is used in answering this question. It is good practice to accompany every test of a hypothesis with a confidence interval whenever possible.

To define a confidence interval, suppose that θ is an unknown parameter. To obtain an interval estimate of θ, we need to find two statistics L and U such that the probability statement

$$P(L \leqslant \theta \leqslant U) = 1 - \alpha \tag{2.27}$$

is true. The interval

$$L \leqslant \theta \leqslant U \tag{2.28}$$

is called a **100(1 − α) percent confidence interval** for the parameter θ. The interpretation of this interval is that if, in repeated random samplings, a large number of such intervals are constructed, 100(1 − α) percent of them will contain the true value of θ. The statistics L and U are called the **lower** and **upper confidence limits**, respectively, and $1 - \alpha$ is called the **confidence coefficient**. If $\alpha = 0.05$, Equation 2.28 is called a 95 percent confidence interval for θ. Note that confidence intervals have a frequency interpretation; that is, we do not know if the statement is true for this specific sample, but we do know that the *method* used to produce the confidence interval yields correct statements 100(1 − α) percent of the time.

Suppose that we wish to find a 100(1 − α) percent confidence interval on the true difference in means $\mu_1 - \mu_2$ for the Portland cement problem. The interval can be derived in the following way. The statistic

$$\frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

is distributed as $t_{n_1+n_2-2}$. Thus,

$$P\left(-t_{\alpha/2,n_1+n_2-2} \leq \frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \leq t_{\alpha/2,n_1+n_2-2}\right) = 1 - \alpha$$

or

$$\begin{aligned} P\Bigg(\bar{y}_1 - \bar{y}_2 - t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} &\leq \mu_1 - \mu_2 \\ &\leq \bar{y}_1 - \bar{y}_2 + t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\Bigg) = 1 - \alpha \end{aligned} \tag{2.29}$$

Comparing Equations 2.29 and 2.27, we see that

$$\begin{aligned} \bar{y}_1 - \bar{y}_2 - t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} &\leq \mu_1 - \mu_2 \\ &\leq \bar{y}_1 - \bar{y}_2 + t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \end{aligned} \tag{2.30}$$

is a 100(1 − α) percent confidence interval for $\mu_1 - \mu_2$.

The actual 95 percent confidence interval estimate for the difference in mean tension bond strength for the formulations of Portland cement mortar is found by substituting in Equation 2.30 as follows:

$$
\begin{aligned}
16.76 - 17.04 - (2.101)0.284\sqrt{\tfrac{1}{10} + \tfrac{1}{10}} \leq \mu_1 - \mu_2 & \\
& \leq 16.76 - 17.04 + (2.101)0.284\sqrt{\tfrac{1}{10} + \tfrac{1}{10}} \\
-0.28 - 0.27 \leq \mu_1 - \mu_2 \leq -0.28 + 0.27 & \\
-0.55 \leq \mu_1 - \mu_2 \leq -0.01 &
\end{aligned}
$$

Thus, the 95 percent confidence interval estimate on the difference in means extends from −0.55 to −0.01 kgf/cm^2. Put another way, the confidence interval is $\mu_1 - \mu_2 = -0.28 \pm 0.27$ kgf/cm^2, or the difference in mean strengths is −0.28 kgf/cm^2, and the accuracy of this estimate is ±0.27 kgf/cm^2. Note that because $\mu_1 - \mu_2 = 0$ is *not* included in this interval, the data do not support the hypothesis that $\mu_1 = \mu_2$ at the 5 percent level of significance (recall that the *P*-value for the two-sample *t*-test was 0.042, just slightly less than 0.05). It is likely that the mean strength of the unmodified formulation exceeds the mean strength of the modified formulation. Notice from Table 2.2 that both Minitab and JMP reported this confidence interval when the hypothesis testing procedure was conducted.

2.4.3 Choice of Sample Size

Selection of an appropriate sample size is one of the most important parts of any experimental design problem. One way to do this is to consider the impact of sample size on the estimate of the difference in two means. From Equation 2.30 we know that the $100(1 - \alpha)\%$ confidence interval on the difference in two means is a measure of the precision of estimation of the difference in the two means. The length of this interval is determined by

$$t_{\alpha/2, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

We consider the case where the sample sizes from the two populations are equal, so that $n_1 = n_2 = n$. Then the length of the CI is determined by

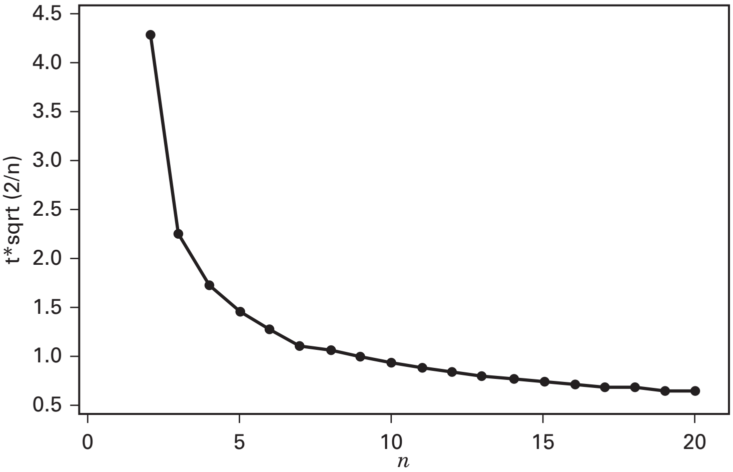
$$t_{\alpha/2, 2n-2} S_p \sqrt{\frac{2}{n}}$$

Consequently, the precision with which the difference in the two means is estimated depends on two quantities—S_p, over which we have no control, and $t_{\alpha/2,2n-2}\sqrt{2/n}$, which we can control by choosing the sample size n. Figure 2.12 is a plot of $t_{\alpha/2,2n-2}\sqrt{2/n}$ versus n for $\alpha = 0.05$. Notice that the curve descends rapidly as n increases up to about $n = 10$ and less rapidly beyond that. Since S_p is relatively constant and $t_{\alpha/2,2n-2}\sqrt{2/n}$ isn't going to change much for sample sizes beyond $n = 10$ or 12, we can conclude that choosing a sample size of $n = 10$ or 12 from each population in a two-sample 95 percent CI will result in a CI that results in about the best precision of estimation for the difference in the two means that is possible given the amount of inherent variability that is present in the two populations.

We can also use a hypothesis testing framework to determine sample size. The choice of sample size and the probability of type II error β are closely connected. Suppose that we are testing the hypotheses

$$
\begin{aligned}
H_0&{:}\mu_1 = \mu_2 \\
H_1&{:}\mu_1 \neq \mu_2
\end{aligned}
$$

and that the means are *not* equal so that $\delta = \mu_1 - \mu_2$. Because $H_0 : \mu_1 = \mu_2$ is not true, we are concerned about wrongly failing to reject H_0. The probability of type II error depends on the true difference in means δ. A graph



■ **FIGURE 2.12** **Plot of $t_{\alpha/2,2n-2}\sqrt{2/n}$ versus sample size in each population n for $\alpha = 0.05$.**

of β versus δ for a particular sample size is called the **operating characteristic curve**, or **OC curve** for the test. The β error is also a function of sample size. Generally, for a given value of δ, the β error decreases as the sample size increases. That is, a specified difference in means is easier to detect for larger sample sizes than for smaller ones.

An alternative to the OC curve is a **power curve**, which typically plots power or $1 - \beta$, versus sample size for a specified difference in the means. Some software packages perform power analysis and will plot power curves. A set of power curves constructed using JMP for the hypotheses

$$H_0{:}\mu_1 = \mu_2$$
$$H_1{:}\mu_1 \neq \mu_2$$

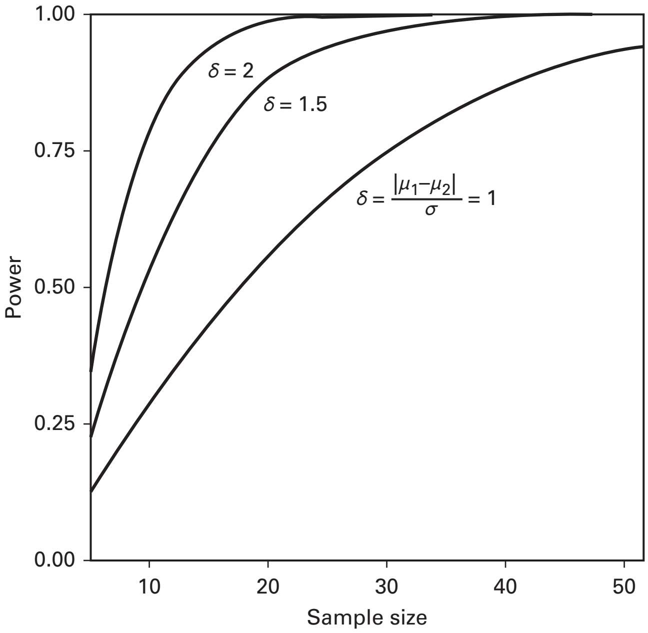
is shown in Figure 2.13 for the case where the two population variances σ_1^2 and σ_2^2 are unknown but equal ($\sigma_1^2 = \sigma_2^2 = \sigma^2$) and for a level of significance of $\alpha = 0.05$. These power curves also assume that the sample sizes from the two populations are equal and that the sample size shown on the horizontal scale (say n) is the total sample size, so that the sample size in each population is $n/2$. Also notice that the difference in means is expressed as a ratio to the common standard deviation; that is

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma}$$

From examining these curves, we observe the following:

1. The greater the difference in means $\mu_1 - \mu_2$, the higher the power (smaller type II error probability). That is, for a specified sample size and significance level α, the test will detect large differences in means more easily than small ones.
2. As the sample size gets larger, the power of the test gets larger (the type II error probability gets smaller) for a given difference in means and significance level α. That is, to detect a specified difference in means we may make the test more powerful by increasing the sample size.

Operating curves and power curves are often helpful in selecting a sample size to use in an experiment. For example, consider the Portland cement mortar problem discussed previously. Suppose that a difference in mean strength of $0.5\ \text{kgf/cm}^2$ has practical impact on the use of the mortar, so if the difference in means is at least this large, we would like to detect it with a high probability. Thus, because $\mu_1 - \mu_2 = 0.5\ \text{kgf/cm}^2$ is the "critical" difference in means



■ **FIGURE 2.13 Power curves (from JMP) for the two-sample *t*-test assuming equal variances and $\alpha = 0.05$. The sample size on the horizontal axis is the total sample size, so the sample size in each population is *n* = sample size from graph/2**

that we wish to detect, we find that the power curve parameter would be $\delta = 0.5/\sigma$. Unfortunately, δ involves the unknown standard deviation σ. However, suppose on the basis of past experience we think that it is very unlikely that the standard deviation will exceed 0.25 kgf/cm^2. Then substituting $\sigma = 0.25$ kgf/cm^2 into the expression for δ results in $\delta = 2$. If we wish to reject the null hypothesis when the difference in means $\mu_1 - \mu_2 = 0.5$ with probability at least 0.95 (power = 0.95) with $\alpha = 0.05$, then referring to Figure 2.13 we find that the required sample size on the horizontal axis is 16 approximately. This is the total sample size, so the sample size in each population should be

$$n = 16/2 = 8.$$

In our example, the experimenter actually used a sample size of 10. The experimenter could have elected to increase the sample size slightly to guard against the possibility that the prior estimate of the common standard deviation σ was too conservative and was likely to be somewhat larger than 0.25.

Operating characteristic curves often play an important role in the choice of sample size in experimental design problems. Their use in this respect is discussed in subsequent chapters. For a discussion of the uses of operating characteristic curves for other simple comparative experiments similar to the two-sample *t*-test, see Montgomery and Runger (2011).

Many statistics software packages can also assist the experimenter in performing power and sample size calculations. The following boxed display illustrates several computations for the Portland cement mortar problem from the power and sample size routine for the two-sample *t*-test in Minitab. The first section of output repeats the analysis performed with the OC curves; find the sample size necessary for detecting the critical difference in means of 0.5 kgf/cm^2, assuming that the standard deviation of strength is 0.25 kgf/cm^2. Notice that the answer obtained from Minitab, $n_1 = n_2 = 8$, is identical to the value obtained from the OC curve analysis. The second section of the output computes the power for the case where the critical difference in means is much smaller, only 0.25 kgf/cm^2. The power has dropped considerably, from over 0.95 to 0.562. The final section determines the sample sizes that would be necessary to detect an actual difference in means of 0.25 kgf/cm^2 with a power of at least 0.9. The required sample size turns out to be considerably larger, $n_1 = n_2 = 23$.

```
Power and Sample Size

2-Sample t-Test
Testing mean 1 = mean 2 (versus not = )
Calculating power for mean 1 = mean 2 + difference
Alpha = 0.05  Sigma = 0.25

                     Sample         Target         Actual
Difference             Size          Power          Power
0.5                       8         0.9500         0.9602

Power and Sample Size
2-Sample t-Test

Testing mean 1 = mean 2 (versus not =)
Calculating power for mean 1 = mean 2 + difference
Alpha = 0.05  Sigma = 0.25

                     Sample
Difference             Size          Power
0.25                     10         0.5620

Power and Sample Size
2-Sample t-Test

Testing mean 1 = mean 2 (versus not =)
Calculating power for mean 1 = mean 2 + difference
Alpha = 0.05  Sigma = 0.25

                     Sample         Target         Actual
Difference             Size          Power          Power
0.25                     23         0.9000         0.9125
```

2.4.4 The Case Where $\sigma_1^2 \neq \sigma_2^2$

If we are testing

$$H_0{:}\mu_1 = \mu_2$$
$$H_1{:}\mu_1 \neq \mu_2$$

and cannot reasonably assume that the variances σ_1^2 and σ_2^2 are equal, then the two-sample t-test must be modified slightly. The test statistic becomes

$$t_0 = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \tag{2.31}$$

This statistic is not distributed exactly as t. However, the distribution of t_0 is well approximated by t if we use

$$v = \frac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{(S_1^2/n_1)^2}{n_1 - 1} + \dfrac{(S_2^2/n_2)^2}{n_2 - 1}} \tag{2.32}$$

as the number of degrees of freedom. A strong indication of unequal variances on a normal probability plot would be a situation calling for this version of the t-test. You should be able to develop an equation for finding the confidence interval on the difference in mean for the unequal variances case easily.

EXAMPLE 2.1

Nerve preservation is important in surgery because accidental injury to the nerve can lead to post-surgical problems such as numbness, pain, or paralysis. Nerves are usually identified by their appearance and relationship to nearby structures or detected by local electrical stimulation (electromyography), but it is relatively easy to overlook them. An article in *Nature Biotechnology* ("Fluorescent Peptides Highlight Peripheral Nerves During Surgery in Mice," Vol. 29, 2011) describes the use of a fluorescently labeled peptide that binds to nerves to assist in identification. Table 2.3 shows the normalized fluorescence after two hours for nerve and muscle tissue for 12 mice (the data were read from a graph in the paper).

We would like to test the hypothesis that the mean normalized fluorescence after two hours is greater for nerve tissue than for muscle tissue. That is, if μ_1 is the mean normalized fluorescence for nerve tissue and μ_2 is the mean normalized fluorescence for muscle tissue, we want to test

$$H_0{:}\mu_1 = \mu_2$$
$$H_1{:}\mu_1 > \mu_2$$

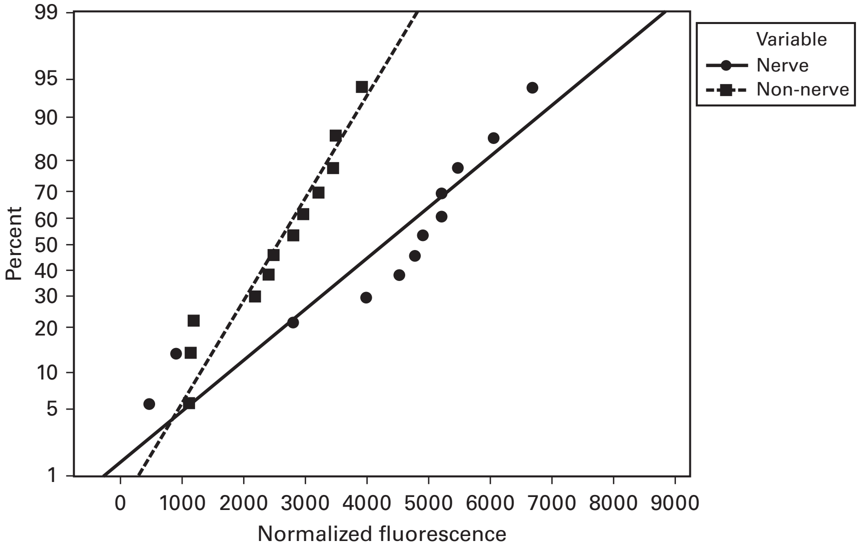
The descriptive statistics output from Minitab is shown below:

```
Variable      N   Mean  StDev  Minimum  Median  Maximum
Nerve        12   4228   1918      450    4825     6625
Non-nerve    12   2534    961     1130    2650     3900
```

■ TABLE 2.3
Normalized Fluorescence After Two Hours

Observation	Nerve	Muscle
1	6625	3900
2	6000	3500
3	5450	3450
4	5200	3200
5	5175	2980
6	4900	2800
7	4750	2500
8	4500	2400
9	3985	2200
10	900	1200
11	450	1150
12	2800	1130

■ **FIGURE 2.14** **Normalized fluorescence data from Table 2.3**



Notice that the two sample standard deviations are quite different, so the assumption of equal variances in the pooled *t*-test may not be appropriate. Figure 2.14 is the normal probability plot from Minitab for the two samples. This plot also indicates that the two population variances are probably not the same.

Because the equal variance assumption is not appropriate here, we will use the two-sample *t*-test described in this section to test the hypothesis of equal means. The test statistic, Equation 2.31, is

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} = \frac{4228 - 2534}{\sqrt{\dfrac{(1918)^2}{12} + \dfrac{(961)^2}{12}}} = 2.7354$$

The number of degrees of freedom are calculated from Equation 2.32:

$$v = \frac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{(S_1^2/n_1)^2}{n_1 - 1} + \dfrac{(S_2^2/n_2)^2}{n_2 - 1}} = \frac{\left(\dfrac{(1918)^2}{12} + \dfrac{(961)^2}{12}\right)^2}{\dfrac{[(1918)^2/12]^2}{11} + \dfrac{[(961)^2/12]^2}{11}} = 16.1955$$

If we are going to find a *P*-value from a table of the *t*-distribution, we should round the degrees of freedom down to 16. Most computer programs interpolate to determine the *P*-value. The Minitab output for the two-sample *t*-test is shown below. Since the *P*-value reported is small (0.007), we would reject the null hypothesis and conclude that the mean normalized fluorescence for nerve tissue is greater than the mean normalized fluorescence for muscle tissue.

```
Difference = mu (Nerve) - mu (Non-nerve)
Estimate for difference: 1694
95% lower bound for difference:  613
T-Test of difference = 0 (vs >): T-Value = 2.74 P-Value = 0.007 DF = 16
```

2.4.5 The Case Where σ_1^2 and σ_2^2 Are Known

If the variances of both populations are **known**, then the hypotheses

$$H_0{:}\mu_1 = \mu_2$$
$$H_1{:}\mu_1 \neq \mu_2$$

may be tested using the statistic

$$Z_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \tag{2.33}$$

If both populations are normal, or if the sample sizes are large enough so that the central limit theorem applies, the distribution of Z_0 is $N(0, 1)$ if the null hypothesis is true. Thus, the critical region would be found using the normal distribution rather than the t. Specifically, we would reject H_0 if $|Z_0| > Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution. This procedure is sometimes called the **two-sample *Z*-test**. A P-value approach can also be used with this test. The P-value would be found as $P = 2[1 - \Phi(|Z_0|)]$, where $\Phi(x)$ is the cumulative standard normal distribution evaluated at the point x.

Unlike the t-test of the previous sections, the test on means with known variances does not require the assumption of sampling from normal populations. One can use the central limit theorem to justify an approximate normal distribution for the difference in sample means $\bar{y}_1 - \bar{y}_2$.

The $100(1 - \alpha)$ percent confidence interval on $\mu_1 - \mu_2$ where the variances are known is

$$\bar{y}_1 - \bar{y}_2 - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{y}_1 - \bar{y}_2 + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{2.34}$$

As noted previously, the confidence interval is often a useful supplement to the hypothesis testing procedure.

2.4.6 Comparing a Single Mean to a Specified Value

Some experiments involve comparing only one population mean μ to a specified value, say, μ_0. The hypotheses are

$$H_0{:}\mu = \mu_0$$
$$H_1{:}\mu \neq \mu_0$$

If the population is normal with known variance, or if the population is nonnormal but the sample size is large enough so that the central limit theorem applies, then the hypothesis may be tested using a direct application of the normal distribution. The **one-sample *Z*-test** statistic is

$$Z_0 = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \tag{2.35}$$

If $H_0 : \mu = \mu_0$ is true, then the distribution of Z_0 is $N(0, 1)$. Therefore, the decision rule for $H_0 : \mu = \mu_0$ is to reject the null hypothesis if $|Z_0| > Z_{\alpha/2}$. A P-value approach could also be used.

The value of the mean μ_0 specified in the null hypothesis is usually determined in one of three ways. It may result from past evidence, knowledge, or experimentation. It may be the result of some theory or model describing the situation under study. Finally, it may be the result of contractual specifications.

The $100(1-\alpha)$ percent confidence interval on the true population mean is

$$\bar{y} - Z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{y} + Z_{\alpha/2}\sigma/\sqrt{n} \tag{2.36}$$

EXAMPLE 2.2

A supplier submits lots of fabric to a textile manufacturer. The customer wants to know if the lot average breaking strength exceeds 200 psi. If so, she wants to accept the lot. Past experience indicates that a reasonable value for the variance of breaking strength is 100(psi)^2. The hypotheses to be tested are

$$H_0 : \mu = 200$$
$$H_1 : \mu > 200$$

Note that this is a one-sided alternative hypothesis. Thus, we would accept the lot only if the null hypothesis $H_0 : \mu = 200$ could be rejected (i.e., if $Z_0 > Z_\alpha$).

Four specimens are randomly selected, and the average breaking strength observed is $\bar{y} = 214$ psi. The value of the test statistic is

$$Z_0 = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{124 - 200}{10/\sqrt{4}} = 2.80$$

If a type I error of $\alpha = 0.05$ is specified, we find $Z_\alpha = Z_{0.05} = 1.645$ from Appendix Table I. The P-value would be computed using only the area in the upper tail of the standard normal distribution, because the alternative hypothesis is one-sided. The P-value is $P = 1 - \Phi(2.80) = 1 - 0.99744 = 0.00256$. Thus H_0 is rejected, and we conclude that the lot average breaking strength exceeds 200 psi.

If the variance of the population is unknown, we must make the additional assumption that the population is normally distributed, although moderate departures from normality will not seriously affect the results.

To test $H_0 : \mu = \mu_0$ in the variance unknown case, the sample variance S^2 is used to estimate σ^2. Replacing σ with S in Equation 2.35, we have the **one-sample *t*-test** statistic

$$t_0 = \frac{\bar{y} - \mu_0}{S/\sqrt{n}} \tag{2.37}$$

The null hypothesis $H_0 : \mu = \mu_0$ would be rejected if $|t_0| > t_{\alpha/2,n-1}$, where $t_{\alpha/2,n-1}$ denotes the upper $\alpha/2$ percentage point of the t distribution with $n-1$ degrees of freedom. A P-value approach could also be used. The $100(1-\alpha)$ percent confidence interval in this case is

$$\bar{y} - t_{\alpha/2,n-1}S/\sqrt{n} \leq \mu \leq \bar{y} + t_{\alpha/2,n-1}S/\sqrt{n} \tag{2.38}$$

2.4.7 Summary

Tables 2.4 and 2.5 summarize the t-test and z-test procedures discussed above for sample means. Critical regions are shown for both two-sided and one-sided alternative hypotheses.

■ **TABLE 2.4**
Tests on Means with Variance Known

Hypothesis	Test Statistic	Fixed Significance Level Criteria for Rejection	P-Value
$H_0{:}\mu = \mu_0$ $H_1{:}\mu \neq \mu_0$	$Z_0 = \dfrac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$	$\lvert Z_0 \rvert > Z_{\alpha/2}$	$P = 2[1 - \Phi(\lvert Z_0 \rvert)]$
$H_0{:}\mu = \mu_0$ $H_1{:}\mu < \mu_0$		$Z_0 < -Z_\alpha$	$P = \Phi(Z_0)$
$H_0{:}\mu = \mu_0$ $H_1{:}\mu > \mu_0$		$Z_0 > Z_\alpha$	$P = 1 - \Phi(Z_0)$
$H_0{:}\mu_1 = \mu_2$ $H_1{:}\mu_1 \neq \mu_2$	$Z_0 = \dfrac{\bar{y}_1 - \bar{y}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$\lvert Z_0 \rvert > Z_{\alpha/2}$	$P = 2[1 - \Phi(\lvert Z_0 \rvert)]$
$H_0{:}\mu_1 = \mu_2$ $H_1{:}\mu_1 < \mu_2$		$Z_0 < -Z_\alpha$	$P = \Phi(Z_0)$
$H_0{:}\mu_1 = \mu_2$ $H_1{:}\mu_1 > \mu_2$		$Z_0 > Z_\alpha$	$P = 1 - \Phi(Z_0)$

■ **TABLE 2.5**
Tests on Means of Normal Distributions, Variance Unknown

Hypothesis	Test Statistic	Fixed Significance Level Criteria for Rejection	P-Value
$H_0{:}\mu = \mu_0$ $H_1{:}\mu \neq \mu_0$	$t_0 = \dfrac{\bar{y} - \mu_0}{S/\sqrt{n}}$	$\lvert t_0 \rvert > t_{\alpha/2,n-1}$	sum of the probability above t_0 and below $-t_0$
$H_0{:}\mu = \mu_0$ $H_1{:}\mu < \mu_0$		$t_0 < -t_{\alpha,n-1}$	probability below t_0
$H_0{:}\mu = \mu_0$ $H_1{:}\mu > \mu_0$		$t_0 > t_{\alpha,n-1}$	probability above t_0
	if $\sigma_1^2 = \sigma_2^2$		
$H_0{:}\mu_1 = \mu_2$ $H_1{:}\mu_1 \neq \mu_2$	$t_0 = \dfrac{\bar{y}_1 - \bar{y}_2}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ $v = n_1 + n_2 - 2$	$\lvert t_0 \rvert > t_{\alpha/2,v}$	sum of the probability above t_0 and below $-t_0$
	if $\sigma_1^2 \neq \sigma_2^2$		
$H_0{:}\mu_1 = \mu_2$ $H_1{:}\mu_1 < \mu_2$	$t_0 = \dfrac{\bar{y}_1 - \bar{y}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$	$t_0 < -t_{\alpha,v}$	probability below t_0
$H_0{:}\mu_1 = \mu_2$ $H_1{:}\mu_1 > \mu_2$	$v = \dfrac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{(S_1^2/n_1)^2}{n_1 - 1} + \dfrac{(S_2^2/n_2)^2}{n_2 - 1}}$	$t_0 > t_{\alpha,v}$	probability above t_0

2.5 Inferences About the Differences in Means, Paired Comparison Designs

2.5.1 The Paired Comparison Problem

In some simple comparative experiments, we can greatly improve the precision by making comparisons within matched pairs of experimental material. For example, consider a hardness testing machine that presses a rod with a pointed tip into a metal specimen with a known force. By measuring the depth of the depression caused by the tip, the hardness of the specimen is determined. Two different tips are available for this machine, and although the precision (variability) of the measurements made by the two tips seems to be the same, it is suspected that one tip produces different mean hardness readings than the other.

An experiment could be performed as follows. A number of metal specimens (e.g., 20) could be randomly selected. Half of these specimens could be tested by tip 1 and the other half by tip 2. The exact assignment of specimens to tips would be randomly determined. Because this is a completely randomized design, the average hardness of the two samples could be compared using the *t*-test described in Section 2.4.

A little reflection will reveal a serious disadvantage in the completely randomized design for this problem. Suppose the metal specimens were cut from different bar stock that were produced in different heats or that were not exactly homogeneous in some other way that might affect the hardness. This lack of homogeneity between specimens will contribute to the variability of the hardness measurements and will tend to inflate the experimental error, thus making a true difference between tips harder to detect.

To protect against this possibility, consider an alternative experimental design. Assume that each specimen is large enough so that *two* hardness determinations may be made on it. This alternative design would consist of dividing each specimen into two parts, then randomly assigning one tip to one-half of each specimen and the other tip to the remaining half. The order in which the tips are tested for a particular specimen would also be randomly selected. The experiment, when performed according to this design with 10 specimens, produced the (coded) data shown in Table 2.6.

We may write a **statistical model** that describes the data from this experiment as

$$y_{ij} = \mu_i + \beta_j + \epsilon_{ij} \begin{cases} i = 1, 2 \\ j = 1, 2, \ldots, 10 \end{cases} \tag{2.39}$$

where y_{ij} is the observation on hardness for tip i on specimen j, μ_i is the true mean hardness of the ith tip, β_j is an effect on hardness due to the jth specimen, and ϵ_{ij} is a random experimental error with mean zero and variance σ_i^2. That is, σ_1^2 is the variance of the hardness measurements from tip 1, and σ_2^2 is the variance of the hardness measurements from tip 2.

■ **TABLE 2.6**
Data for the Hardness Testing Experiment

Specimen	Tip 1	Tip 2
1	7	6
2	3	3
3	3	5
4	4	3
5	8	8
6	3	2
7	2	4
8	9	9
9	5	4
10	4	5

Note that if we compute the *j*th paired difference

$$d_j = y_{1j} - y_{2j} \qquad j = 1, 2, \ldots, 10 \tag{2.40}$$

the expected value of this difference is

$$\begin{aligned}\mu_d &= E(d_j)\\ &= E(y_{1j} - y_{2j})\\ &= E(y_{1j}) - E(y_{2j})\\ &= \mu_1 + \beta_j - (\mu_2 + \beta_j)\\ &= \mu_1 - \mu_2\end{aligned}$$

That is, we may make inferences about the difference in the mean hardness readings of the two tips $\mu_1 - \mu_2$ by making inferences about the mean of the differences μ_d. Notice that the additive effect of the specimens β_j cancels out when the observations are paired in this manner.

Testing $H_0 : \mu_1 = \mu_2$ is equivalent to testing

$$\begin{aligned}H_0&{:}\mu_d = 0\\ H_1&{:}\mu_d \neq 0\end{aligned}$$

This is a single-sample *t*-test. The test statistic for this hypothesis is

$$t_0 = \frac{\bar{d}}{S_d/\sqrt{n}} \tag{2.41}$$

where

$$\bar{d} = \frac{1}{n}\sum_{j=1}^{n} d_j \tag{2.42}$$

is the sample mean of the differences and

$$S_d = \left[\frac{\sum_{j=1}^{n}(d_j - \bar{d})^2}{n-1}\right]^{1/2} = \left[\frac{\sum_{j=1}^{n} d_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} d_j\right)^2}{n-1}\right]^{1/2} \tag{2.43}$$

is the sample standard deviation of the differences. $H_0 : \mu_d = 0$ would be rejected if $|t_0| > t_{\alpha/2,n-1}$. A *P*-value approach could also be used. Because the observations from the factor levels are "paired" on each experimental unit, this procedure is usually called the **paired *t*-test**.

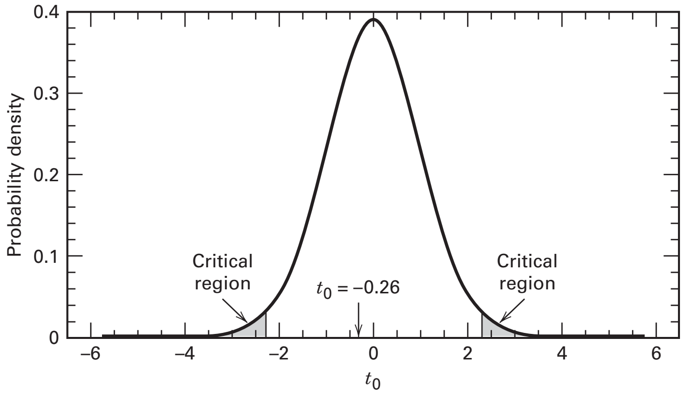
For the data in Table 2.6, we find

$$\begin{aligned}d_1 &= 7 - 6 = 1 & d_6 &= 3 - 2 = 1\\ d_2 &= 3 - 3 = 0 & d_7 &= 2 - 4 = -2\\ d_3 &= 3 - 5 = -2 & d_8 &= 9 - 9 = 0\\ d_4 &= 4 - 3 = 1 & d_9 &= 5 - 4 = 1\\ d_5 &= 8 - 8 = 0 & d_{10} &= 4 - 5 = -1\end{aligned}$$

Thus,

$$\bar{d} = \frac{1}{n}\sum_{j=1}^{n} d_j = \frac{1}{10}(-1) = -0.10$$

$$S_d = \left[\frac{\sum_{j=1}^{n} d_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} d_j\right)^2}{n-1}\right]^{1/2} = \left[\frac{13 - \frac{1}{10}(-1)^2}{10-1}\right]^{1/2} = 1.20$$



■ **FIGURE 2.15 The reference distribution (*t* with 9 degrees of freedom) for the hardness testing problem**

Suppose we choose $\alpha = 0.05$. Now to make a decision, we would compute t_0 and reject H_0 if $|t_0| > t_{0.025,9} = 2.262$.

The computed value of the paired *t*-test statistic is

$$t_0 = \frac{\overline{d}}{S_d/\sqrt{n}} = \frac{-0.10}{1.20/\sqrt{10}} = -0.26$$

and because $|t_0| = 0.26 \ngtr t_{0.025,9} = 2.262$, we cannot reject the hypothesis $H_0 : \mu_d = 0$. That is, there is no evidence to indicate that the two tips produce different hardness readings. Figure 2.15 shows the t_0 distribution with 9 degrees of freedom, the reference distribution for this test, with the value of t_0 shown relative to the critical region.

Table 2.7 shows the computer output from the Minitab paired *t*-test procedure for this problem. Notice that the *P*-value for this test is $P \simeq 0.80$, implying that we cannot reject the null hypothesis at *any* reasonable level of significance.

2.5.2 Advantages of the Paired Comparison Design

The design actually used for this experiment is called the **paired comparison design**, and it illustrates the blocking principle discussed in Section 1.3. Actually, it is a special case of a more general type of design called the **randomized block design**. The term *block* refers to a relatively homogeneous experimental unit (in our case, the metal specimens are the blocks), and the block represents a restriction on complete randomization because the treatment combinations are only randomized within the block. We look at designs of this type in Chapter 4. In that chapter, the mathematical model for the design, Equation 2.39, is written in a slightly different form.

Before leaving this experiment, several points should be made. Note that, although $2n = 2(10) = 20$ observations have been taken, only $n - 1 = 9$ degrees of freedom are available for the *t* statistic. (We know that as the degrees of

■ **TABLE 2.7**
Minitab Paired *t*-Test Results for the Hardness Testing Example

```
Paired T for Tip 1-Tip 2

                     N          Mean       Std. Dev.      SE Mean
Tip 1               10         4.800           2.394        0.757
Tip 2               10         4.900           2.234        0.706
Difference          10        -0.100           1.197        0.379

95% CI for mean difference: (-0.956, 0.756)
t-Test of mean difference = 0 (vs not = 0):
T-Value = -0.26 P-Value = 0.798
```

freedom for *t* increase, the test becomes more sensitive.) By blocking or pairing we have effectively "lost" $n - 1$ degrees of freedom, but we hope we have gained a better knowledge of the situation by eliminating an additional source of variability (the difference between specimens).

We may obtain an indication of the quality of information produced from the paired design by comparing the standard deviation of the differences S_d with the pooled standard deviation S_p that would have resulted had the experiment been conducted in a completely randomized manner and the data of Table 2.5 been obtained. Using the data in Table 2.5 as two independent samples, we compute the pooled standard deviation from Equation 2.25 to be $S_p = 2.32$. Comparing this value to $S_d = 1.20$, we see that blocking or pairing has reduced the estimate of variability by nearly 50 percent.

Generally, when we don't block (or pair the observations) when we really should have, S_p will always be larger than S_d. It is easy to show this formally. If we pair the observations, it is easy to show that S_d^2 is an unbiased estimator of the variance of the differences d_j under the model in Equation 2.39 because the block effects (the β_j) cancel out when the differences are computed. However, if we don't block (or pair) and treat the observations as two independent samples, then S_p^2 is not an unbiased estimator of σ^2 under the model in Equation 2.39. In fact, assuming that both population variances are equal,

$$E(S_p^2) = \sigma^2 + \sum_{j=1}^{n} \beta_j^2$$

That is, the block effects β_j inflate the variance estimate. This is why blocking serves as a **noise reduction** design technique.

We may also express the results of this experiment in terms of a confidence interval on $\mu_1 - \mu_2$. Using the paired data, a 95 percent confidence interval on $\mu_1 - \mu_2$ is

$$\begin{gathered}\overline{d} \pm t_{0.025,9} S_d/\sqrt{n} \\ -0.10 \pm (2.262)(1.20)/\sqrt{10} \\ -0.10 \pm 0.86\end{gathered}$$

Conversely, using the pooled or independent analysis, a 95 percent confidence interval on $\mu_1 - \mu_2$ is

$$\begin{gathered}\overline{y}_1 - \overline{y}_2 \pm t_{0.025,18} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ 4.80 - 4.90 \pm (2.101)(2.32)\sqrt{\tfrac{1}{10} + \tfrac{1}{10}} \\ -0.10 \pm 2.18\end{gathered}$$

The confidence interval based on the paired analysis is much narrower than the confidence interval from the independent analysis. This again illustrates the **noise reduction** property of blocking.

Blocking is not always the best design strategy. If the within-block variability is the same as the between-block variability, the variance of $\overline{y}_1 - \overline{y}_2$ will be the same regardless of which design is used. Actually, blocking in this situation would be a poor choice of design because blocking results in the loss of $n - 1$ degrees of freedom and will actually lead to a wider confidence interval on $\mu_1 - \mu_2$. A further discussion of blocking is given in Chapter 4.

2.6 Inferences About the Variances of Normal Distributions

In many experiments, we are interested in possible differences in the mean response for two treatments. However, in some experiments it is the comparison of variability in the data that is important. In the food and beverage industry, for example, it is important that the variability of filling equipment be small so that all packages have close to the nominal net weight or volume of content. In chemical laboratories, we may wish to compare the variability of two analytical methods. We now briefly examine tests of hypotheses and confidence intervals for variances of normal distributions. Unlike the tests on means, the procedures for tests on variances are rather sensitive to the normality assumption. A good discussion of the normality assumption is in Appendix 2A of Davies (1956).

Suppose we wish to test the hypothesis that the variance of a normal population equals a constant, for example, σ_0^2. Stated formally, we wish to test

$$\begin{aligned} H_0&:\sigma^2 = \sigma_0^2 \\ H_1&:\sigma^2 \neq \sigma_0^2 \end{aligned} \tag{2.44}$$

The test statistic for Equation 2.44 is

$$\chi_0^2 = \frac{SS}{\sigma_0^2} = \frac{(n-1)S^2}{\sigma_0^2} \tag{2.45}$$

where $SS = \sum_{i=1}^{n} (y_i - \overline{y})^2$ is the corrected sum of squares of the sample observations. The appropriate reference distribution for χ_0^2 is the chi-square distribution with $n-1$ degrees of freedom. The null hypothesis is rejected if $\chi_0^2 > \chi^2_{\alpha/2,n-1}$ or if $\chi_0^2 < \chi^2_{1-(\alpha/2),n-1}$, where $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-(\alpha/2),n-1}$ are the upper $\alpha/2$ and lower $1-(\alpha/2)$ percentage points of the chi-square distribution with $n-1$ degrees of freedom, respectively. Table 2.8 gives the critical regions for the one-sided alternative hypotheses. The $100(1-\alpha)$ percent confidence interval on σ^2 is

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-(\alpha/2),n-1}} \tag{2.46}$$

Now consider testing the equality of the variances of two normal populations. If independent random samples of size n_1 and n_2 are taken from populations 1 and 2, respectively, the test statistic for

$$\begin{aligned} H_0&:\sigma_1^2 = \sigma_2^2 \\ H_1&:\sigma_1^2 \neq \sigma_2^2 \end{aligned} \tag{2.47}$$

is the ratio of the sample variances

$$F_0 = \frac{S_1^2}{S_2^2} \tag{2.48}$$

The appropriate reference distribution for F_0 is the F distribution with $n_1 - 1$ numerator degrees of freedom and $n_2 - 1$ denominator degrees of freedom. The null hypothesis would be rejected if $F_0 > F_{\alpha/2,n_1-1,n_2-1}$ or if

■ **TABLE 2.8**

Tests on Variances of Normal Distributions

Hypothesis	Test Statistic	Fixed Significance Level Criteria for Rejection
$H_0:\sigma^2 = \sigma_0^2$ $H_1:\sigma^2 \neq \sigma_0^2$	$\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$	$\chi_0^2 > \chi^2_{\alpha/2,n-1}$ or $\chi_0^2 < \chi^2_{1-\alpha/2,n-1}$
$H_0:\sigma^2 = \sigma_0^2$ $H_1:\sigma^2 < \sigma_0^2$		$\chi_0^2 < \chi^2_{1-\alpha,n-1}$
$H_0:\sigma^2 = \sigma_0^2$ $H_1:\sigma^2 > \sigma_0^2$		$\chi_0^2 > \chi^2_{\alpha,n-1}$
$H_0:\sigma_1^2 = \sigma_2^2$ $H_1:\sigma_1^2 \neq \sigma_2^2$	$F_0 = \frac{S_1^2}{S_2^2}$	$F_0 > F_{\alpha/2,n_1-1,n_2-1}$ or $F_0 < F_{1-\alpha/2,n_1-1,n_2-1}$
$H_0:\sigma_1^2 = \sigma_2^2$ $H_1:\sigma_1^2 < \sigma_2^2$	$F_0 = \frac{S_2^2}{S_1^2}$	$F_0 > F_{\alpha,n_2-1,n_1-1}$
$H_0:\sigma_1^2 = \sigma_2^2$ $H_1:\sigma_1^2 > \sigma_2^2$	$F_0 = \frac{S_1^2}{S_2^2}$	$F_0 > F_{\alpha,n_1-1,n_2-1}$

$F_0 < F_{1-(\alpha/2),n_1-1,n_2-1}$, where $F_{\alpha/2,n_1-1,n_2-1}$ and $F_{1-(\alpha/2),n_1-1,n_2-1}$ denote the upper $\alpha/2$ and lower $1-(\alpha/2)$ percentage points of the F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. Table IV of the Appendix gives only upper-tail percentage points of F; however, the upper- and lower-tail points are related by

$$F_{1-\alpha,v_1,v_2} = \frac{1}{F_{\alpha,v_2,v_1}} \tag{2.49}$$

Critical values for the one-sided alternative hypothesis are given in Table 2.8. Test procedures for more than two variances are discussed in Section 3.4.3. We will also discuss the use of the variance or standard deviation as a response variable in more general experimental settings.

EXAMPLE 2.3

A chemical engineer is investigating the inherent variability of two types of test equipment that can be used to monitor the output of a production process. He suspects that the old equipment, type 1, has a larger variance than the new one. Thus, he wishes to test the hypothesis

$$H_0{:}\sigma_1^2 = \sigma_2^2$$
$$H_1{:}\sigma_1^2 > \sigma_2^2$$

Two random samples of $n_1 = 12$ and $n_2 = 10$ observations are taken, and the sample variances are $S_1^2 = 14.5$ and $S_2^2 = 10.8$. The test statistic is

$$F_0 = \frac{S_1^2}{S_2^2} = \frac{14.5}{10.8} = 1.34$$

From Appendix Table IV we find that $F_{0.05,11,9} = 3.10$, so the null hypothesis cannot be rejected. That is, we have found insufficient statistical evidence to conclude that the variance of the old equipment is greater than the variance of the new equipment.

The $100(1 - \alpha)$ confidence interval for the ratio of the population variances σ_1^2/σ_2^2 is

$$\frac{S_1^2}{S_2^2}F_{1-\alpha/2,n_2-1,n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}F_{\alpha/2,n_2-1,n_1-1} \tag{2.50}$$

To illustrate the use of Equation 2.50, the 95 percent confidence interval for the ratio of variances σ_1^2/σ_2^2 in Example 2.2 is, using $F_{0.025,9,11} = 3.59$ and $F_{0.975,9,11} = 1/F_{0.025,11,9} = 1/3.92 = 0.255$,

$$\frac{14.5}{10.8}(0.225) \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{14.5}{10.8}(3.59)$$

$$0.34 \le \frac{\sigma_1^2}{\sigma_2^2} \le 4.82$$

2.7 Problems

2.1 Computer output for a random sample of data is shown below. Some of the quantities are missing. Compute the values of the missing quantities.

```
Variable N  Mean SE Mean Std. Dev. Variance Minimum Maximum
  Y      9 19.96   ?      3.12       ?       15.94   27.16
```

2.2 Computer output for a random sample of data is shown below. Some of the quantities are missing. Compute the values of the missing quantities.

```
Variable      N     Mean    SE Mean    Std. Dev.     Sum
  Y          16      ?       0.159        ?        399.851
```

2.3 Suppose that we are testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. Calculate the *P*-value for the following observed values of the test statistic:

(a) $Z_0 = 2.25$ **(b)** $Z_0 = 1.55$ **(c)** $Z_0 = 2.10$

(d) $Z_0 = 1.95$ **(e)** $Z_0 = -0.10$

2.4 Suppose that we are testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$. Calculate the *P*-value for the following observed values of the test statistic:

(a) $Z_0 = 2.45$ **(b)** $Z_0 = -1.53$ **(c)** $Z_0 = 2.15$

(d) $Z_0 = 1.95$ **(e)** $Z_0 = -0.25$

2.5 Consider the computer output shown below.

One-Sample Z

```
Test of mu = 30 vs not = 30
The assumed standard deviation = 1.2
 N      Mean   SE Mean        95% CI          Z    P
16   31.2000    0.3000  (30.6120, 31.7880)    ?    ?
```

(a) Fill in the missing values in the output. What conclusion would you draw?

(b) Is this a one-sided or two-sided test?

(c) Use the output and the normal table to find a 99 percent CI on the mean.

(d) What is the *P*-value if the alternative hypothesis is $H_1 : \mu > 30$?

2.6 Suppose that we are testing $H_0 : \mu_1 = \mu_2$ versus $H_0 : \mu_1 \neq \mu_2$ where the two sample sizes are $n_1 = n_2 = 12$. Both sample variances are unknown but assumed equal. Find bounds on the *P*-value for the following observed values of the test statistic.

(a) $t_0 = 2.30$ **(b)** $t_0 = 3.41$ **(c)** $t_0 = 1.95$ **(d)** $t_0 = -2.45$

2.7 Suppose that we are testing $H_0 : \mu_1 = \mu_2$ versus $H_0 : \mu_1 > \mu_2$ where the two sample sizes are $n_1 = n_2 = 10$. Both sample variances are unknown but assumed equal. Find bounds on the *P*-value for the following observed values of the test statistic.

(a) $t_0 = 2.31$ **(b)** $t_0 = 3.60$ **(c)** $t_0 = 1.95$ **(d)** $t_0 = 2.19$

2.8 Consider the following sample data: 9.37, 13.04, 11.69, 8.21, 11.18, 10.41, 13.15, 11.51, 13.21, and 7.75. Is it reasonable to assume that this data is a sample from a normal distribution? Is there evidence to support a claim that the mean of the population is 10?

2.9 A computer program has produced the following output for a hypothesis-testing problem:

```
Difference in sample means: 2.35
Degrees of freedom: 18
Standard error of the difference in sample means: ?
Test statistic: t0 = 2.01
P-value: 0.0298
```

(a) What is the missing value for the standard error?

(b) Is this a two-sided or a one-sided test?

(c) If $\alpha = 0.05$, what are your conclusions?

(d) Find a 90% two-sided CI on the difference in means.

2.10 A computer program has produced the following output for a hypothesis-testing problem:

```
Difference in sample means: 11.5
Degrees of freedom: 24
Standard error of the difference in sample means: ?
Test statistic: t0 = -1.88
P-value: 0.0723
```

(a) What is the missing value for the standard error?

(b) Is this a two-sided or a one-sided test?

(c) If $\alpha = 0.05$, what are your conclusions?

(d) Find a 95% two-sided CI on the difference in means.

2.11 A two-sample *t*-test has been conducted and the sample sizes are $n_1 = n_2 = 10$. The computed value of the test statistic is $t_0 = 2.15$. If the null hypothesis is two-sided, an upper bound on the *P*-value is

(a) 0.10 **(b)** 0.05 **(c)** 0.025

(d) 0.01 **(e)** none of the above.

2.12 A two-sample *t*-test has been conducted and the sample sizes are $n_1 = n_2 = 12$ The computed value of the test statistic is $t_0 = 2.27$. If the null hypothesis is two-sided, an upper bound on the *P*-value is

(a) 0.10 **(b)** 0.05 **(c)** 0.025

(d) 0.01 **(e)** none of the above.

2.13 Suppose that we are testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$ with a sample size of $n = 15$. Calculate bounds on the *P*-value for the following observed values of the test statistic:

(a) $t_0 = 2.35$ **(b)** $t_0 = 3.55$ **(c)** $t_0 = 2.00$ **(d)** $t_0 = 1.55$

2.14 Suppose that we are testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ with a sample size of $n = 10$. Calculate bounds on the *P*-value for the following observed values of the test statistic:

(a) $t_0 = 2.48$ **(b)** $t_0 = -3.95$ **(c)** $t_0 = 2.69$

(d) $t_0 = 1.88$ **(e)** $t_0 = -1.25$

2.15 Consider the computer output shown below.

```
One-Sample T: Y
Test of mu = 91 vs. not = 91
Variable  N  Mean   Std. Dev. SE Mean    95% CI       T    P
  Y      25 92.5805     ?      0.4673 (91.6160, ?) 3.38 0.002
```

(a) Fill in the missing values in the output. Can the null hypothesis be rejected at the 0.05 level? Why?

(b) Is this a one-sided or a two-sided test?

(c) If the hypotheses had been $H_0 : \mu = 90$ versus $H_1 : \mu \neq 90$, would you reject the null hypothesis at the 0.05 level?,

(d) Use the output and the *t* table to find a 99 percent two-sided CI on the mean.

(e) What is the *P*-value if the alternative hypothesis is $H_1 : \mu > 91$?

2.16 Consider the computer output shown below.

```
One-Sample T: Y

Test of mu = 25 vs. > 25
                                                95% Lower
Variable    N  Mean    Std. Dev.  SE Mean   Bound   T     P
Y          12  25.6818     ?       0.3360     ?     ?   0.034
```

(a) How many degrees of freedom are there on the *t*-test statistic?

(b) Fill in the missing information.

2.17 Consider the computer output shown below.

```
Two-Sample T-Test and Cl: Y1, Y2
Two-sample T for Y1 vs Y2
          N      Mean      Std. Dev.     SE Mean
Y1       20     50.19        1.71          0.38
Y2       20     52.52        2.48          0.55

Difference = mu (X1) - mu (X2)
Estimate for difference: - 2.33341
95% CI for difference: (- 3.69547, - 0.97135)
T-Test of difference = 0 (vs not =): T-Value = -3.47
P-Value = 0.001 DF = 38
Both use Pooled Std. Dev. = 2.1277
```

(a) Can the null hypothesis be rejected at the 0.05 level? Why?

(b) Is this a one-sided or a two-sided test?

(c) If the hypotheses had been $H_0 : \mu_1 - \mu_2 = 2$ versus $H_1 : \mu_1 - \mu_2 \neq 2$, would you reject the null hypothesis at the 0.05 level?,

(d) If the hypotheses had been $H_0 : \mu_1 - \mu_2 = 2$ versus $H_1 : \mu_1 - \mu_2 < 2$, would you reject the null hypothesis at the 0.05 level? Can you answer this question without doing any additional calculations? Why?

(e) Use the output and the *t* table to find a 95 percent upper confidence bound on the difference in means.

(f) What is the *P*-value if the hypotheses are $H_0 : \mu_1 - \mu_2 = 2$ versus $H_1 : \mu_1 - \mu_2 \neq 2$?

2.18 The breaking strength of a fiber is required to be at least 150 psi. Past experience has indicated that the standard deviation of breaking strength is $\sigma = 3$ psi. A random sample of four specimens is tested, and the results are $y_1 = 145, y_2 = 153, y_3 = 150,$ and $y_4 = 147$.

(a) State the hypotheses that you think should be tested in this experiment.

(b) Test these hypotheses using $\alpha = 0.05$. What are your conclusions?

(c) Find the *P*-value for the test in part (b).

(d) Construct a 95 percent confidence interval on the mean breaking strength.

2.19 The viscosity of a liquid detergent is supposed to average 800 centistokes at 25°C. A random sample of 16 batches of detergent is collected, and the average viscosity is 812. Suppose we know that the standard deviation of viscosity is $\sigma = 25$ centistokes.

(a) State the hypotheses that should be tested.

(b) Test these hypotheses using $\alpha = 0.05$. What are your conclusions?

(c) What is the *P*-value for the test?

(d) Find a 95 percent confidence interval on the mean.

2.20 The diameters of steel shafts produced by a certain manufacturing process should have a mean diameter of 0.255 inches. The diameter is known to have a standard deviation of $\sigma = 0.0001$ inch. A random sample of 10 shafts has an average diameter of 0.2545 inches.

(a) Set up appropriate hypotheses on the mean μ.

(b) Test these hypotheses using $\alpha = 0.05$. What are your conclusions?

(c) Find the P-value for this test.

(d) Construct a 95 percent confidence interval on the mean shaft diameter.

2.21 A normally distributed random variable has an unknown mean μ and a known variance $\sigma^2 = 9$. Find the sample size required to construct a 95 percent confidence interval on the mean that has a total length of 1.0.

2.22 The shelf life of a carbonated beverage is of interest. Ten bottles are randomly selected and tested, and the following results are obtained:

Days	
108	138
124	163
124	159
106	134
115	139

(a) We would like to demonstrate that the mean shelf life exceeds 120 days. Set up appropriate hypotheses for investigating this claim.

(b) Test these hypotheses using $\alpha = 0.01$. What are your conclusions?

(c) Find the P-value for the test in part (b).

(d) Construct a 99 percent confidence interval on the mean shelf life.

2.23 Consider the shelf life data in Problem 2.22. Can shelf life be described or modeled adequately by a normal distribution? What effect would the violation of this assumption have on the test procedure you used in solving Problem 2.17?

2.24 The time to repair an electronic instrument is a normally distributed random variable measured in hours. The repair times for 16 such instruments chosen at random are as follows:

Hours			
159	280	101	212
224	379	179	264
222	362	168	250
149	260	485	170

(a) You wish to know if the mean repair time exceeds 225 hours. Set up appropriate hypotheses for investigating this issue.

(b) Test the hypotheses you formulated in part (a). What are your conclusions? Use $\alpha = 0.05$.

(c) Find the P-value for the test.

(d) Construct a 95 percent confidence interval on mean repair time.

2.25 Reconsider the repair time data in Problem 2.24. Can repair time, in your opinion, be adequately modeled by a normal distribution?

2.26 Two machines are used for filling plastic bottles with a net volume of 16.0 ounces. The filling processes can be assumed to be normal, with standard deviations of $\sigma_1 = 0.015$ and $\sigma_2 = 0.018$. The quality engineering department suspects that both machines fill to the same net volume, whether or not this volume is 16.0 ounces. An experiment is performed by taking a random sample from the output of each machine.

Machine 1		**Machine 2**	
16.03	16.01	16.02	16.03
16.04	15.96	15.97	16.04
16.05	15.98	15.96	16.02
16.05	16.02	16.01	16.01
16.02	15.99	15.99	16.00

(a) State the hypotheses that should be tested in this experiment.

(b) Test these hypotheses using $\alpha = 0.05$. What are your conclusions?

(c) Find the P-value for this test.

(d) Find a 95 percent confidence interval on the difference in mean fill volume for the two machines.

2.27 Two types of plastic are suitable for use by an electronic calculator manufacturer. The breaking strength of this plastic is important. It is known that $\sigma_1 = \sigma_2 = 1.0$ psi. From random samples of $n_1 = 10$ and $n_2 = 12$, we obtain $\bar{y}_1 = 162.5$ and $\bar{y}_2 = 155.0$. The company will not adopt plastic 1 unless its breaking strength exceeds that of plastic 2 by at least 10 psi. Based on the sample information, should they use plastic 1? In answering this question, set up and test appropriate hypotheses using $\alpha = 0.01$. Construct a 99 percent confidence interval on the true mean difference in breaking strength.

2.28 The following are the burning times (in minutes) of chemical flares of two different formulations. The design engineers are interested in both the mean and variance of the burning times.

Type 1		Type 2	
65	82	64	56
81	67	71	69
57	59	83	74
66	75	59	82
82	70	65	79

(a) Test the hypothesis that the two variances are equal. Use $\alpha = 0.05$.

(b) Using the results of (a), test the hypothesis that the mean burning times are equal. Use $\alpha = 0.05$. What is the *P*-value for this test?

(c) Discuss the role of the normality assumption in this problem. Check the assumption of normality for both types of flares.

2.29 An article in *Solid State Technology*, "Orthogonal Design for Process Optimization and Its Application to Plasma Etching" by G. Z. Yin and D. W. Jillie (May 1987) describes an experiment to determine the effect of the C_2F_6 flow rate on the uniformity of the etch on a silicon wafer used in integrated circuit manufacturing. All of the runs were made in random order. Data for two flow rates are as follows:

C_2F_6 Flow	Uniformity Observation					
(SCCM)	1	2	3	4	5	6
125	2.7	4.6	2.6	3.0	3.2	3.8
200	4.6	3.4	2.9	3.5	4.1	5.1

(a) Does the C_2F_6 flow rate affect average etch uniformity? Use $\alpha = 0.05$.

(b) What is the *P*-value for the test in part (a)?

(c) Does the C_2F_6 flow rate affect the wafer-to-wafer variability in etch uniformity? Use $\alpha = 0.05$.

(d) Draw box plots to assist in the interpretation of the data from this experiment.

2.30 A new filtering device is installed in a chemical unit. Before its installation, a random sample yielded the following information about the percentage of impurity: $\bar{y}_1 = 12.5$, $S_1^2 = 101.17$, and $n_1 = 8$. After installation, a random sample yielded $\bar{y}_2 = 10.2$, $S_2^2 = 94.73, n_2 = 9$.

(a) Can you conclude that the two variances are equal? Use $\alpha = 0.05$.

(b) Has the filtering device reduced the percentage of impurity significantly? Use $\alpha = 0.05$.

2.31 Photoresist is a light-sensitive material applied to semiconductor wafers so that the circuit pattern can be imaged on to the wafer. After application, the coated wafers are baked to remove the solvent in the photoresist mixture and to harden the resist. Here are measurements of photoresist thickness (in kA) for eight wafers baked at two different temperatures. Assume that all of the runs were made in random order.

95°C	100°C
11.176	5.263
7.089	6.748
8.097	7.461
11.739	7.015
11.291	8.133
10.759	7.418
6.467	3.772
8.315	8.963

(a) Is there evidence to support the claim that the higher baking temperature results in wafers with a lower mean photoresist thickness? Use $\alpha = 0.05$.

(b) What is the *P*-value for the test conducted in part (a)?

(c) Find a 95 percent confidence interval on the difference in means. Provide a practical interpretation of this interval.

(d) Draw dot diagrams to assist in interpreting the results from this experiment.

(e) Check the assumption of normality of the photoresist thickness.

(f) Find the power of this test for detecting an actual difference in means of 2.5 kA.

(g) What sample size would be necessary to detect an actual difference in means of 1.5 kA with a power of at least 0.9?

2.32 Front housings for cell phones are manufactured in an injection molding process. The time the part is allowed to cool in the mold before removal is thought to influence the occurrence of a particularly troublesome cosmetic defect, flow lines, in the finished housing. After manufacturing, the housings are inspected visually and assigned a score between 1 and 10 based on their appearance, with 10 corresponding to a perfect part and 1 corresponding to a completely defective part. An experiment was conducted using two cool-down times, 10 and 20 seconds, and 20 housings were evaluated at each level of cool-down time. All 40 observations in this experiment were run in random order. The data are as follows.

10 seconds		20 seconds	
1	3	7	6
2	6	8	9
1	5	5	5
3	3	9	7
5	2	5	4
1	1	8	6
5	6	6	8
2	8	4	5
3	2	6	8
5	3	7	7

(a) Is there evidence to support the claim that the longer cool-down time results in fewer appearance defects? Use $\alpha = 0.05$.

(b) What is the P-value for the test conducted in part (a)?

(c) Find a 95 percent confidence interval on the difference in means. Provide a practical interpretation of this interval.

(d) Draw dot diagrams to assist in interpreting the results from this experiment.

(e) Check the assumption of normality for the data from this experiment.

2.33 Twenty observations on etch uniformity on silicon wafers are taken during a qualification experiment for a plasma etcher. The data are as follows:

5.34	6.65	4.76	5.98	7.25
6.00	7.55	5.54	5.62	6.21
5.97	7.35	5.44	4.39	4.98
5.25	6.35	4.61	6.00	5.32

(a) Construct a 95 percent confidence interval estimate of σ^2.

(b) Test the hypothesis that $\sigma^2 = 1.0$. Use $\alpha = 0.05$. What are your conclusions?

(c) Discuss the normality assumption and its role in this problem.

(d) Check normality by constructing a normal probability plot. What are your conclusions?

2.34 The diameter of a ball bearing was measured by 12 inspectors, each using two different kinds of calipers. The results are as follows:

Inspector	Caliper 1	Caliper 2
1	0.265	0.264
2	0.265	0.265
3	0.266	0.264
4	0.267	0.266
5	0.267	0.267
6	0.265	0.268
7	0.267	0.264
8	0.267	0.265
9	0.265	0.265
10	0.268	0.267
11	0.268	0.268
12	0.265	0.269

(a) Is there a significant difference between the means of the population of measurements from which the two samples were selected? Use $\alpha = 0.05$.

(b) Find the P-value for the test in part (a).

(c) Construct a 95 percent confidence interval on the difference in mean diameter measurements for the two types of calipers.

2.35 An article in the journal *Neurology* (1998, Vol. 50, pp. 1246–1252) observed that monozygotic twins share numerous physical, psychological, and pathological traits. The investigators measured an intelligence score of 10 pairs of twins. The data obtained are as follows:

Pair	Birth Order: 1	Birth Order: 2
1	6.08	5.73
2	6.22	5.80
3	7.99	8.42
4	7.44	6.84
5	6.48	6.43
6	7.99	8.76
7	6.32	6.32
8	7.60	7.62
9	6.03	6.59
10	7.52	7.67

(a) Is the assumption that the difference in score is normally distributed reasonable?

(b) Find a 95% confidence interval on the difference in mean score. Is there any evidence that mean score depends on birth order?

(c) Test an appropriate set of hypotheses indicating that the mean score does not depend on birth order.

2.36 An article in the *Journal of Strain Analysis* (Vol. 18, no. 2, 1983) compares several procedures for predicting the shear strength for steel plate girders. Data for nine girders in the form of the ratio of predicted to observed load for two of these procedures, the Karlsruhe and Lehigh methods, are as follows:

Girder	Karlsruhe Method	Lehigh Method
S1/1	1.186	1.061
S2/1	1.151	0.992
S3/1	1.322	1.063
S4/1	1.339	1.062
S5/1	1.200	1.065
S2/1	1.402	1.178
S2/2	1.365	1.037
S2/3	1.537	1.086
S2/4	1.559	1.052

(a) Is there any evidence to support a claim that there is a difference in mean performance between the two methods? Use $\alpha = 0.05$.

(b) What is the P-value for the test in part (a)?

(c) Construct a 95 percent confidence interval for the difference in mean predicted to observed load.

(d) Investigate the normality assumption for both samples.

(e) Investigate the normality assumption for the difference in ratios for the two methods.

(f) Discuss the role of the normality assumption in the paired t-test.

2.37 The deflection temperature under load for two different formulations of ABS plastic pipe is being studied. Two samples of 12 observations each are prepared using each formulation and the deflection temperatures (in °F) are reported below:

Formulation 1			Formulation 2		
206	193	192	177	176	198
188	207	210	197	185	188
205	185	194	206	200	189
187	189	178	201	197	203

(a) Construct normal probability plots for both samples. Do these plots support assumptions of normality and equal variance for both samples?

(b) Do the data support the claim that the mean deflection temperature under load for formulation 1 exceeds that of formulation 2? Use $\alpha = 0.05$.

(c) What is the P-value for the test in part (a)?

2.38 Refer to the data in Problem 2.37. Do the data support a claim that the mean deflection temperature under load for formulation 1 exceeds that of formulation 2 by at least 3°F?

2.39 In semiconductor manufacturing, wet chemical etching is often used to remove silicon from the backs of wafers prior to metalization. The etch rate is an important characteristic of this process. Two different etching solutions are being evaluated. Eight randomly selected wafers have been etched in each solution, and the observed etch rates (in mils/min) are as follows.

Solution 1		Solution 2	
9.9	10.6	10.2	10.6
9.4	10.3	10.0	10.2
10.0	9.3	10.7	10.4
10.3	9.8	10.5	10.3

(a) Do the data indicate that the claim that both solutions have the same mean etch rate is valid? Use $\alpha = 0.05$ and assume equal variances.

(b) Find a 95 percent confidence interval on the difference in mean etch rates.

(c) Use normal probability plots to investigate the adequacy of the assumptions of normality and equal variances.

2.40 Two popular pain medications are being compared on the basis of the speed of absorption by the body. Specifically, tablet 1 is claimed to be absorbed twice as fast as tablet 2. Assume that σ_1^2 and σ_2^2 are known. Develop a test statistic for

$$H_0{:}2\mu_1 = \mu_2$$
$$H_1{:}2\mu_1 \neq \mu_2$$

2.41 **Continuation of Problem 2.40.** An article in *Nature* (1972, pp. 225–226) reported on the levels of monoamine oxidase in blood platelets for a sample of 43 schizophrenic patients resulting in $\bar{y}_1 = 2.69$ and $s_1 = 2.30$ while for a sample of 45 normal patients the results were $\bar{y}_2 = 6.35$ and $s_2 = 4.03$. The units are nm/mg protein/h. Use the results of the previous problem to test the claim that the mean monoamine oxidase level for normal patients is at least twice the mean level for schizophrenic patients. Assume that the sample sizes are large enough to use the sample standard deviations as the true parameter values.

2.42 Suppose we are testing

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

where $\sigma_1^2 > \sigma_2^2$ are known. Our sampling resources are constrained such that $n_1 + n_2 = N$. Show that an allocation of the observations n_1 and n_2 to the two samples leads to the most powerful test is in the ratio $n_1/n_2 = \sigma_1/\sigma_2$.

2.43 **Continuation of Problem 2.42.** Suppose that we want to construct a 95% two-sided confidence interval on the difference in two means where the two sample standard deviations are known to be $\sigma_1 = 4$ and $\sigma_2 = 8$. The total sample size is restricted to $N = 30$. What is the length of the 95% CI if the sample sizes used by the experimenter are $n_1 = n_2 = 15$? How much shorter would the 95% CI have been if the experimenter had used an optimal sample size allocation?

2.44 Develop Equation 2.46 for a $100(1 - \alpha)$ percent confidence interval for the variance of a normal distribution.

2.45 Develop Equation 2.50 for a $100(1 - \alpha)$ percent confidence interval for the ratio σ_1^2/σ_2^2, where σ_1^2 and σ_1^2 are the variances of two normal distributions.

2.46 Develop an equation for finding a $100(1 - \alpha)$ percent confidence interval on the difference in the means of two normal distributions where $\sigma_1^2 \neq \sigma_2^2$. Apply your equation to the Portland cement experiment data, and find a 95 percent confidence interval.

2.47 Construct a data set for which the paired t-test statistic is very large, but for which the usual two-sample or pooled t-test statistic is small. In general, describe how you created the data. Does this give you any insight regarding how the paired t-test works?

2.48 Consider the experiment described in Problem 2.28. If the mean burning times of the two flares differ by as much as 2 minutes, find the power of the test. What sample size would be required to detect an actual difference in mean burning time of 1 minute with a power of at least 0.90?

2.49 Reconsider the bottle filling experiment described in Problem 2.26. Rework this problem assuming that the two population variances are unknown but equal.

2.50 Consider the data from Problem 2.26. If the mean fill volume of the two machines differ by as much as 0.25 ounces, what is the power of the test used in Problem 2.21? What sample size would result in a power of at least 0.9 if the actual difference in mean fill volume is 0.25 ounces?

2.51 An experiment has been performed with a factor that has only two levels. Samples of size $n_1 = n_2 = 12$ have been taken, and the resulting sample data is as follows:

$$\bar{y}_1 = 12.5,\ \bar{y}_2 = 13.1,\ S_1 = 1.8,\ S_2 = 2.1.$$

Can you conclude that there is no difference in means using $\alpha = 0.05$? What are bounds on the P-value for this test? Find a 95 percent confidence interval on the difference in the two means. Does the confidence interval provide any information that is useful in interpreting the test of the hypothesis on the difference in the two means?

2.52 Reconsider the situation in Problem 2.51. Suppose that the two sample sizes were $n_1 = n_2 = 5$. What difference in conclusions (if any) would you have obtained from the hypothesis test? From the confidence interval?

2.53 Suppose that you are testing the hypothesis $H_0 : \mu = 50$ against the usual two-sided alternative. The data are normally distributed with known standard deviation $\sigma = 1$. The sample average obtained in the experiment is 50.5, and it is known that if the true population mean is actually 50.5, then this has no practical significance on the problem that motivated the experiment. Find the P-value for the t-test for the following sample sizes:

(a) $n = 5$ **(b)** $n = 10$ **(c)** $n = 25$

(d) $n = 50$ **(e)** $n = 100$ **(f)** $n = 1000$

Discuss your findings. What does this tell you about relying on P-values in hypothesis testing situations when sample sizes are large?

2.54 Consider the situation in Problem 2.53. Calculate the 95 percent confidence interval on the mean for each of the sample sizes given. How does the length of the confidence interval change with sample size?

2.55 Is the assumption of sampling from a normal distribution critical in the application of the t-test? Justify your answer.

2.56 Why is the random sampling assumption important in statistical inference? Suppose that you had to select a random sample of 100 items from a production line. How would you propose to do this? Should you take into account factors such as the production rate, or whether the line operates continuously or only intermittently?

2.57 An experiment has been performed with a factor that has only two levels. Samples of size $n_1 = n_2 = 10$ have been taken, and the resulting sample data is as follows:

$$\bar{y}_1 = 10.7,\ \bar{y}_2 = 15.1,\ S_1 = 1.5,\ S_2 = 4.1.$$

It seems likely that the two population variances are not the same. Can you conclude that there is no difference in means using $\alpha = 0.05$? What are bounds on the P-value for this test? Find a 95 percent confidence interval on the difference in the two means. Does the confidence interval provide any information that is useful in interpreting the test of the hypothesis on the difference in the two means?

2.58 Do you think that using a significance level of $\alpha = 0.05$ is appropriate for all experiments? In the early stages of research and development work, is there a lot of harm in identifying a factor as important when it really isn't? Would that seem to justify higher levels of significance such as $\alpha = 0.10$ or perhaps even $\alpha = 0.15$ in some situations?

2.59 Power calculation for hypothesis testing are relatively easy to do with modern statistical software. What do you think "adequate power" should be for an experiment? What issues need to be considered in answering this question?

2.60 In the early stages of research and development experimentation, which type of error do you think is most important, type I or type II? Justify your answer.

CHAPTER 3

Experiments with a Single Factor: The Analysis of Variance

CHAPTER OUTLINE

The supplemental material is on the textbook website www.wiley.com/college/montgomery.

CHAPTER LEARNING OBJECTIVES

1. Understand how to set up and run a completely randomized experiment.
2. Understand how to perform a single-factor analysis of variance for a completely randomized design.

3. Know the assumptions underlying the ANOVA and how to check for departures from these assumptions.
4. Know how to apply methods for post-ANOVA comparisons for individual differences between means.
5. Know how to interpret computer output from some standard statistics packages.
6. Understand several approaches for determining appropriate sample sizes in designed experiments.

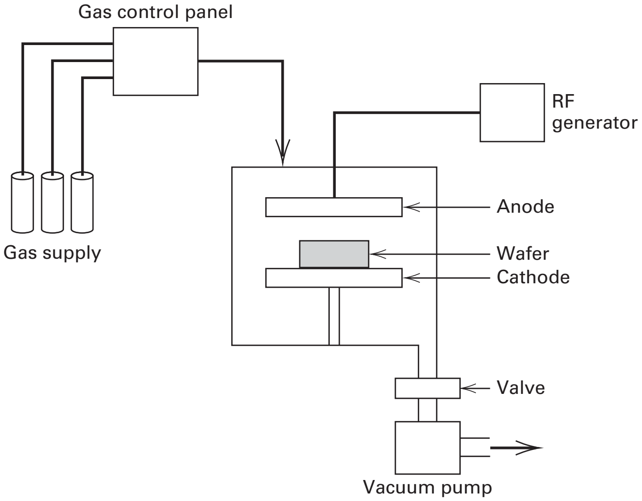
In Chapter 2, we discussed methods for comparing two conditions or treatments. For example, the Portland cement tension bond experiment involved two different mortar formulations. Another way to describe this experiment is as a single-factor experiment with two levels of the factor, where the factor is mortar formulation and the two levels are the two different formulation methods. Many experiments of this type involve more than two levels of the factor. This chapter focuses on methods for the design and analysis of single-factor experiments with an arbitrary number *a* levels of the factor (or *a* treatments). We will assume that the experiment has been completely randomized.

3.1 An Example

In many integrated circuit manufacturing steps, wafers are completely coated with a layer of material such as silicon dioxide or a metal. The unwanted material is then selectively removed by etching through a mask, thereby creating circuit patterns, electrical interconnects, and areas in which diffusions or metal depositions are to be made. A plasma etching process is widely used for this operation, particularly in small geometry applications. Figure 3.1 shows the important features of a typical single-wafer etching tool. Energy is supplied by a radio-frequency (RF) generator causing plasma to be generated in the gap between the electrodes. The chemical species in the plasma are determined by the particular gases used. Fluorocarbons, such as CF_4 (tetrafluoromethane) or C_2F_6 (hexafluoroethane), are often used in plasma etching, but other gases and mixtures of gases are relatively common, depending on the application.

An engineer is interested in investigating the relationship between the RF power setting and the etch rate for this tool. The objective of an experiment like this is to model the relationship between etch rate and RF power and to specify the power setting that will give a desired target etch rate. She is interested in a particular gas (C_2F_6) and gap (0.80 cm) and wants to test four levels of RF power: 160, 180, 200, and 220 W. She decided to test five wafers at each level of RF power.

This is an example of a single-factor experiment with $a = 4$ **levels** of the factor and $n = 5$ **replicates**. The 20 runs should be made in random order. A very efficient way to generate the run order is to enter the 20 runs in a spreadsheet (Excel), generate a column of random numbers using the RAND () function, and then sort by that column.



■ **FIGURE 3.1** **A single-wafer plasma etching tool**

Suppose that the test sequence obtained from this process is given as below:

Test Sequence	Excel Random Number (Sorted)	Power
1	12417	200
2	18369	220
3	21238	220
4	24621	160
5	29337	160
6	32318	180
7	36481	200
8	40062	160
9	43289	180
10	49271	200
11	49813	220
12	52286	220
13	57102	160
14	63548	160
15	67710	220
16	71834	180
17	77216	180
18	84675	180
19	89323	200
20	94037	200

This randomized test sequence is necessary to prevent the effects of unknown nuisance variables, perhaps varying out of control during the experiment, from contaminating the results. To illustrate this, suppose that we were to run the 20 test wafers in the original nonrandomized order (that is, all five 160 W power runs are made first, all five 180 W power runs are made next, and so on). If the etching tool exhibits a warm-up effect such that the longer it is on, the lower the observed etch rate readings will be, the warm-up effect will potentially contaminate the data and destroy the validity of the experiment.

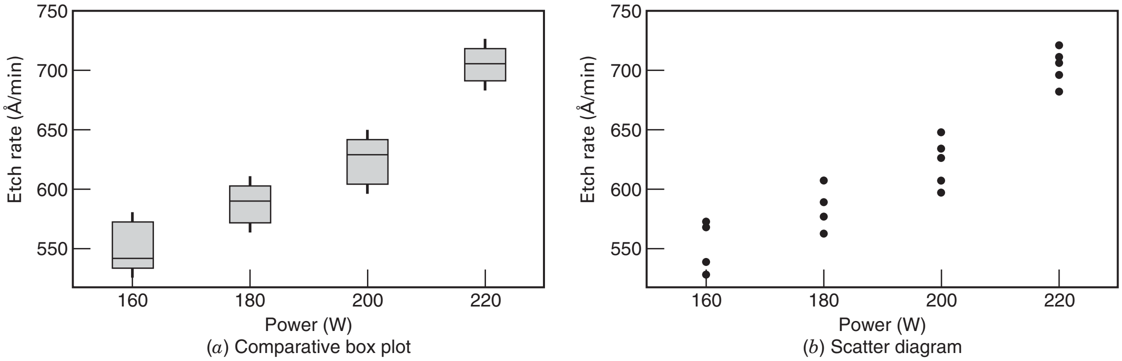
Suppose that the engineer runs the experiment that we have designed in the indicated random order. The observations that she obtains on etch rate are shown in Table 3.1.

It is always a good idea to examine experimental data **graphically**. Figure 3.2*a* presents **box plots** for etch rate at each level of RF power and Figure 3.2*b* presents a **scatter diagram** of etch rate versus RF power. Both graphs indicate that etch rate increases as the power setting increases. There is no strong evidence to suggest that the variability in etch rate around the average depends on the power setting. On the basis of this simple graphical analysis, we strongly suspect that (1) RF power setting affects the etch rate and (2) higher power settings result in increased etch rate.

Suppose that we wish to be more **objective** in our analysis of the data. Specifically, suppose that we wish to test for differences between the mean etch rates at all $a = 4$ levels of RF power. Thus, we are interested in testing the equality of all four means. It might seem that this problem could be solved by performing a t-test for all six possible pairs of means. However, this is not the best solution to this problem. First of all, performing all six pairwise t-tests is inefficient. It takes a lot of effort. Second, conducting all these pairwise comparisons inflates the type I error. Suppose that all four means are equal, so if we select $\alpha = 0.05$, the probability of reaching the correct decision on any single comparison is 0.95. However, the probability of reaching the correct conclusion on all six comparisons is considerably less than 0.95, so the type I error is inflated.

■ **TABLE 3.1**
Etch Rate Data (in Å/min) from the Plasma Etching Experiment

Power	Observations						
(W)	1	2	3	4	5	Totals	Averages
160	575	542	530	539	570	2756	551.2
180	565	593	590	579	610	2937	587.4
200	600	651	610	637	629	3127	625.4
220	725	700	715	685	710	3535	707.0



(*a*) Comparative box plot

(*b*) Scatter diagram

■ **FIGURE 3.2** **Box plots and scatter diagram of the etch rate data**

The appropriate procedure for testing the equality of several means is the **analysis of variance**. However, the analysis of variance has a much wider application than the problem above. It is probably the most useful technique in the field of statistical inference.

3.2 The Analysis of Variance

Suppose we have a **treatments** or different **levels** of a **single factor** that we wish to compare. The observed response from each of the a treatments is a random variable. The data would appear as in Table 3.2. An entry in Table 3.2 (e.g., y_{ij}) represents the jth observation taken under factor level or treatment i. There will be, in general, n observations under the ith treatment. Notice that Table 3.2 is the general case of the data from the plasma etching experiment in Table 3.1.

Models for the Data. We will find it useful to describe the observations from an experiment with a **model.** One way to write this model is

$$y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases} \tag{3.1}$$

where y_{ij} is the ijth observation, μ_i is the mean of the ith factor level or treatment, and ϵ_{ij} is a **random error** component that incorporates all other sources of variability in the experiment including measurement, variability arising from uncontrolled factors, differences between the experimental units (such as test material) to which the treatments are applied, and the general background noise in the process (such as variability over time, effects of environmental variables). It is convenient to think of the errors as having mean zero, so that $E(y_{ij}) = \mu_i$.

Equation 3.1 is called the **means model.** An alternative way to write a model for the data is to define

$$\mu_i = \mu + \tau_i, \qquad i = 1, 2, \ldots, a$$

■ **TABLE 3.2**
Typical Data for a Single-Factor Experiment

Treatment						
(Level)	**Observations**				**Totals**	**Averages**
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
			$\cdots$			
a	y_{a1}	y_{a2}	$\cdots$	y_{an}	$\underline{y_{a.}}$	$\underline{\bar{y}_{a.}}$
					$y_{..}$	$y_{..}$

so that Equation 3.1 becomes

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases} \tag{3.2}$$

In this form of the model, μ is a parameter common to all treatments called the **overall mean**, and τ_i is a parameter unique to the ith treatment called the ***i*th treatment effect.** Equation 3.2 is usually called the **effects model.**

Both the means model and the effects model are **linear statistical models**; that is, the response variable y_{ij} is a linear function of the model parameters. Although both forms of the model are useful, the effects model is more widely encountered in the experimental design literature. It has some intuitive appeal in that μ is a constant and the treatment effects τ_i represent deviations from this constant when the specific treatments are applied.

Equation 3.2 (or 3.1) is also called the **one-way** or **single-factor analysis of variance** (**ANOVA**) model because only one factor is investigated. Furthermore, we will require that the experiment be performed in random order so that the environment in which the treatments are applied (often called the **experimental units**) is as uniform as possible. Thus, the experimental design is a **completely randomized design.** Our objectives will be to test appropriate hypotheses about the treatment means and to estimate them. For hypothesis testing, the model errors are assumed to be normally and independently distributed random variables with mean zero and variance σ^2. The variance σ^2 is assumed to be constant for all levels of the factor. This implies that the observations

$$y_{ij} \sim N(\mu + \tau_i, \sigma^2)$$

and that the observations are mutually independent.

Fixed or Random Factor? The statistical model, Equation 3.2, describes two different situations with respect to the treatment effects. First, the a treatments could have been specifically chosen by the experimenter. In this situation, we wish to test hypotheses about the treatment means, and our conclusions will apply only to the factor levels considered in the analysis. The conclusions cannot be extended to similar treatments that were not explicitly considered. We may also wish to estimate the model parameters (μ, τ_i, σ^2). This is called the **fixed effects model.** Alternatively, the a treatments could be a **random sample** from a larger population of treatments. In this situation, we should like to be able to extend the conclusions (which are based on the sample of treatments) to all treatments in the population, whether or not they were explicitly considered in the analysis. Here, the τ_i are random variables, and knowledge about the particular ones investigated is relatively useless. Instead, we test hypotheses about the variability of the τ_i and try to estimate this variability. This is called the **random effects model** or **components of variance model.** We discuss the single-factor random effects model in Section 3.9. However, we will defer a more complete discussion of experiments with random factors to Chapter 13.

3.3 Analysis of the Fixed Effects Model

In this section, we develop the single-factor analysis of variance for the fixed effects model. Recall that $y_{i.}$ represents the total of the observations under the ith treatment. Let $\bar{y}_{i.}$ represent the average of the observations under the ith treatment. Similarly, let $y_{..}$ represent the grand total of all the observations and $\bar{y}_{..}$ represent the grand average of all the observations. Expressed symbolically,

$$\begin{aligned} y_{i.} &= \sum_{j=1}^{n} y_{ij} \qquad \bar{y}_{i.} = y_{i.}/n \qquad i = 1, 2, \ldots, a \\ y_{..} &= \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij} \qquad \bar{y}_{..} = y_{..}/N \end{aligned} \tag{3.3}$$

where $N = an$ is the total number of observations. We see that the "dot" subscript notation implies summation over the subscript that it replaces.

We are interested in testing the equality of the a treatment means; that is, $E(y_{ij}) = \mu + \tau_i = \mu_i, i = 1, 2, \ldots, a$. The appropriate hypotheses are

$$\begin{aligned} H_0&: \mu_1 = \mu_2 = \cdots = \mu_a \\ H_1&: \mu_i \neq \mu_j \qquad \text{for at least one pair } (i, j) \end{aligned} \tag{3.4}$$

In the effects model, we break the ith treatment mean μ_i into two components such that $\mu_i = \mu + \tau_i$. We usually think of μ as an overall mean so that

$$\frac{\sum_{i=1}^{a} \mu_i}{a} = \mu$$

This definition implies that

$$\sum_{i=1}^{a} \tau_i = 0$$

That is, the treatment or factor effects can be thought of as deviations from the overall mean.[1] Consequently, an equivalent way to write the above hypotheses is in terms of the treatment effects τ_i, say

$$\begin{aligned} H_0&: \tau_1 = \tau_2 = \cdots \tau_a = 0 \\ H_1&: \tau_i \neq 0 \qquad \text{for at least one } i \end{aligned}$$

Thus, we speak of testing the equality of treatment means or testing that the treatment effects (the τ_i) are zero. The appropriate procedure for testing the equality of a treatment means is the analysis of variance.

3.3.1 Decomposition of the Total Sum of Squares

The name **analysis of variance** is derived from a partitioning of total variability into its component parts. The total corrected sum of squares

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2$$

is used as a measure of overall variability in the data. Intuitively, this is reasonable because if we were to divide SS_T by the appropriate number of degrees of freedom (in this case, $an - 1 = N - 1$), we would have the **sample variance** of the y's. The sample variance is, of course, a standard measure of variability.

[1] For more information on this subject, refer to the supplemental text material for Chapter 3.

Note that the total corrected sum of squares SS_T may be written as

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2=\sum_{i=1}^{a}\sum_{j=1}^{n}[(\bar{y}_{i.}-\bar{y}_{..})+(y_{ij}-\bar{y}_{i.})]^2 \tag{3.5}$$

or

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2=n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2+\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2 \\ +2\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{y}_{i.}-\bar{y}_{..})(y_{ij}-\bar{y}_{i.})$$

However, the cross-product term in this last equation is zero, because

$$\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})=y_{i.}-n\bar{y}_{i.}=y_{i.}-n(y_{i.}/n)=0$$

Therefore, we have

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2=n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2+\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2 \tag{3.6}$$

Equation 3.6 is the fundamental ANOVA identity. It states that the total variability in the data, as measured by the total corrected sum of squares, can be partitioned into a sum of squares of the differences **between** the treatment averages and the grand average plus a sum of squares of the differences of observations **within** treatments from the treatment average. Now, the difference between the observed treatment averages and the grand average is a measure of the differences between treatment means, whereas the differences of observations within a treatment from the treatment average can be due to only random error. Thus, we may write Equation 3.6 symbolically as

$$SS_T=SS_{\text{Treatments}}+SS_E$$

where $SS_{\text{Treatments}}$ is called the sum of squares due to treatments (i.e., between treatments) and SS_E is called the sum of squares due to error (i.e., within treatments). There are $an=N$ total observations; thus, SS_T has $N-1$ degrees of freedom. There are a levels of the factor (and a treatment means), so $SS_{\text{Treatments}}$ has $a-1$ degrees of freedom. Finally, there are n replicates within any treatment providing $n-1$ degrees of freedom with which to estimate the experimental error. Because there are a treatments, we have $a(n-1)=an-a=N-a$ degrees of freedom for error.

It is instructive to examine explicitly the two terms on the right-hand side of the fundamental ANOVA identity. Consider the error sum of squares

$$SS_E=\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2=\sum_{i=1}^{a}\left[\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2\right]$$

In this form, it is easy to see that the term within square brackets, if divided by $n-1$, is the sample variance in the ith treatment, or

$$S_i^2=\frac{\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2}{n-1}\qquad i=1,2,\ldots,a$$

Now a sample variances may be combined to give a single estimate of the common population variance as follows:

$$\frac{(n-1)S_1^2+(n-1)S_2^2+\cdots+(n-1)S_a^2}{(n-1)+(n-1)+\cdots+(n-1)}=\frac{\sum_{i=1}^{a}\left[\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2\right]}{\sum_{i=1}^{a}(n-1)}$$
$$=\frac{SS_E}{(N-a)}$$

Thus, $SS_E/(N-a)$ is a **pooled estimate** of the common variance within each of the a treatments.

Similarly, if there were no differences between the a treatment means, we could use the variation of the treatment averages from the grand average to estimate σ^2. Specifically,

$$\frac{SS_{\text{Treatments}}}{a-1}=\frac{n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2}{a-1}$$

is an estimate of σ^2 if the treatment means are equal. The reason for this may be intuitively seen as follows: The quantity $\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2/(a-1)$ estimates σ^2/n, the variance of the treatment averages, so $n\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2/(a-1)$ must estimate σ^2 if there are no differences in treatment means.

We see that the ANOVA identity (Equation 3.6) provides us with two estimates of σ^2—one based on the inherent variability within treatments and the other based on the variability between treatments. If there are no differences in the treatment means, these two estimates should be very similar, and if they are not, we suspect that the observed difference must be caused by differences in the treatment means. Although we have used an intuitive argument to develop this result, a somewhat more formal approach can be taken.

The quantities

$$MS_{\text{Treatments}}=\frac{SS_{\text{Treatments}}}{a-1}$$

and

$$MS_E=\frac{SS_E}{N-a}$$

are called **mean squares.** We now examine the **expected values** of these mean squares. Consider

$$E(MS_E)=E\left(\frac{SS_E}{N-a}\right)=\frac{1}{N-a}E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2\right]$$
$$=\frac{1}{N-a}E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}^2-2y_{ij}\bar{y}_{i.}+\bar{y}_{i.}^2)\right]$$
$$=\frac{1}{N-a}E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}^2-2n\sum_{i=1}^{a}\bar{y}_{i.}^2+n\sum_{i=1}^{a}\bar{y}_{i.}^2\right]$$
$$=\frac{1}{N-a}E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}^2-\frac{1}{n}\sum_{i=1}^{a}\bar{y}_{i.}^2\right]$$

Substituting the model (Equation 3.1) into this equation, we obtain

$$E(MS_E)=\frac{1}{N-a}E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(\mu+\tau_i+\epsilon_{ij})^2-\frac{1}{n}\sum_{i=1}^{a}\left(\sum_{i=1}^{n}\mu+\tau_i+\epsilon_{ij}\right)^2\right]$$

Now when squaring and taking expectation of the quantity within the brackets, we see that terms involving ϵ_{ij}^2 and $\epsilon_{i.}^2$ are replaced by σ^2 and $n\sigma^2$, respectively, because $E(\epsilon_{ij}) = 0$. Furthermore, all cross products involving ϵ_{ij} have zero expectation. Therefore, after squaring and taking expectation, the last equation becomes

$$E(MS_E) = \frac{1}{N-a}\left[N\mu^2 + n\sum_{i=1}^{a}\tau_i^2 + N\sigma^2 - N\mu^2 - n\sum_{i=1}^{a}\tau_i^2 - a\sigma^2\right]$$

or

$$E(MS_E) = \sigma^2$$

By a similar approach, we may also show that[2]

$$E(MS_{\text{Treatments}}) = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}$$

Thus, as we argued heuristically, $MS_E = SS_E/(N-a)$ estimates σ^2, and, if there are no differences in treatment means (which implies that $\tau_i = 0$), $MS_{\text{Treatments}} = SS_{\text{Treatments}}/(a-1)$ also estimates σ^2. However, note that if treatment means do differ, the expected value of the treatment mean square is greater than σ^2.

It seems clear that a test of the hypothesis of no difference in treatment means can be performed by comparing $MS_{\text{Treatments}}$ and MS_E. We now consider how this comparison may be made.

3.3.2 Statistical Analysis

We now investigate how a formal test of the hypothesis of no differences in treatment means ($H_0\colon \mu_1 = \mu_2 = \cdots = \mu_a$, or equivalently, $H_0\colon \tau_1 = \tau_2 = \cdots = \tau_a = 0$) can be performed. Because we have assumed that the errors ϵ_{ij} are normally and independently distributed with mean zero and variance σ^2, the observations y_{ij} are normally and independently distributed with mean $\mu + \tau_i$ and variance σ^2. Thus, SS_T is a sum of squares in normally distributed random variables; consequently, it can be shown that SS_T/σ^2 is distributed as chi-square with $N-1$ degrees of freedom. Furthermore, we can show that SS_E/σ^2 is chi-square with $N-a$ degrees of freedom and that $SS_{\text{Treatments}}/\sigma^2$ is chi-square with $a-1$ degrees of freedom if the null hypothesis $H_0\colon \tau_i = 0$ is true. However, all three sums of squares are not necessarily independent because $SS_{\text{Treatments}}$ and SS_E add to SS_T. The following theorem, which is a special form of one attributed to William G. Cochran, is useful in establishing the independence of SS_E and $SS_{\text{Treatments}}$.

THEOREM 3-1
Cochran's Theorem

Let Z_i be NID(0, 1) for $i = 1, 2, \ldots, v$ and

$$\sum_{i=1}^{v} Z_i^2 = Q_1 + Q_2 + \cdots + Q_s$$

where $s \le v$, and Q_i has v_i degrees of freedom ($i = 1, 2, \ldots, s$). Then $Q_1, Q_2, \ldots, Q_s$ are independent chi-square random variables with $v_1, v_2, \ldots, v_s$ degrees of freedom, respectively, if and only if

$$v = v_1 + v_2 + \cdots + v_s$$

[2] Refer to the supplemental text material for Chapter 3.

Because the degrees of freedom for $SS_{\text{Treatments}}$ and SS_E add to $N - 1$, the total number of degrees of freedom, Cochran's theorem implies that $SS_{\text{Treatments}}/\sigma^2$ and SS_E/σ^2 are independently distributed chi-square random variables. Therefore, if the null hypothesis of no difference in treatment means is true, the ratio

$$F_0 = \frac{SS_{\text{Treatments}}/(a-1)}{SS_E/(N-a)} = \frac{MS_{\text{Treatments}}}{MS_E} \tag{3.7}$$

is distributed as F with $a - 1$ and $N - a$ degrees of freedom. Equation 3.7 is the **test statistic** for the hypothesis of no differences in treatment means.

From the expected mean squares we see that, in general, MS_E is an unbiased estimator of σ^2. Also, under the null hypothesis, $MS_{\text{Treatments}}$ is an unbiased estimator of σ^2. However, if the null hypothesis is false, the expected value of $MS_{\text{Treatments}}$ is greater than σ^2. Therefore, under the alternative hypothesis, the expected value of the numerator of the test statistic (Equation 3.7) is greater than the expected value of the denominator, and we should reject H_0 on values of the test statistic that are too large. This implies an upper-tail, one-tail critical region. Therefore, we should reject H_0 and conclude that there are differences in the treatment means if

$$F_0 > F_{\alpha,a-1,N-a}$$

where F_0 is computed from Equation 3.7. Alternatively, we could use the P-value approach for decision making. The table of F percentages in the Appendix (Table IV) can be used to find bounds on the P-value.

The sums of squares may be computed in several ways. One direct approach is to make use of the definition

$$y_{ij} - \bar{y}_{..} = (\bar{y}_{.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})$$

Use a spreadsheet to compute these three terms for each observation. Then, sum up the squares to obtain SS_T, $SS_{\text{Treatments}}$, and SS_E. Another approach is to rewrite and simplify the definitions of $SS_{\text{Treatments}}$ and SS_T in Equation 3.6, which results in

$$SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{N} \tag{3.8}$$

$$SS_{\text{Treatments}} = \frac{1}{n} \sum_{i=1}^{a} y_{i.}^2 - \frac{y_{..}^2}{N} \tag{3.9}$$

and

$$SS_E = SS_T - SS_{\text{Treatments}} \tag{3.10}$$

This approach is nice because some calculators are designed to accumulate the sum of entered numbers in one register and the sum of the squares of those numbers in another, so each number only has to be entered once. In practice, we use computer software to do this.

The test procedure is summarized in Table 3.3. This is called an **analysis of variance** (or **ANOVA**) table.

■ TABLE 3.3
The Analysis of Variance Table for the Single-Factor, Fixed Effects Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Between treatments	$SS_{\text{Treatments}} = n \sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2$	$a - 1$	$MS_{\text{Treatments}}$	$F_0 = \frac{MS_{\text{Treatments}}}{MS_E}$
Error (within treatments)	$SS_E = SS_T - SS_{\text{Treatments}}$	$N - a$	MS_E	
Total	$SS_{\text{T}} = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2$	$N - 1$		

EXAMPLE 3.1 The Plasma Etching Experiment

To illustrate the analysis of variance, return to the first example discussed in Section 3.1. Recall that the engineer is interested in determining if the RF power setting affects the etch rate, and she has run a completely randomized experiment with four levels of RF power and five replicates. For convenience, we repeat here the data from Table 3.1:

RF Power	Observed Etch Rate (Å/min)					Totals	Averages
(W)	1	2	3	4	5	$y_{i.}$	$\bar{y}_{i.}$
160	575	542	530	539	570	2756	551.2
180	565	593	590	579	610	2937	587.4
200	600	651	610	637	629	3127	625.4
220	725	700	715	685	710	3535	707.0
						$y_{i.} = 12{,}355$	$\bar{y}_{..} = 617.75$

$$SS_T = \sum_{i=1}^{4}\sum_{j=1}^{5} y_{ij}^2 - \frac{y_{..}^2}{N}$$

$$= (575)^2 + (542)^2 + \cdots + (710)^2 - \frac{(12{,}355)^2}{20}$$

$$= 72{,}209.75$$

$$SS_{\text{Treatments}} = \frac{1}{n}\sum_{i=1}^{4} y_{i.}^2 - \frac{y_{..}^2}{N}$$

$$= \frac{1}{5}[(2756)^2 + \cdots + (3535)^2] - \frac{(12{,}355)^2}{20}$$

$$= 66{,}870.55$$

We will use the analysis of variance to test $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ against the alternative H_1 : some means are different. The sums of squares required are computed using Equations 3.8, 3.9, and 3.10 as follows:

$$SS_E = SS_T - SS_{\text{Treatments}}$$

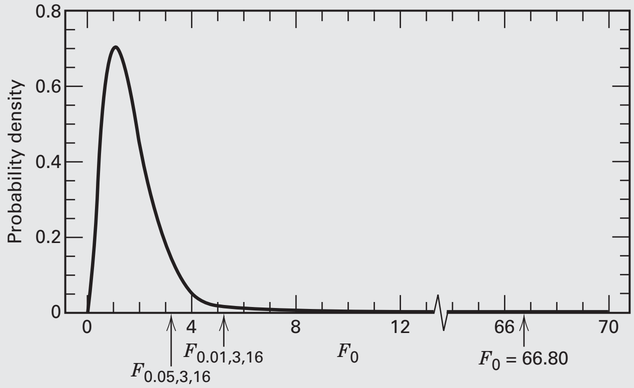
$$= 72{,}209.75 - 66{,}870.55 = 5339.20$$

Usually, these calculations would be performed on a computer, using a software package with the capability to analyze data from designed experiments.

The ANOVA is summarized in Table 3.4. Note that the RF power or between-treatment mean square (22,290.18) is many times larger than the within-treatment or error mean square (333.70). This indicates that it is unlikely that the treatment means are equal. More formally, we can compute the F ratio $F_0 = 22{,}290.18/333.70 = 66.80$ and compare this to an appropriate upper-tail percentage point of the $F_{3,16}$ distribution. To use a fixed significance level approach, suppose that the experimenter has selected $\alpha = 0.05$. From Appendix Table IV we find that $F_{0.05,3,16} = 3.24$. Because $F_0 = 66.80 > 3.24$, we reject H_0 and conclude that the treatment means differ; that is, the RF power setting significantly affects the mean etch rate. We could also compute a P-value for this test statistic. Figure 3.3 shows the reference distribution ($F_{3,16}$) for the test statistic F_0. Clearly, the P-value is very small in this case. From Appendix Table A-4, we find that $F_{0.01,3,16} = 5.29$ and because $F_0 > 5.29$, we can conclude that an upper bound for the P-value is 0.01; that is, $P < 0.01$ (the exact P-value is $P = 2.88 \times 10^{-9}$).

■ **TABLE 3.4**
ANOVA for the Plasma Etching Experiment

Source of Variation	Sum of Square	Degrees of Freedom	Mean Squares	F_0	P-Value
RF Power	66,870.55	3	22,290.18	$F_0 = 66.80$	< 0.01
Error	5339.20	16	333.70		
Total	72,209.75	19			



■ **FIGURE 3.3** **The reference distribution ($F_{3,16}$) for the test statistic F_0 in Example 3.1**

Coding the Data. Generally, we need not be too concerned with computing because there are many widely available computer programs for performing the calculations. These computer programs are also helpful in performing many other analyses associated with experimental design (such as residual analysis and model adequacy checking). In many cases, these programs will also assist the experimenter in setting up the design.

However, when hand calculations are necessary, it is sometimes helpful to code the observations. This is illustrated in Example 3.2.

EXAMPLE 3.2 Coding the Observations

The ANOVA calculations may often be made more easily or accurately by **coding** the observations. For example, consider the plasma etching data in Example 3.1. Suppose we subtract 600 from each observation. The coded data are shown in Table 3.5. It is easy to verify that

$$SS_T = (-25)^2 + (-58)^2 + \cdots + (110)^2 - \frac{(355)^2}{20} = 72{,}209.75$$

$$SS_{\text{Treatment}} = \frac{(-244)^2 + (-63)^2 + (127)^2 + (535)^2}{5} - \frac{(355)^2}{20} = 66{,}870.55$$

and

$$SS_E = 5339.20$$

Comparing these sums of squares to those obtained in Example 3.1, we see that subtracting a constant from the original data does not change the sums of squares.

Now suppose that we multiply each observation in Example 3.1 by 2. It is easy to verify that the sums of squares for the transformed data are $SS_T = 288{,}839.00$, $SS_{\text{Treatments}} = 267{,}482.20$, and $SS_E = 21{,}356.80$. These sums of squares appear to differ considerably from those obtained in Example 3.1. However, if they are divided by 4 (i.e., 2^2), the results are identical. For example, for the treatment sum of squares $267{,}482.20/4 = 66{,}870.55$. Also, for the coded data, the F ratio is $F = (267{,}482.20/3)/(21{,}356.80/16) = 66.80$, which is identical to the F ratio for the original data. Thus, the ANOVAs are equivalent.

■ **TABLE 3.5**
Coded Etch Rate Data for Example 3.2

RF Power	Observations					Totals
(W)	1	2	3	4	5	$y_{i.}$
160	−25	−58	−70	−61	−30	−244
180	−35	−7	−10	−21	10	−63
200	0	51	10	37	29	127
220	125	100	115	85	110	535

Randomization Tests and Analysis of Variance. In our development of the ANOVA F-test, we have used the assumption that the random errors ϵ_{ij} are normally and independently distributed random variables. The F-test can also be justified as an approximation to a **randomization test.** To illustrate this, suppose that we have five observations on each of two treatments and that we wish to test the equality of treatment means. The data would look like this:

Treatment 1	Treatment 2
y_{11}	y_{21}
y_{12}	y_{22}
y_{13}	y_{23}
y_{14}	y_{24}
y_{15}	y_{25}

We could use the ANOVA F-test to test $H_0: \mu_1 = \mu_2$. Alternatively, we could use a somewhat different approach. Suppose we consider all the possible ways of allocating the 10 numbers in the above sample to the two treatments. There are $10!/5!5! = 252$ possible arrangements of the 10 observations. If there is no difference in treatment means, all 252 arrangements are equally likely. For each of the 252 arrangements, we calculate the value of the F-statistic using Equation 3.7. The distribution of these F values is called a **randomization distribution**, and a large value of F indicates that the data are not consistent with the hypothesis $H_0: \mu_1 = \mu_2$. For example, if the value of F actually observed was exceeded by only five of the values of the randomization distribution, this would correspond to rejection of $H_0: \mu_1 = \mu_2$ at a significance level of $\alpha = 5/252 = 0.0198$ (or 1.98 percent). Notice that no normality assumption is required in this approach.

The difficulty with this approach is that, even for relatively small problems, it is computationally prohibitive to enumerate the exact randomization distribution. However, numerous studies have shown that the exact randomization distribution is well approximated by the usual normal-theory F distribution. Thus, even without the normality assumption, the ANOVA F-test can be viewed as an approximation to the randomization test. For further reading on randomization tests in the analysis of variance, see Box, Hunter, and Hunter (2005).

3.3.3 Estimation of the Model Parameters

We now present estimators for the parameters in the single-factor model

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}.$$

and confidence intervals on the treatment means. We will prove later that reasonable estimates of the overall mean and the treatment effects are given by

$$\begin{aligned} \hat{\mu} &= \bar{y}_{..} \\ \hat{\tau}_i &= \bar{y}_{i.} - \bar{y}_{..}, \qquad i = 1, 2, \ldots, a \end{aligned} \tag{3.11}$$

These estimators have considerable intuitive appeal; note that the overall mean is estimated by the grand average of the observations and that any treatment effect is just the difference between the treatment average and the grand average.

A **confidence interval** estimate of the *i*th treatment mean may be easily determined. The mean of the *i*th treatment is

$$\mu_i = \mu + \tau_i$$

A point estimator of μ_i would be $\hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \bar{y}_i$. Now, if we assume that the errors are normally distributed, each treatment average $\bar{y}_{i.}$ is distributed NID$(\mu_i, \sigma^2/n)$. Thus, if σ^2 were known, we could use the normal distribution to define the confidence interval. Using the MS_E as an estimator of σ^2, we would base the confidence interval on the t distribution. Therefore, a $100(1 - \alpha)$ percent confidence interval on the *i*th treatment mean μ_i is

$$\bar{y}_{i.} - t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}} \le \mu_i \le \bar{y}_{i.} + t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}} \tag{3.12}$$

Differences in treatments are frequently of great practical interest. A $100(1 - \alpha)$ percent confidence interval on the difference in any two treatment means, say $\mu_i - \mu_j$, would be

$$\bar{y}_{i.} - \bar{y}_{j.} - t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}} \le \mu_i - \mu_j \le \bar{y}_{i.} - \bar{y}_{j.} + t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}} \tag{3.13}$$

EXAMPLE 3.3

Using the data in Example 3.1, we may find the estimates of the overall mean and the treatment effects as $\hat{\mu} = 12{,}355/20 = 617.75$ and

$$\begin{aligned} \hat{\tau}_1 &= \bar{y}_{1.} - \bar{y}_{..} = 551.20 - 617.75 = -66.55 \\ \hat{\tau}_2 &= \bar{y}_{2.} - \bar{y}_{..} = 587.40 - 617.75 = -30.35 \\ \hat{\tau}_3 &= \bar{y}_{3.} - \bar{y}_{..} = 625.40 - 617.75 = 7.65 \\ \hat{\tau}_4 &= \bar{y}_{4.} - \bar{y}_{..} = 707.00 - 617.75 = 89.25 \end{aligned}$$

A 95 percent confidence interval on the mean of treatment 4 (220 W of RF power) is computed from Equation 3.12 as

$$707.00 - 2.120\sqrt{\frac{333.70}{5}} \le \mu_4 \le 707.00 + 2.120\sqrt{\frac{333.70}{5}}$$

or

$$707.00 - 17.32 \le \mu_4 \le 707.00 + 17.32$$

Thus, the desired 95 percent confidence interval is $689.68 \le \mu_4 \le 724.32$.

Simultaneous Confidence Intervals. The confidence interval expressions given in Equations 3.12 and 3.13 are **one-at-a-time** confidence intervals. That is, the confidence level $1 - \alpha$ applies to only one particular estimate. However, in many problems, the experimenter may wish to calculate several confidence intervals, one for each of a number of means or differences between means. If there are r such $100(1 - \alpha)$ percent confidence intervals of interest, the probability that the r intervals will **simultaneously** be correct is at least $1 - r\alpha$. The probability $r\alpha$ is often called the **experimentwise error rate** or overall confidence coefficient. The number of intervals r does not have to be large before the set of confidence intervals becomes relatively uninformative. For example, if there are $r = 5$ intervals and $\alpha = 0.05$ (a typical choice), the simultaneous confidence level for the set of five confidence intervals is at least 0.75, and if $r = 10$ and $\alpha = 0.05$, the simultaneous confidence level is at least 0.50.

One approach to ensuring that the simultaneous confidence level is not too small is to replace $\alpha/2$ in the one-at-a-time confidence interval Equations 3.12 and 3.13 with $\alpha/(2r)$. This is called the **Bonferroni method**, and it allows the experimenter to construct a set of r simultaneous confidence intervals on treatment means or differences in treatment means for which the overall confidence level is at least $100(1-\alpha)$ percent. When r is not too large, this is a very nice method that leads to reasonably short confidence intervals. For more information, refer to the **supplemental text material** for Chapter 3.

3.3.4 Unbalanced Data

In some single-factor experiments, the number of observations taken within each treatment may be different. We then say that the design is **unbalanced.** The analysis of variance described may still be used, but slight modifications must be made in the sum of squares formulas. Let n_i observations be taken under treatment i $(i = 1, 2, \ldots, a)$ and $N = \sum_{i=1}^{a} n_i$. The manual computational formulas for SS_T and $SS_{\text{Treatments}}$ become

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N} \tag{3.14}$$

and

$$SS_{\text{Treatments}} = \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N} \tag{3.15}$$

No other changes are required in the analysis of variance.

There are two advantages in choosing a balanced design. First, the test statistic is relatively insensitive to small departures from the assumption of equal variances for the a treatments if the sample sizes are equal. This is not the case for unequal sample sizes. Second, the power of the test is maximized if the samples are of equal size.

3.4 Model Adequacy Checking

The decomposition of the variability in the observations through an analysis of variance identity (Equation 3.6) is a purely algebraic relationship. However, the use of the partitioning to test formally for no differences in treatment means requires that certain assumptions be satisfied. Specifically, these assumptions are that the observations are adequately described by the model

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

and that the errors are normally and independently distributed with mean zero and constant but unknown variance σ^2. If these assumptions are valid, the analysis of variance procedure is an exact test of the hypothesis of no difference in treatment means.

In practice, however, these assumptions will usually not hold exactly. Consequently, it is usually unwise to rely on the analysis of variance until the validity of these assumptions has been checked. Violations of the basic assumptions and model adequacy can be easily investigated by the examination of **residuals.** We define the residual for observation j in treatment i as

$$e_{ij} = y_{ij} - \hat{y}_{ij} \tag{3.16}$$

where $\hat{y}_{ij}$ is an estimate of the corresponding observation y_{ij} obtained as follows:

$$\begin{aligned}\hat{y}_{ij} &= \hat{\mu} + \hat{\tau}_i\\ &= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..})\\ &= \bar{y}_{i.}\end{aligned} \tag{3.17}$$

Equation 3.17 gives the intuitively appealing result that the estimate of any observation in the *i*th treatment is just the corresponding treatment average.

Examination of the residuals should be an automatic part of any analysis of variance. If the model is adequate, the residuals should be **structureless**; that is, they should contain no obvious patterns. Through analysis of residuals, many types of model inadequacies and violations of the underlying assumptions can be discovered. In this section, we show how model diagnostic checking can be done easily by graphical analysis of residuals and how to deal with several commonly occurring abnormalities.

3.4.1 The Normality Assumption

A check of the normality assumption could be made by plotting a histogram of the residuals. If the NID$(0, \sigma^2)$ assumption on the errors is satisfied, this plot should look like a sample from a normal distribution centered at zero. Unfortunately, with small samples, considerable fluctuation in the shape of a histogram often occurs, so the appearance of a moderate departure from normality does not necessarily imply a serious violation of the assumptions. Gross deviations from normality are potentially serious and require further analysis.

An extremely useful procedure is to construct a **normal probability plot** of the residuals. Recall from Chapter 2 that we used a normal probability plot of the raw data to check the assumption of normality when using the *t*-test. In the analysis of variance, it is usually more effective (and straightforward) to do this with the **residuals.** If the underlying error distribution is normal, this plot will resemble a straight line. In visualizing the straight line, place more emphasis on the central values of the plot than on the extremes.

Table 3.6 shows the original data and the residuals for the etch rate data in Example 3.1. The normal probability plot is shown in Figure 3.4. The general impression from examining this display is that the error distribution is approximately normal. The tendency of the normal probability plot to bend down slightly on the left side and upward slightly on the right side implies that the tails of the error distribution are somewhat *thinner* than would be anticipated in a normal distribution; that is, the largest residuals are not quite as large (in absolute value) as expected. This plot is not grossly nonnormal, however.

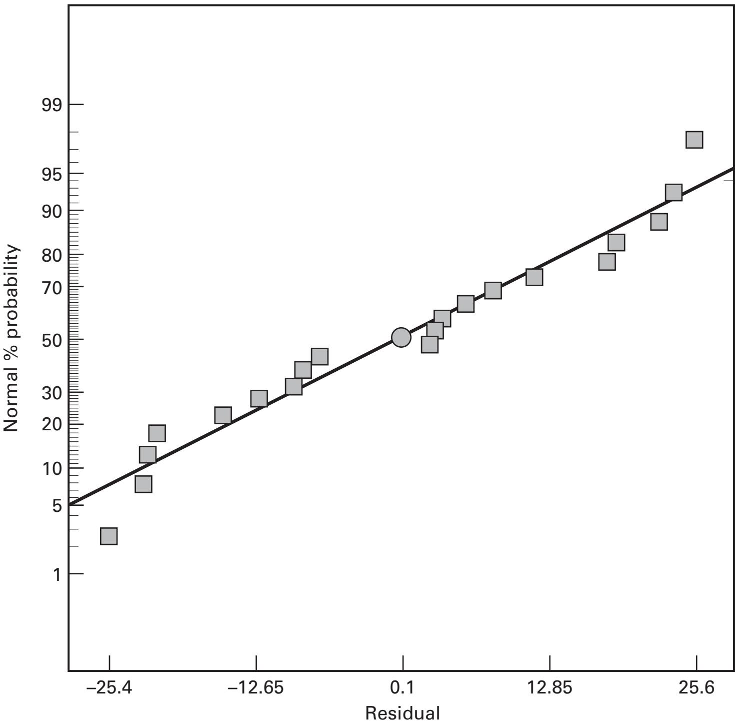
In general, moderate departures from normality are of little concern in the fixed effects analysis of variance (recall our discussion of randomization tests in Section 3.3.2). An error distribution that has considerably thicker or thinner tails than the normal is of more concern than a skewed distribution. Because the F-test is only slightly affected, we say that the analysis of variance (and related procedures such as multiple comparisons) is **robust** to the normality assumption. Departures from normality usually cause both the true significance level and the power to differ slightly from the advertised values, with the power generally being lower. The random effects model that we will discuss in Section 3.9 and Chapter 13 is more severely affected by nonnormality.

■ **TABLE 3.6**
Etch Rate Data and Residuals from Example 3.1[a]

	Observations (*j*)					
Power (w)	**1**	**2**	**3**	**4**	**5**	$\hat{y}_{ij} = \bar{y}_{i.}$
160	23.8	–9.2	–21.2	–12.2	18.8	551.2
	575 (13)	542 (14)	530 (8)	539 (5)	570 (4)	
180	–22.4	5.6	2.6	–8.4	22.6	587.4
	565 (18)	593 (9)	590 (6)	579 (16)	610 (17)	
200	–25.4	25.6	–15.4	11.6	3.6	625.4
	600 (7)	651 (19)	610 (10)	637 (20)	629 (1)	
220	18.0	–7.0	8.0	–22.0	3.0	707.0
	725 (2)	700 (3)	715 (15)	685 (11)	710 (12)	

[a] The residuals are shown in the box in each cell. The numbers in parentheses indicate the order in which each experimental run was made.

■ **FIGURE 3.4 Normal probability plot of residuals for Example 3.1**



A very common defect that often shows up on normal probability plots is one residual that is very much larger than any of the others. Such a residual is often called an **outlier.** The presence of one or more outliers can seriously distort the analysis of variance, so when a potential outlier is located, careful investigation is called for. Frequently, the cause of the outlier is a mistake in calculations or a data coding or copying error. If this is not the cause, the experimental circumstances surrounding this run must be carefully studied. If the outlying response is a particularly desirable value (high strength, low cost, etc.), the outlier may be more informative than the rest of the data. We should be careful not to reject or discard an outlying observation unless we have reasonably nonstatistical grounds for doing so. At worst, you may end up with two analyses: one with the outlier and one without.

Several formal statistical procedures may be used for detecting outliers [e.g., see Stefansky (1972), John and Prescott (1975), and Barnett and Lewis (1994)]. Some statistical software packages report the results of a statistical test for normality (such as the Anderson–Darling test) on the normal probability plot of residuals. This should be viewed with caution as those tests usually assume that the data to which they are applied are independent and residuals are not independent.

A rough check for outliers may be made by examining the **standardized residuals**

$$d_{ij} = \frac{e_{ij}}{\sqrt{MS_E}} \tag{3.18}$$

If the errors ϵ_{ij} are $N(0, \sigma^2)$, the standardized residuals should be approximately normal with mean zero and unit variance. Thus, about 68 percent of the standardized residuals should fall within the limits ± 1, about 95 percent of them should fall within ± 2, and virtually all of them should fall within ± 3. A residual bigger than 3 or 4 standard deviations from zero is a potential outlier.

For the tensile strength data of Example 3.1, the normal probability plot gives no indication of outliers. Furthermore, the largest standardized residual is

$$d_1 = \frac{e_1}{\sqrt{MS_E}} = \frac{25.6}{\sqrt{333.70}} = \frac{25.6}{18.27} = 1.40$$

which should cause no concern.

3.4.2 Plot of Residuals in Time Sequence

Plotting the residuals in time order of data collection is helpful in detecting strong **correlation** between the residuals. A tendency to have runs of positive and negative residuals indicates positive correlation. This would imply that the **independence assumption** on the errors has been violated. This is a potentially serious problem and one that is difficult to correct, so it is important to prevent the problem if possible when the data are collected. Proper randomization of the experiment is an important step in obtaining independence.

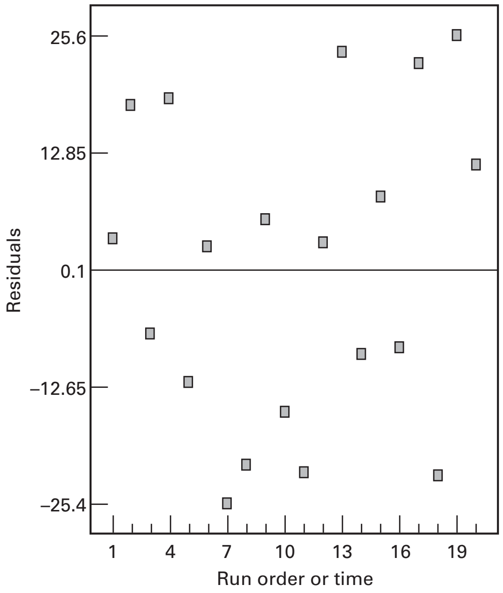
Sometimes the skill of the experimenter (or the subjects) may change as the experiment progresses, or the process being studied may "drift" or become more erratic. This will often result in a change in the error variance over time. This condition often leads to a plot of residuals versus time that exhibits more spread at one end than at the other. Nonconstant variance is a potentially serious problem. We will have more to say on the subject in Sections 3.4.3 and 3.4.4.

Table 3.6 displays the residuals and the time sequence of data collection for the tensile strength data. A plot of these residuals versus run order or time is shown in Figure 3.5. There is no reason to suspect any violation of the independence or constant variance assumptions.

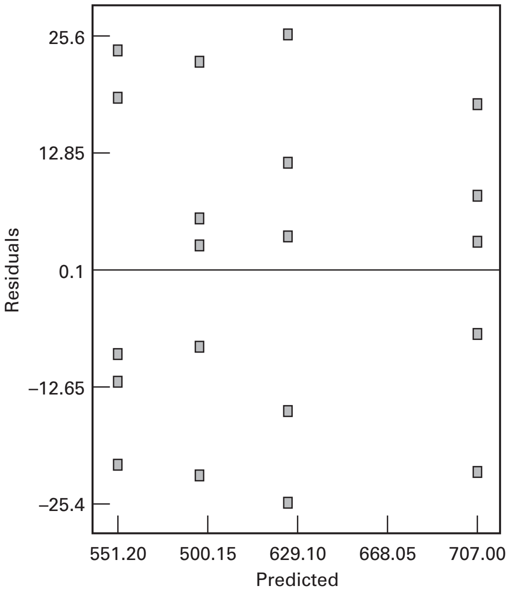
3.4.3 Plot of Residuals Versus Fitted Values

If the model is correct and the assumptions are satisfied, the residuals should be structureless; in particular, they should be unrelated to any other variable including the predicted response. A simple check is to plot the residuals versus the fitted values $\hat{y}_{ij}$. (For the single-factor experiment model, remember that $\hat{y}_{ij} = \bar{y}_{i.}$, the ith treatment average.) This plot should not reveal any obvious pattern. Figure 3.6 plots the residuals versus the fitted values for the tensile strength data of Example 3.1. No unusual structure is apparent.

A defect that occasionally shows up on this plot is **nonconstant variance.** Sometimes the variance of the observations increases as the magnitude of the observation increases. This would be the case if the error or background noise in the experiment was a constant percentage of the size of the observation. (This commonly happens with many measuring instruments—error is a percentage of the scale reading.) If this were the case, the residuals would get larger as y_{ij} gets larger, and the plot of residuals versus $\hat{y}_{ij}$ would look like an outward-opening funnel or megaphone. Nonconstant variance also arises in cases where the data follow a nonnormal, skewed distribution because in skewed distributions the variance tends to be a function of the mean.



■ **FIGURE 3.5 Plot of residuals versus run order or time**



■ **FIGURE 3.6 Plot of residuals versus fitted values**

If the assumption of homogeneity of variances is violated, the F-test is only slightly affected in the balanced (equal sample sizes in all treatments) fixed effects model. However, in unbalanced designs or in cases where one variance is very much larger than the others, the problem is more serious. Specifically, if the factor levels having the larger variances also have the smaller sample sizes, the actual type I error rate is larger than anticipated (or confidence intervals have lower actual confidence levels than were specified). Conversely, if the factor levels with larger variances also have the larger sample sizes, the significance levels are smaller than anticipated (confidence levels are higher). This is a good reason for choosing **equal sample sizes** whenever possible. For random effects models, unequal error variances can significantly disturb inferences on variance components even if balanced designs are used.

Inequality of variance also shows up occasionally on the plot of residuals versus run order. An outward-opening funnel pattern indicates that variability is increasing over time. This could result from operator/subject fatigue, accumulated stress on equipment, changes in material properties such as catalyst degradation, or tool wear, or any of a number of causes.

The usual approach to dealing with nonconstant variance when it occurs for the aforementioned reasons is to apply a **variance-stabilizing transformation** and then to run the analysis of variance on the transformed data. In this approach, one should note that the conclusions of the analysis of variance apply to the *transformed* populations.

Considerable research has been devoted to the selection of an appropriate transformation. If experimenters know the theoretical distribution of the observations, they may utilize this information in choosing a transformation. For example, if the observations follow the Poisson distribution, the **square root transformation** $y_{ij}^* = \sqrt{y_{ij}}$ or $y_{ij}^* = \sqrt{1 + y_{ij}}$ would be used. If the data follow the lognormal distribution, the **logarithmic transformation** $y_{ij}^* = \log y_{ij}$ is appropriate. For binomial data expressed as fractions, the **arcsin transformation** $y_{ij}^* = \arcsin \sqrt{y_{ij}}$ is useful. When there is no obvious transformation, the experimenter usually *empirically* seeks a transformation that equalizes the variance regardless of the value of the mean. We offer some guidance on this at the conclusion of this section. In factorial experiments, which we introduce in Chapter 5, another approach is to select a transformation that minimizes the interaction mean square, resulting in an experiment that is easier to interpret. In Chapter 15, we discuss methods for analytically selecting the form of the transformation in more detail. Transformations made for inequality of variance also affect the form of the error distribution. In most cases, the transformation brings the error distribution closer to normal. For more discussion of transformations, refer to Bartlett (1947), Dolby (1963), Box and Cox (1964), and Draper and Hunter (1969).

Statistical Tests for Equality of Variance. Although residual plots are frequently used to diagnose inequality of variance, several statistical tests have also been proposed. These tests may be viewed as formal tests of the hypotheses

$$\begin{aligned} &H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_a^2 \\ &H_1 : \text{above not true for at least one } \sigma_i^2 \end{aligned}$$

A widely used procedure is **Bartlett's test.** The procedure involves computing a statistic whose sampling distribution is closely approximated by the chi-square distribution with $a - 1$ degrees of freedom when the a random samples are from independent normal populations. The test statistic is

$$\chi_0^2 = 2.3026 \frac{q}{c} \tag{3.19}$$

where

$$q = (N - a) \log_{10} S_p^2 - \sum_{i=1}^{a} (n_i - 1) \log_{10} S_i^2$$

$$c = 1 + \frac{1}{3(a-1)} \left(\sum_{i=1}^{a} (n_i - 1)^{-1} - (N - a)^{-1} \right)$$

$$S_p^2 = \frac{\sum_{i=1}^{a} (n_i - 1) S_i^2}{N - a}$$

and S_i^2 is the sample variance of the ith population.

The quantity q is large when the sample variances S_i^2 differ greatly and is equal to zero when all S_i^2 are equal. Therefore, we should reject H_0 on values of χ_0^2 that are too large; that is, we reject H_0 only when

$$\chi_0^2 > \chi_{\alpha,a-1}^2$$

where $\chi_{\alpha,a-1}^2$ is the upper α percentage point of the chi-square distribution with $a - 1$ degrees of freedom. The P-value approach to decision making could also be used.

Bartlett's test is very sensitive to the normality assumption. Consequently, when the validity of this assumption is doubtful, Bartlett's test should not be used.

EXAMPLE 3.4

In the plasma etch experiment, the normality assumption is not in question, so we can apply Bartlett's test to the etch rate data. We first compute the sample variances in each treatment and find that $S_1^2 = 400.7$, $S_2^2 = 280.3$, $S_3^2 = 421.3$, and $S_4^2 = 232.5$. Then

$$S_p^2 = \frac{4(400.7) + 4(280.3) + 4(421.3) + 4(232.5)}{16} = 333.7$$

$$q = 16\log_{10}(333.7) - 4[\log_{10}400.7 + \log_{10}280.3 + \log_{10}421.3 + \log_{10}232.5] = 0.21$$

$$c = 1 + \frac{1}{3(3)}\left(\frac{4}{4} - \frac{1}{16}\right) = 1.10$$

and the test statistic is

$$\chi_0^2 = 2.3026\frac{(0.21)}{(1.10)} = 0.43$$

From Appendix Table III, we find that $\chi_{0.05,3}^2 = 7.81$ (the P-value is $P = 0.934$), so we cannot reject the null hypothesis. There is no evidence to counter the claim that all five variances are the same. This is the same conclusion reached by analyzing the plot of residuals versus fitted values.

Because Bartlett's test is sensitive to the normality assumption, there may be situations where an alternative procedure would be useful. Anderson and McLean (1974) present a useful discussion of statistical tests for equality of variance. The **modified Levene test** [see Levene (1960) and Conover, Johnson, and Johnson (1981)] is a very nice procedure that is robust to departures from normality. To test the hypothesis of equal variances in all treatments, the modified Levene test uses the absolute deviation of the observations y_{ij} in each treatment from the treatment median, say, $\tilde{y}_i$. Denote these deviations by

$$d_{ij} = |y_{ij} - \tilde{y}_i| \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots n_i \end{cases}$$

The modified Levene test then evaluates whether or not the means of these deviations are equal for all treatments. It turns out that if the mean deviations are equal, the variances of the observations in all treatments will be the same. The test statistic for Levene's test is simply the usual ANOVA F-statistic for testing equality of means applied to the absolute deviations.

EXAMPLE 3.5

A civil engineer is interested in determining whether four different methods of estimating flood flow frequency produce equivalent estimates of peak discharge when applied to the same watershed. Each procedure is used six times on the watershed, and the resulting discharge data (in cubic feet per second) are shown in the upper panel of Table 3.7. The analysis of variance for the data, summarized in Table 3.8, implies that there is a difference in mean peak discharge estimates given by the four procedures. The plot of residuals versus fitted values, shown in Figure 3.7, is disturbing because the outward-opening funnel shape indicates that the constant variance assumption is not satisfied.

We will apply the modified Levene test to the peak discharge data. The upper panel of Table 3.7 contains the treatment medians $\tilde{y}_i$ and the lower panel contains the deviations d_{ij} around the medians. Levene's test consists of conducting a standard analysis of variance on the d_{ij}.

The F-test statistic that results from this is $F_0 = 4.55$, for which the P-value is $P = 0.0137$. Therefore, Levene's test rejects the null hypothesis of equal variances, essentially confirming the diagnosis we made from visual examination of Figure 3.7. The peak discharge data are a good candidate for data transformation.

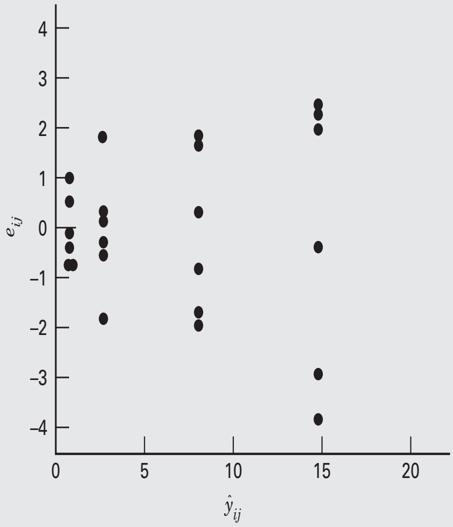
■ **TABLE 3.7**
Peak Discharge Data

Estimation Method	Observations						$\bar{y}_{i.}$	$\tilde{y}_i$	S_i
1	0.34	0.12	1.23	0.70	1.75	0.12	0.71	0.520	0.66
2	0.91	2.94	2.14	2.36	2.86	4.55	2.63	2.610	1.09
3	6.31	8.37	9.75	6.09	9.82	7.24	7.93	7.805	1.66
4	17.15	11.82	10.95	17.20	14.35	16.82	14.72	15.59	2.77

Estimation Method	Deviations d_{ij} for the Modified Levene Test					
1	0.18	0.40	0.71	0.18	1.23	0.40
2	1.70	0.33	0.47	0.25	0.25	1.94
3	1.495	0.565	1.945	1.715	2.015	0.565
4	1.56	3.77	4.64	1.61	1.24	1.23

■ **TABLE 3.8**
Analysis of Variance for Peak Discharge Data

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-Value
Methods	708.3471	3	236.1157	76.07	< 0.001
Error	62.0811	20	3.1041		
Total	770.4282	23			



■ **FIGURE 3.7 Plot of residuals versus $\hat{y}_{ij}$ for Example 3.5**

Empirical Selection of a Transformation. We observed above that if experimenters knew the relationship between the variance of the observations and the mean, they could use this information to guide them in selecting the form of the transformation. We now elaborate on this point and show one method for empirically selecting the form of the required transformation from the data.

Let $E(y) = \mu$ be the mean of y, and suppose that the standard deviation of y is proportional to a power of the mean of y such that

$$\sigma_y \propto \mu^\alpha$$

We want to find a transformation on y that yields a constant variance. Suppose that the transformation is a power of the original data, say

$$y^* = y^\lambda \tag{3.20}$$

Then it can be shown that

$$\sigma_{y*} \propto \mu^{\lambda+\alpha-1} \tag{3.21}$$

Clearly, if we set $\lambda = 1 - \alpha$, the variance of the transformed data y^* is constant.

Several of the common transformations discussed previously are summarized in Table 3.9. Note that $\lambda = 0$ implies the log transformation. These transformations are arranged in order of increasing **strength.** By the strength of a transformation, we mean the amount of curvature it induces. A mild transformation applied to data spanning a narrow range has little effect on the analysis, whereas a strong transformation applied over a large range may have dramatic results. Transformations often have little effect unless the ratio $y_{\max}/y_{\min}$ is larger than 2 or 3.

In many experimental design situations where there is replication, we can empirically estimate α from the data. Because in the ith treatment combination $\sigma_{y_i} \propto \mu_i^\alpha = \theta\mu_i^\alpha$, where θ is a constant of proportionality, we may take logs to obtain

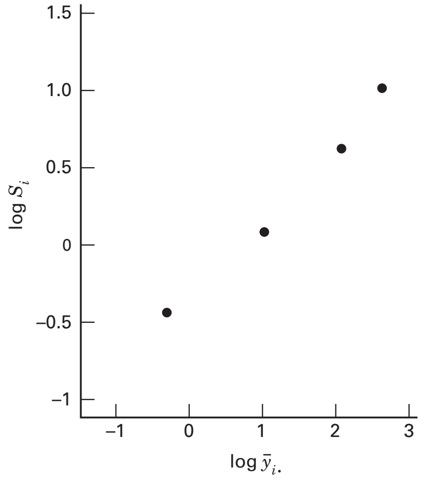
$$\log\ \sigma_{y_i} = \log\ \theta + \alpha\ \log\ \mu_i \tag{3.22}$$

Therefore, a plot of $\log \sigma_{y_i}$ versus $\log\ \mu_i$ would be a straight line with slope α. Because we don't know σ_{y_i} and μ_i, we may substitute reasonable estimates of them in Equation 3.22 and use the slope of the resulting straight line fit as an estimate of α. Typically, we would use the standard deviation S_i and the average $\bar{y}_{i.}$ of the ith treatment (or, more generally, the ith treatment combination or set of experimental conditions) to estimate σ_{y_i} and μ_i.

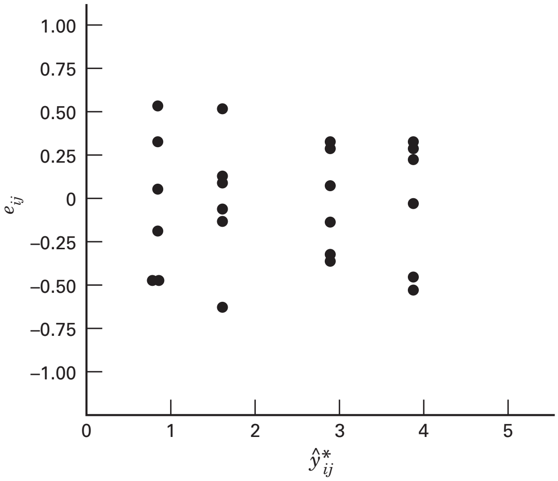
To investigate the possibility of using a variance-stabilizing transformation on the peak discharge data from Example 3.5, we plot $\log S_i$ versus $\log \bar{y}_{i.}$ in Figure 3.8. The slope of a straight line passing through these four points is close to 1/2 and from Table 3.9 this implies that the square root transformation may be appropriate. The analysis of variance for the transformed data $y^* = \sqrt{y}$ is presented in Table 3.10, and a plot of residuals versus the predicted response is shown in Figure 3.9. This residual plot is much improved in comparison to Figure 3.7, so we conclude that the square root transformation has been helpful. Note that in Table 3.10 we have reduced the degrees of freedom for error and total by one to account for the use of the data to estimate the transformation parameter α.

■ **TABLE 3.9**
Variance-Stabilizing Transformations

Relationship Between σ_y and μ	α	$\lambda = 1 - \alpha$	Transformation	Comment
$\sigma_y \propto$ constant	0	1	No transformation	
$\sigma_y \propto \mu^{1/2}$	1/2	1/2	Square root	Poisson (count) data
$\sigma_y \propto \mu$	1	0	Log	
$\sigma_y \propto \mu^{3/2}$	3/2	−1/2	Reciprocal square root	
$\sigma_y \propto \mu^2$	2	−1	Reciprocal	



■ **FIGURE 3.8 Plot of log S_i versus log $\bar{y}_{i.}$ for the peak discharge data from Example 3.5**



■ **FIGURE 3.9 Plot of residuals from transformed data versus $\hat{y}^*_{ij}$ for the peak discharge data in Example 3.5**

■ **TABLE 3.10**
Analysis of Variance for Transformed Peak Discharge Data, $y^* = \sqrt{y}$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	*P*-Value
Methods	32.6842	3	10.8947	76.99	< 0.001
Error	2.6884	19	0.1415		
Total	35.3726	22			

In practice, many experimenters select the form of the transformation by simply trying several alternatives and observing the effect of each transformation on the plot of residuals versus the predicted response. The transformation that produced the most satisfactory residual plot is then selected. Alternatively, there is a formal method called the **Box-Cox Method** for selecting a variance-stability transformation. In Chapter 15 we discuss and illustrate this procedure. It is widely used and implemented in many software packages.

3.4.4 Plots of Residuals Versus Other Variables

If data have been collected on any other variables that might possibly affect the response, the residuals should be plotted against these variables. For example, in the tensile strength experiment of Example 3.1, strength may be significantly affected by the thickness of the fiber, so the residuals should be plotted versus fiber thickness. If different testing machines were used to collect the data, the residuals should be plotted against machines. Patterns in such residual plots imply that the variable affects the response. This suggests that the variable should be either controlled more carefully in future experiments or included in the analysis.

3.5 Practical Interpretation of Results

After conducting the experiment, performing the statistical analysis, and investigating the underlying assumptions, the experimenter is ready to draw practical conclusions about the problem he or she is studying. Often this is relatively easy, and certainly in the simple experiments we have considered so far, this might be done somewhat informally, perhaps by inspection of graphical displays such as the box plots and scatter diagram in Figures 3.1 and 3.2. However, in some cases, more formal techniques need to be applied. We present some of these techniques in this section.

3.5.1 A Regression Model

The factors involved in an experiment can be either **quantitative** or **qualitative.** A quantitative factor is one whose levels can be associated with points on a numerical scale, such as temperature, pressure, or time. Qualitative factors, on the other hand, are factors for which the levels cannot be arranged in order of magnitude. Operators, batches of raw material, and shifts are typical qualitative factors because there is no reason to rank them in any particular numerical order.

Insofar as the initial design and analysis of the experiment are concerned, both types of factors are treated identically. The experimenter is interested in determining the differences, if any, between the levels of the factors. In fact, the analysis of variance treats the design factor as if it were qualitative or categorical. If the factor is really qualitative, such as operators, it is meaningless to consider the response for a subsequent run at an intermediate level of the factor. However, with a quantitative factor such as time, the experimenter is usually interested in the entire range of values used, particularly the response from a subsequent run at an intermediate factor level. That is, if the levels 1.0, 2.0, and 3.0 hours are used in the experiment, we may wish to predict the response at 2.5 hours. Thus, the experimenter is frequently interested in developing an interpolation equation for the response variable in the experiment. This equation is an **empirical model** of the process that has been studied.

The general approach to fitting empirical models is called **regression analysis**, which is discussed extensively in Chapter 10. See also the **supplemental text material** for this chapter. This section briefly illustrates the technique using the etch rate data of Example 3.1.

Figure 3.10 presents scatter diagrams of etch rate y versus the power x for the experiment in Example 3.1. From examining the scatter diagram, it is clear that there is a strong relationship between the etch rate and power. As a first approximation, we could try fitting a **linear model** to the data, say

$$y = \beta_0 + \beta_1 x + \epsilon$$

where β_0 and β_1 are unknown parameters to be estimated and ϵ is a random error term. The method often used to estimate the parameters in a model such as this is the **method of least squares.** This consists of choosing estimates of the β's such that the sum of the squares of the errors (the ϵ's) is minimized. The least squares fit in our example is

$$\hat{y} = 137.62 + 2.527x$$

(If you are unfamiliar with regression methods, see Chapter 10 and the supplemental text material for this chapter.)

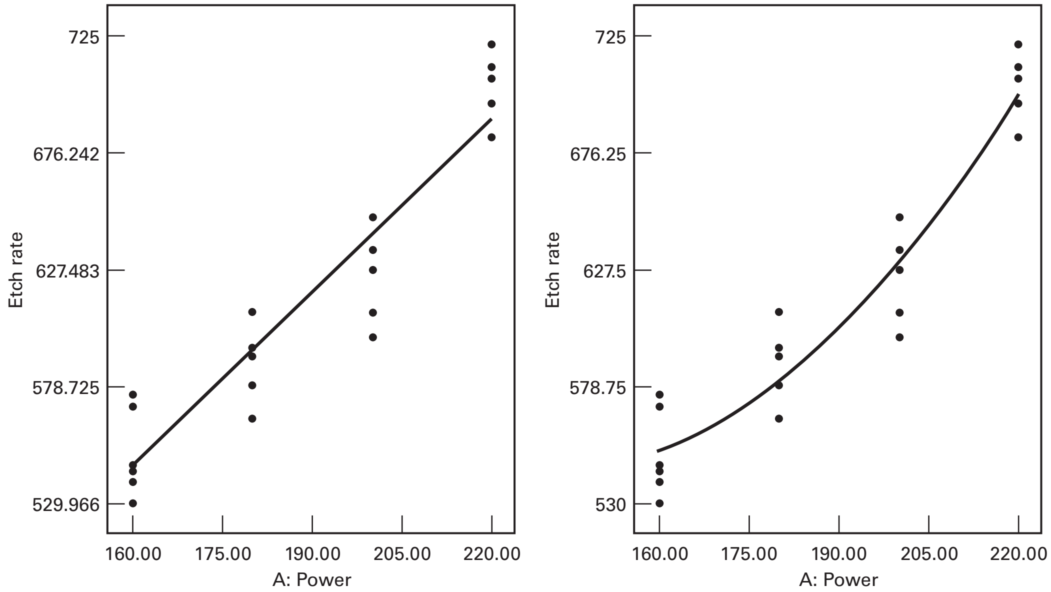
This linear model is shown in Figure 3.10*a*. It does not appear to be very satisfactory at the higher power settings. Perhaps an improvement can be obtained by adding a quadratic term in x. The resulting **quadratic model** fit is

$$\hat{y} = 1147.77 - 8.2555\, x + 0.028375\, x^2$$

This quadratic fit is shown in Figure 3.10*b*. The quadratic model appears to be superior to the linear model because it provides a better fit at the higher power settings.

In general, we would like to fit the lowest order polynomial that adequately describes the system or process. In this example, the quadratic polynomial seems to fit better than the linear model, so the extra complexity of the quadratic model is justified. Selecting the order of the approximating polynomial is not always easy, however, and it is relatively easy to overfit, that is, to add high-order polynomial terms that do not really improve the fit but increase the complexity of the model and often damage its usefulness as a predictor or interpolation equation.

In this example, the empirical model could be used to predict etch rate at power settings within the region of experimentation. In other cases, the empirical model could be used for **process optimization**, that is, finding the levels of the design variables that result in the best values of the response. We will discuss and illustrate these problems extensively later in the book.



(*a*) Linear model

(*b*) Quadratic model

■ **FIGURE 3.10 Scatter diagrams and regression models for the etch rate data of Example 3.1**

3.5.2 Comparisons Among Treatment Means

Suppose that in conducting an analysis of variance for the fixed effects model the null hypothesis is rejected. Thus, there are differences between the treatment means but exactly *which* means differ is not specified. Sometimes in this situation, further comparisons and analysis among **groups** of treatment means may be useful. The ith treatment mean is defined as $\mu_i = \mu + \tau_i$, and μ_i is estimated by $\overline{y}_i$. Comparisons between treatment means are made in terms of either the treatment totals $\{y_{i.}\}$ or the treatment averages $\{\overline{y}_{i.}\}$. The procedures for making these comparisons are usually called **multiple comparison methods.** In the next several sections, we discuss methods for making comparisons among individual treatment means or groups of these means.

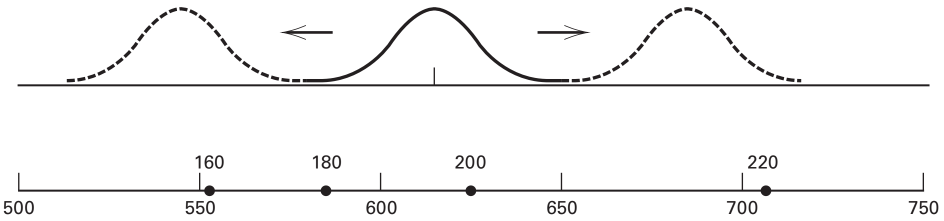
3.5.3 Graphical Comparisons of Means

It is very easy to develop a graphical procedure for the comparison of means following an analysis of variance. Suppose that the factor of interest has a levels and that $\overline{y}_{1.}, \overline{y}_{2.}, \ldots, \overline{y}_{a.}$, are the treatment averages. If we know σ, any treatment average would have a standard deviation $\sigma/\sqrt{n}$. Consequently, if all factor level means are identical, the observed sample means $\overline{y}_{i.}$ would behave as if they were a set of observations drawn at random from a normal distribution with mean $\overline{y}_{..}$ and standard deviation $\sigma/\sqrt{n}$. Visualize a normal distribution capable of being slid along an axis below which the $\overline{y}_{1.}, \overline{y}_{2.}, \ldots, \overline{y}_{a.}$, are plotted. If the treatment means are all equal, there should be some position for this distribution that makes it obvious that the $\overline{y}_{i.}$ values were drawn from the same distribution. If this is not the case, the $\overline{y}_{i.}$ values that appear *not* to have been drawn from this distribution are associated with factor levels that produce different mean responses.

The only flaw in this logic is that σ is unknown. Box, Hunter, and Hunter (2005) point out that we can replace σ with $\sqrt{MS_E}$ from the analysis of variance and use a t distribution with a scale factor $\sqrt{MS_E/n}$ instead of the normal. Such an arrangement for the etch rate data of Example 3.1 is shown in Figure 3.11. Focus on the t distribution shown as a solid line curve in the middle of the display.

To sketch the t distribution in Figure 3.11, simply multiply the abscissa t value by the scale factor

$$\sqrt{MS_E/n} = \sqrt{330.70/5} = 8.13$$



■ **FIGURE 3.11** **Etch rate averages from Example 3.1 in relation to a *t* distribution with scale factor $\sqrt{MS_E/n} = \sqrt{330.70/5} = 8.13$**

and plot this against the ordinate of *t* at that point. Because the *t* distribution looks much like the normal, except that it is a little flatter near the center and has longer tails, this sketch is usually easily constructed by eye. If you wish to be more precise, there is a table of abscissa *t* values and the corresponding ordinates in Box, Hunter, and Hunter (2005). The distribution can have an arbitrary origin, although it is usually best to choose one in the region of the $\bar{y}_{i.}$ values to be compared. In Figure 3.11, the origin is 615 Å/min.

Now visualize sliding the *t* distribution in Figure 3.11 along the horizontal axis as indicated by the dashed lines and examine the four means plotted in the figure. Notice that there is no location for the distribution such that all four averages could be thought of as typical, randomly selected observations from the distribution. This implies that all four means are not equal; thus, the figure is a graphical display of the ANOVA results. Furthermore, the figure indicates that all four levels of power (160, 180, 200, 220 W) produce mean etch rates that differ from each other. In other words, $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$.

This simple procedure is a rough but effective technique for many multiple comparison problems. However, there are more formal methods. We now give a brief discussion of some of these procedures.

3.5.4 Contrasts

Many multiple comparison methods use the idea of a **contrast.** Consider the plasma etching experiment of Example 3.1. Because the null hypothesis was rejected, we know that some power settings produce different etch rates than others, but which ones actually cause this difference? We might suspect at the outset of the experiment that 200 W and 220 W produce the same etch rate, implying that we would like to test the hypothesis

$$H_0 : \mu_3 = \mu_4$$
$$H_1 : \mu_3 \neq \mu_4$$

or equivalently

$$\begin{aligned} H_0 &: \mu_3 - \mu_4 = 0 \\ H_1 &: \mu_3 - \mu_4 \neq 0 \end{aligned} \tag{3.23}$$

If we had suspected at the start of the experiment that the *average* of the lowest levels of power did not differ from the *average* of the highest levels of power, then the hypothesis would have been

$$H_0 : \mu_1 + \mu_2 = \mu_3 + \mu_4$$
$$H_1 : \mu_1 + \mu_2 \neq \mu_3 + \mu_4$$

or

$$\begin{aligned} H_0 &: \mu_1 + \mu_2 - \mu_3 - \mu_4 = 0 \\ H_1 &: \mu_1 + \mu_2 - \mu_3 - \mu_4 \neq 0 \end{aligned} \tag{3.24}$$

In general, a **contrast** is a linear combination of parameters of the form

$$\Gamma = \sum_{i=1}^{a} c_i \mu_i$$

where the **contrast constants** $c_1, c_2, \ldots, c_a$ sum to zero; that is, $\sum_{i=1}^{a} c_i = 0$. Both of the above hypotheses can be expressed in terms of contrasts:

$$\begin{aligned} H_0 &: \sum_{i=1}^{a} c_i \mu_i = 0 \\ H_1 &: \sum_{i=1}^{a} c_i \mu_i \neq 0 \end{aligned} \tag{3.25}$$

The contrast constants for the hypotheses in Equation 3.23 are $c_1 = c_2 = 0$, $c_3 = +1$, and $c_4 = -1$, whereas for the hypotheses in Equation 3.24, they are $c_1 = c_2 = +1$ and $c_3 = c_4 = -1$.

Testing hypotheses involving contrasts can be done in two basic ways. The first method uses a t-test. Write the contrast of interest in terms of the **treatment averages**, giving

$$C = \sum_{i=1}^{a} c_i \bar{y}_{i.}$$

The variance of C is

$$V(C) = \frac{\sigma^2}{n} \sum_{i=1}^{a} c_i^2 \tag{3.26}$$

when the sample sizes in each treatment are equal. If the null hypothesis in Equation 3.25 is true, the ratio

$$\frac{\sum_{i=1}^{a} c_i \bar{y}_{i.}}{\sqrt{\frac{\sigma^2}{n} \sum_{i=1}^{a} c_i^2}}$$

has the $N(0, 1)$ distribution. Now we would replace the unknown variance σ^2 by its estimate, the mean square error MS_E and use the statistic

$$t_0 = \frac{\sum_{i=1}^{a} c_i \bar{y}_{i.}}{\sqrt{\frac{MS_E}{n} \sum_{i=1}^{a} c_i^2}} \tag{3.27}$$

to test the hypotheses in Equation 3.25. The null hypothesis would be rejected if $|t_0|$ in Equation 3.27 exceeds $t_{\alpha/2,N-a}$.

The second approach uses an F-test. Now the square of a t random variable with v degrees of freedom is an F random variable with 1 numerator and v denominator degrees of freedom. Therefore, we can obtain

$$F_0 = t_0^2 = \frac{\left(\sum_{i=1}^{a} c_i \bar{y}_{i.}\right)^2}{\frac{MS_E}{n} \sum_{i=1}^{a} c_i^2} \tag{3.28}$$

as an F-statistic for testing Equation 3.25. The null hypothesis would be rejected if $F_0 > F_{\alpha,1,N-a}$. We can write the test statistic of Equation 3.28 as

$$F_0 = \frac{MS_C}{MS_E} = \frac{SS_C/1}{MS_E}$$

where the single-degree-of-freedom contrast sum of squares is

$$SS_C = \frac{\left(\sum_{i=1}^{a} c_i \bar{y}_{i.}\right)^2}{\frac{1}{n}\sum_{i=1}^{a} c_i^2} \tag{3.29}$$

Confidence Interval for a Contrast. Instead of testing hypotheses about a contrast, it may be more useful to construct a confidence interval. Suppose that the contrast of interest is

$$\Gamma = \sum_{i=1}^{a} c_i \mu_i$$

Replacing the treatment means with the treatment averages yields

$$C = \sum_{i=1}^{a} c_i \bar{y}_{i.}$$

Because

$$E\left(\sum_{i=1}^{a} c_i \bar{y}_{i.}\right) = \sum_{i=1}^{a} c_i \mu_i \qquad \text{and} \qquad V(C) = \sigma^2/n \sum_{i=1}^{a} c_i^2$$

the 100(1 − α) percent confidence interval on the contrast $\Sigma_{i=1}^{a} c_i \mu_i$ is

$$\sum_{i=1}^{a} c_i \bar{y}_{i.} - t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}\sum_{i=1}^{a} c_i^2} \le \sum_{i=1}^{a} c_i \mu_i \le \sum_{i=1}^{a} c_i \bar{y}_{i.} + t_{\alpha/2,N-a}\sqrt{\frac{MS_E}{n}\sum_{i=1}^{a} c_i^2} \tag{3.30}$$

Note that we have used MS_E to estimate σ^2. Clearly, if the confidence interval in Equation 3.30 includes zero, we would be unable to reject the null hypothesis in Equation 3.25.

Standardized Contrast. When more than one contrast is of interest, it is often useful to evaluate them on the same scale. One way to do this is to standardize the contrast so that it has variance σ^2. If the contrast $\Sigma_{i=1}^{a} c_i \mu_i$ is written in terms of treatment averages as $\Sigma_{i=1}^{a} c_i \bar{y}_{i.}$, dividing it by $\sqrt{(1/n)\Sigma_{i=1}^{a} c_i^2}$ will produce a standardized contrast with variance σ^2. Effectively, then, the **standardized contrast** is

$$\sum_{i=1}^{a} c_i^* \bar{y}_{i.}$$

where

$$c_i^* = \frac{c_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{a} c_i^2}}$$

Unequal Sample Sizes. When the sample sizes in each treatment are different, minor modifications are made in the above results. First, note that the definition of a contrast now requires that

$$\sum_{i=1}^{a} n_i c_i = 0$$

Other required changes are straightforward. For example, the t statistic in Equation 3.27 becomes

$$t_0 = \frac{\sum_{i=1}^{a} c_i \bar{y}_{i.}}{\sqrt{MS_E \sum_{i=1}^{a} \frac{c_i^2}{n_i}}}$$

and the contrast sum of squares from Equation 3.29 becomes

$$SS_C = \frac{\left(\sum_{i=1}^{a} c_i \bar{y}_{i.}\right)^2}{\sum_{i=1}^{a} \frac{c_i^2}{n_i}}$$

3.5.5 Orthogonal Contrasts

A useful special case of the procedure in Section 3.5.4 is that of **orthogonal contrasts.** Two contrasts with coefficients $\{c_i\}$ and $\{d_i\}$ are orthogonal if

$$\sum_{i=1}^{a} c_i d_i = 0$$

or, for an unbalanced design, if

$$\sum_{i=1}^{a} c_i d_i / n_i = 0$$

For a treatments, the set of $a - 1$ orthogonal contrasts partition the sum of squares due to treatments into $a - 1$ independent single-degree-of-freedom components. Thus, tests performed on orthogonal contrasts are independent.

There are many ways to choose the orthogonal contrast coefficients for a set of treatments. Usually, something in the nature of the experiment should suggest which comparisons will be of interest. For example, if there are $a = 3$ treatments, with treatment 1 a control and treatments 2 and 3 actual levels of the factor of interest to the experimenter, appropriate orthogonal contrasts might be as follows:

Treatment	Coefficients for Orthogonal Contrasts	
1 (control)	−2	0
2 (level 1)	1	−1
3 (level 2)	1	1

Note that contrast 1 with $c_i = -2, 1, 1$ compares the average effect of the factor with the control, whereas contrast 2 with $d_i = 0, -1, 1$ compares the two levels of the factor of interest.

Generally, the method of contrasts (or orthogonal contrasts) is useful for what are called **preplanned comparisons.** That is, the contrasts are specified prior to running the experiment and examining the data. The reason for this is that if comparisons are selected after examining the data, most experimenters would construct tests that correspond to large observed differences in means. These large differences could be the result of the presence of real effects, or they could be the result of random error. If experimenters consistently pick the largest differences to compare, they will inflate the type I error of the test because it is likely that, in an unusually high percentage of the comparisons selected, the observed differences will be the result of error. Examining the data to select comparisons of potential interest is often called **data snooping.** The Scheffé method for all comparisons, discussed in the next section, permits data snooping.

EXAMPLE 3.6

Consider the plasma etching experiment in Example 3.1. There are four treatment means and three degrees of freedom between these treatments. Suppose that prior to running the experiment the following set of comparisons among the treatment means (and their associated contrasts) were specified:

Hypothesis	Contrast
$H_0 : \mu_1 = \mu_2$	$C_1 = \bar{y}_{1.} - \bar{y}_{2.}$
$H_0 : \mu_1 + \mu_2 = \mu_3 + \mu_4$	$C_2 = \bar{y}_{1.} + \bar{y}_{2.} - \bar{y}_{3.} - \bar{y}_{4.}$
$H_0 : \mu_3 = \mu_4$	$C_3 = \bar{y}_{3.} - \bar{y}_{4.}$

Notice that the contrast coefficients are orthogonal. Using the data in Table 3.4, we find the numerical values of the contrasts and the sums of squares to be as follows:

$$C_1 = +1(551.2) - 1(587.4) = -36.2$$

$$SS_{C_1} = \frac{(-36.2)^2}{\frac{1}{5}(2)} = 3276.10$$

$$C_2 = \begin{array}{l} +1(551.2) + 1(587.4) \\ -1(625.4) - 1(707.0) \end{array} = -193.8$$

$$SS_{C_2} = \frac{(-193.8)^2}{\frac{1}{5}(4)} = 46{,}948.05$$

$$C_3 = +1(625.4) - 1(707.6) = -81.6$$

$$SS_{C_3} = \frac{(-81.6)^2}{\frac{1}{5}(2)} = 16{,}646.40$$

These contrast sums of squares completely partition the treatment sum of squares. The tests on such orthogonal contrasts are usually incorporated in the ANOVA, as shown in Table 3.11. We conclude from the *P*-values that there are significant differences in mean etch rates between levels 1 and 2 and between levels 3 and 4 of the power settings, and that the *average* of levels 1 and 2 does differ significantly from the average of levels 3 and 4 at the $\alpha = 0.05$ level.

■ **TABLE 3.11**
Analysis of Variance for the Plasma Etching Experiment

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	*P*-Value
Power setting	66,870.55	3	22,290.18	66.80	< 0.001
Orthogonal contrasts					
$C_1 : \mu_1 = \mu_2$	(3276.10)	1	3276.10	9.82	< 0.01
$C_2 : \mu_1 + \mu_3 = \mu_3 + \mu_4$	(46,948.05)	1	46,948.05	140.69	< 0.001
$C_3 : \mu_3 = \mu_4$	(16,646.40)	1	16,646.40	49.88	< 0.001
Error	5,339.20	16	333.70		
Total	72,209.75	19			

3.5.6 Scheffé's Method for Comparing All Contrasts

In many situations, experimenters may not know in advance which contrasts they wish to compare, or they may be interested in more than $a - 1$ possible comparisons. In many exploratory experiments, the comparisons of interest are discovered only after preliminary examination of the data. Scheffé (1953) has proposed a method for comparing any and all possible contrasts between treatment means. In the Scheffé method, the type I error is at most α for any of the possible comparisons.

Suppose that a set of m contrasts in the treatment means

$$\Gamma_u = c_{1u}\mu_1 + c_{2u}\mu_2 + \cdots + c_{au}\mu_a \quad u = 1, 2, \ldots, m \tag{3.31}$$

of interest have been determined. The corresponding contrast in the treatment averages $\overline{y}_{i.}$ is

$$C_u = c_{1u}\overline{y}_{1.} + c_{2u}\overline{y}_{2.} + \cdots + c_{au}\overline{y}_{a.} \quad u = 1, 2, \ldots, m \tag{3.32}$$

and the **standard error** of this contrast is

$$S_{C_u} = \sqrt{MS_E \sum_{i=1}^{a}(c_{iu}^2/n_i)} \tag{3.33}$$

where n_i is the number of observations in the ith treatment. It can be shown that the critical value against which C_u should be compared is

$$S_{\alpha,u} = S_{C_u}\sqrt{(a-1)F_{\alpha,a-1,N-a}} \tag{3.34}$$

To test the hypothesis that the contrast Γ_u differs significantly from zero, refer C_u to the critical value. If $|C_u| > S_{\alpha,u}$, the hypothesis that the contrast Γ_u equals zero is rejected.

The Scheffé procedure can also be used to form confidence intervals for all possible contrasts among treatment means. The resulting intervals, say $C_u - S_{\alpha,u} \leq \Gamma_u \leq C_u + S_{\alpha,u}$, are **simultaneous confidence intervals** in that the probability that all of them are simultaneously true is at least $1 - \alpha$.

To illustrate the procedure, consider the data in Example 3.1 and suppose that the contrasts of interests are

$$\Gamma_1 = \mu_1 + \mu_2 - \mu_3 - \mu_4$$

and

$$\Gamma_2 = \mu_1 - \mu_4$$

The numerical values of these contrasts are

$$\begin{aligned} C_1 &= \overline{y}_{1.} + \overline{y}_{2.} - \overline{y}_{3.} - \overline{y}_{4.} \\ &= 551.2 + 587.4 - 625.4 - 707.0 = -193.80 \end{aligned}$$

and

$$\begin{aligned} C_2 &= \overline{y}_{1.} - \overline{y}_{4.} \\ &= 551.2 - 707.0 = -155.8 \end{aligned}$$

The standard errors are found from Equation 3.33 as

$$S_{C_1} = \sqrt{MS_E \sum_{i=1}^{5}(c_{i1}^2/n_i)} = \sqrt{333.70(1+1+1+1)/5} = 16.34$$

and

$$S_{C_2} = \sqrt{MS_E \sum_{i=1}^{5}(c_{i2}^2/n_i)} = \sqrt{333.70(1+1)/5} = 11.55$$

From Equation 3.34, the 1 percent critical values are

$$S_{0.01,1} = S_{C_1}\sqrt{(a-1)F_{0.01,a-1,N-a}} = 16.34\sqrt{3(5.29)} = 65.09$$

and

$$S_{0.01,2} = S_{C_2}\sqrt{(a-1)F_{0.01,a-1,N-a}} = 11.55\sqrt{3(5.29)} = 45.97$$

Because $|C_1| > S_{0.01,1}$, we conclude that the contrast $\Gamma_1 = \mu_1 + \mu_2 - \mu_3 - \mu_4$ does not equal zero; that is, we conclude that the mean etch rates of power settings 1 and 2 as a group differ from the means of power settings 3 and 4 as a group. Furthermore, because $|C_2| > S_{0.01,2}$, we conclude that the contrast $\Gamma_2 = \mu_1 - \mu_4$ does not equal zero; that is, the mean etch rates of treatments 1 and 4 differ significantly.

3.5.7 Comparing Pairs of Treatment Means

In many practical situations, we will wish to compare only **pairs of means.** Frequently, we can determine which means differ by testing the differences between *all* pairs of treatment means. Thus, we are interested in contrasts of the form $\Gamma = \mu_j - \mu_j$ for all $i \neq j$. Although the Scheffé method described in the previous section could be easily applied to this problem, it is not the most sensitive procedure for such comparisons. We now turn to a consideration of methods specifically designed for pairwise comparisons between all a population means.

Suppose that we are interested in comparing all pairs of a treatment means and that the null hypotheses that we wish to test are $H_0 : \mu_i = \mu_j$ for all $i \neq j$. There are numerous procedures available for this problem. We now present two popular methods for making such comparisons.

Tukey's Test. Suppose that, following an ANOVA in which we have rejected the null hypothesis of equal treatment means, we wish to test all pairwise mean comparisons:

$$H_0 : \mu_i = \mu_j$$
$$H_1 : \mu_i \neq \mu_j$$

for all $i \neq j$. Tukey (1953) proposed a procedure for testing hypotheses for which the overall significance level is exactly α when the sample sizes are equal and at most α when the sample sizes are unequal. His procedure can also be used to construct confidence intervals on the differences in all pairs of means. For these intervals, the simultaneous confidence level is $100(1 - \alpha)$ percent when the sample sizes are equal and at least $100(1 - \alpha)$ percent when sample sizes are unequal. In other words, the Tukey procedure controls the **experimentwise** or "family" error rate at the selected level α. This is an excellent data snooping procedure when interest focuses on pairs of means.

Tukey's procedure makes use of the distribution of the **studentized range statistic**

$$q = \frac{\bar{y}_{\max} - \bar{y}_{\min}}{\sqrt{MS_E/n}}$$

where $\bar{y}_{\max}$ and $\bar{y}_{\min}$ are the largest and smallest sample means, respectively, out of a group of p sample means. Appendix Table V contains values of $q_\alpha(p,f)$, the upper α percentage points of q, where f is the number of degrees of freedom associated with the MS_E. For equal sample sizes, Tukey's test declares two means significantly different if the absolute value of their sample differences exceeds

$$T_\alpha = q_\alpha(a,f)\sqrt{\frac{MS_E}{n}} \tag{3.35}$$

Equivalently, we could construct a set of $100(1 - \alpha)$ percent confidence intervals for all pairs of means as follows:

$$\begin{aligned} \bar{y}_{i.} - \bar{y}_{j.} - q_\alpha(a,f)\sqrt{\frac{MS_E}{n}} &\leq \mu_i - \mu_j \\ &\leq \bar{y}_{i.} - \bar{y}_{j.} + q_\alpha(a,f)\sqrt{\frac{MS_E}{n}}, i \neq j. \end{aligned} \tag{3.36}$$

When sample sizes are not equal, Equations 3.35 and 3.36 become

$$T_\alpha = \frac{q_\alpha(a,f)}{\sqrt{2}}\sqrt{MS_E\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{3.37}$$

and

$$\bar{y}_{i.} - \bar{y}_{j.} - \frac{q_\alpha(a,f)}{\sqrt{2}}\sqrt{MS_E\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \le \mu_i - \mu_j$$
$$\le \bar{y}_{i.} - \bar{y}_{j.} + \frac{q_\alpha(a,f)}{\sqrt{2}}\sqrt{MS_E\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}, i \neq j \qquad \textbf{(3.38)}$$

respectively. The unequal sample size version is sometimes called the **Tukey–Kramer procedure.**

EXAMPLE 3.7

To illustrate Tukey's test, we use the data from the plasma etching experiment in Example 3.1. With $\alpha = 0.05$ and $f = 16$ degrees of freedom for error, Appendix Table V gives $q_{0.05}(4, 16) = 4.05$. Therefore, from Equation 3.35,

$$T_{0.05} = q_{0.05}(4, 16)\sqrt{\frac{MS_E}{n}} = 4.05\sqrt{\frac{333.70}{5}} = 33.09$$

Thus, any pairs of treatment averages that differ in absolute value by more than 33.09 would imply that the corresponding pair of population means are significantly different. The four treatment averages are

$$\bar{y}_{1.} = 551.2 \qquad \bar{y}_{2.} = 587.4$$
$$\bar{y}_{3.} = 625.4 \qquad \bar{y}_{4.} = 707.0$$

and the differences in averages are

$$\bar{y}_{1.} - \bar{y}_{2.} = 551.2 - 587.4 = -36.20^*$$
$$\bar{y}_{1.} - \bar{y}_{3.} = 551.2 - 625.4 = -74.20^*$$
$$\bar{y}_{1.} - \bar{y}_{4.} = 551.2 - 707.0 = -155.8^*$$
$$\bar{y}_{2.} - \bar{y}_{3.} = 587.4 - 625.4 = -38.0^*$$
$$\bar{y}_{2.} - \bar{y}_{4.} = 587.4 - 707.0 = -119.6^*$$
$$\bar{y}_{3.} - \bar{y}_{4.} = 625.4 - 707.0 = -81.60^*$$

The starred values indicate the pairs of means that are significantly different. Note that the Tukey procedure indicates that all pairs of means differ. Therefore, each power setting results in a mean etch rate that differs from the mean etch rate at any other power setting.

When using any procedure for pairwise testing of means, we occasionally find that the overall F-test from the ANOVA is significant, but the pairwise comparison of means fails to reveal any significant differences. This situation occurs because the F-test is simultaneously considering all possible contrasts involving the treatment means, not just pairwise comparisons. That is, in the data at hand, the significant contrasts may not be of the form $\mu_i - \mu_j$.

The derivation of the Tukey confidence interval of Equation 3.36 for equal sample sizes is straightforward. For the studentized range statistic q, we have

$$P\left(\frac{\max(\bar{y}_{i.} - \mu_i) - \min(\bar{y}_{i.} - \mu_i)}{\sqrt{MS_E/n}} \le q_\alpha(a,f)\right) = 1 - \alpha$$

If $\max(\bar{y}_{i.} - \mu_i) - \min(\bar{y}_{i.} - \mu_i)$ is less than or equal to $q_\alpha(a,f)\sqrt{MS_E/n}$, it must be true that $|(\bar{y}_{i.} - \mu_i) - (\bar{y}_{j.} - \mu_j)| \le q_\alpha(a,f)\sqrt{MS_E/n}$ for every pair of means. Therefore

$$P\left(-q_\alpha(a,f)\sqrt{\frac{MS_E}{n}} \le \bar{y}_{i.} - \bar{y}_{j.} - (\mu_i - \mu_j) \le q_\alpha(a,f)\sqrt{\frac{MS_E}{n}}\right) = 1 - \alpha$$

Rearranging this expression to isolate $\mu_i - \mu_j$ between the inequalities will lead to the set of $100(1 - \alpha)$ percent simultaneous confidence intervals given in Equation 3.38.

The Fisher Least Significant Difference (LSD) Method. The Fisher method for comparing all pairs of means controls the error rate α for each individual pairwise comparison but does not control the experimentwise or family error rate. This procedure uses the t statistic for testing $H_0: \mu_i = \mu_j$

$$t_0 = \frac{\bar{y}_{i.} - \bar{y}_{j.}}{\sqrt{MS_E\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{3.39}$$

Assuming a two-sided alternative, the pair of means μ_i and μ_j would be declared significantly different if $|\bar{y}_{i.} - \bar{y}_{j.}| > t_{\alpha/2,N-a}\sqrt{MS_E(1/n_i + 1/n_j)}$. The quantity

$$\text{LSD} = t_{\alpha/2,N-a}\sqrt{MS_E\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{3.40}$$

is called the **least significant difference.** If the design is balanced, $n_1 = n_2 = \cdots = n_a = n$, and

$$\text{LSD} = t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}} \tag{3.41}$$

To use the Fisher LSD procedure, we simply compare the observed difference between each pair of averages to the corresponding LSD. If $|\bar{y}_{i.} - \bar{y}_{j.}| > \text{LSD}$, we conclude that the population means μ_i and μ_j differ. The t statistic in Equation 3.39 could also be used.

EXAMPLE 3.8

To illustrate the procedure, if we use the data from the experiment in Example 3.1, the LSD at $\alpha = 0.05$ is

$$\text{LSD} = t_{.025,16}\sqrt{\frac{2MS_E}{n}} = 2.120\sqrt{\frac{2(333.70)}{5}} = 24.49$$

Thus, any pair of treatment averages that differ in absolute value by more than 24.49 would imply that the corresponding pair of population means are significantly different. The differences in averages are

$$\bar{y}_{1.} - \bar{y}_{2.} = 551.2 - 587.4 = -36.2^*$$
$$\bar{y}_{1.} - \bar{y}_{3.} = 551.2 - 625.4 = -74.2^*$$
$$\bar{y}_{1.} - \bar{y}_{4.} = 551.2 - 707.0 = -155.8^*$$
$$\bar{y}_{2.} - \bar{y}_{3.} = 587.4 - 625.4 = -38.0^*$$
$$\bar{y}_{2.} - \bar{y}_{4.} = 587.4 - 707.0 = -119.6^*$$
$$\bar{y}_{3.} - \bar{y}_{4.} = 625.4 - 707.0 = -81.6^*$$

The starred values indicate pairs of means that are significantly different. Clearly, all pairs of means differ significantly.

Note that the overall α risk may be considerably inflated using this method. Specifically, as the number of treatments a gets larger, the experimentwise or family type I error rate (the ratio of the number of experiments in which at least one type I error is made to the total number of experiments) becomes large.

Which Pairwise Comparison Method Do I Use? Certainly, a logical question at this point is as follows: Which one of these procedures should I use? Unfortunately, there is no clear-cut answer to this question, and professional statisticians often disagree over the utility of the various procedures. Carmer and Swanson (1973) have conducted Monte Carlo simulation studies of a number of multiple comparison procedures, including others not discussed here. They report that the least significant difference method is a very effective test for detecting true differences in means if it is applied *only after* the F-test in the ANOVA is significant at 5 percent. However, this method does not contain the experimentwise error rate. Because the Tukey method does control the overall error rate, many statisticians prefer to use it.

As indicated above, there are several other multiple comparison procedures. For articles describing these methods, see O'Neill and Wetherill (1971), Miller (1977), and Nelson (1989). The books by Miller (1991) and Hsu (1996) are also recommended.

3.5.8 Comparing Treatment Means with a Control

In many experiments, one of the treatments is a **control**, and the analyst is interested in comparing each of the other $a - 1$ treatment means with the control. Thus, only $a - 1$ comparisons are to be made. A procedure for making these comparisons has been developed by Dunnett (1964). Suppose that treatment a is the control and we wish to test the hypotheses

$$H_0\colon \mu_i = \mu_a$$
$$H_1\colon \mu_i \neq \mu_a$$

for $i = 1, 2, \ldots, a - 1$. Dunnett's procedure is a modification of the usual t-test. For each hypothesis, we compute the observed differences in the sample means

$$|\bar{y}_{i.} - \bar{y}_{a.}| \qquad i = 1, 2, \ldots, a - 1$$

The null hypothesis $H_0\colon \mu_i = \mu_a$ is rejected using a type I error rate α if

$$|\bar{y}_{i.} - \bar{y}_{a.}| > d_\alpha(a-1, f)\sqrt{MS_E\left(\frac{1}{n_i} + \frac{1}{n_a}\right)} \tag{3.42}$$

where the constant $d_\alpha(a - 1, f)$ is given in Appendix Table VI. (Both two- and one-sided tests are possible.) Note that α is the **joint significance level** associated with all $a - 1$ tests.

EXAMPLE 3.9

To illustrate Dunnett's test, consider the experiment from Example 3.1 with treatment 4 considered as the control. In this example, $a = 4, a - 1 = 3, f = 16$, and $n_i = n = 5$. At the 5 percent level, we find from Appendix Table VI that $d_{0.05}(3, 16) = 2.59$. Thus, the critical difference becomes

$$d_{0.05}(3, 16)\sqrt{\frac{2MS_E}{n}} = 2.59\sqrt{\frac{2(333.70)}{5}} = 29.92$$

(Note that this is a simplification of Equation 3.42 resulting from a balanced design.) Thus, any treatment mean that differs in absolute value from the control by more than 29.92 would be declared significantly different. The observed differences are

$$1 \text{ vs. } 4\colon \bar{y}_{1.} - \bar{y}_{4.} = 551.2 - 707.0 = -155.8$$
$$2 \text{ vs. } 4\colon \bar{y}_{2.} - \bar{y}_{4.} = 587.4 - 707.0 = -119.6$$
$$3 \text{ vs. } 4\colon \bar{y}_{3.} - \bar{y}_{4.} = 625.4 - 707.0 = -81.6$$

Note that all differences are significant. Thus, we would conclude that all power settings are different from the control.

When comparing treatments with a control, it is a good idea to use more observations for the control treatment (say n_a) than for the other treatments (say n), assuming equal numbers of observations for the remaining $a - 1$ treatments. The ratio n_a/n should be chosen to be approximately equal to the square root of the total number of treatments. That is, choose $n_a/n = \sqrt{a}$.

3.6 Sample Computer Output

Computer programs for supporting experimental design and performing the analysis of variance are widely available. The output from one such program, Design-Expert, is shown in Figure 3.12, using the data from the plasma etching experiment in Example 3.1. The sum of squares corresponding to the "Model" is the usual $SS_{\text{Treatments}}$ for a single-factor design. That source is further identified as "*A*." When there is more than one factor in the experiment, the model sum of squares will be decomposed into several sources (A, B, etc.). Notice that the analysis of variance summary at the top of the computer output contains the usual sums of squares, degrees of freedom, mean squares, and test statistic F_0. The column "Prob > F" is the *P*-value (actually, the upper bound on the *P*-value because probabilities less than 0.0001 are defaulted to 0.0001).

In addition to the basic analysis of variance, the program displays some other useful information. The quantity "R-squared" is defined as

$$R^2 = \frac{SS_{\text{Model}}}{SS_{\text{Total}}} = \frac{66{,}870.55}{72{,}209.75} = 0.9261$$

and is loosely interpreted as the proportion of the variability in the data "explained" by the ANOVA model. Thus, in the plasma etching experiment, the factor "power" explains about 92.61 percent of the variability in etch rate. Clearly, we must have $0 \leq R^2 \leq 1$, with larger values being more desirable. There are also some other R^2-like statistics displayed in the output. The "adjusted" R^2 is a variation of the ordinary R^2 statistic that reflects the number of factors in the model. It can be a useful statistic for more complex experiments with several design factors when we wish to evaluate the impact of increasing or decreasing the number of model terms. "Std. Dev." is the square root of the error mean square, $\sqrt{333.70} = 18.27$, and "C.V." is the coefficient of variation, defined as $(\sqrt{MS_E}/\bar{y})100$. The coefficient of variation measures the unexplained or residual variability in the data as a percentage of the mean of the response variable. "PRESS" stands for "prediction error sum of squares," and it is a measure of how well the model for the experiment is likely to predict the responses in a *new experiment*. Small values of PRESS are desirable. Alternatively, one can calculate an R^2 for prediction based on PRESS (we will show how to do this later). This R^2_{Pred} in our problem is 0.8845, which is not unreasonable, considering that the model accounts for about 93 percent of the variability in the current experiment. The "adequate precision" statistic is computed by dividing the difference between the maximum predicted response and the minimum predicted response by the average standard deviation of all predicted responses. Large values of this quantity are desirable, and values that exceed four usually indicate that the model will give reasonable performance in prediction.

Treatment means are estimated, and the standard error (or sample standard deviation of each treatment mean, $\sqrt{MS_E/n}$) is displayed. Differences between pairs of treatment means are investigated by using a hypothesis testing version of the Fisher LSD method described in Section 3.5.7.

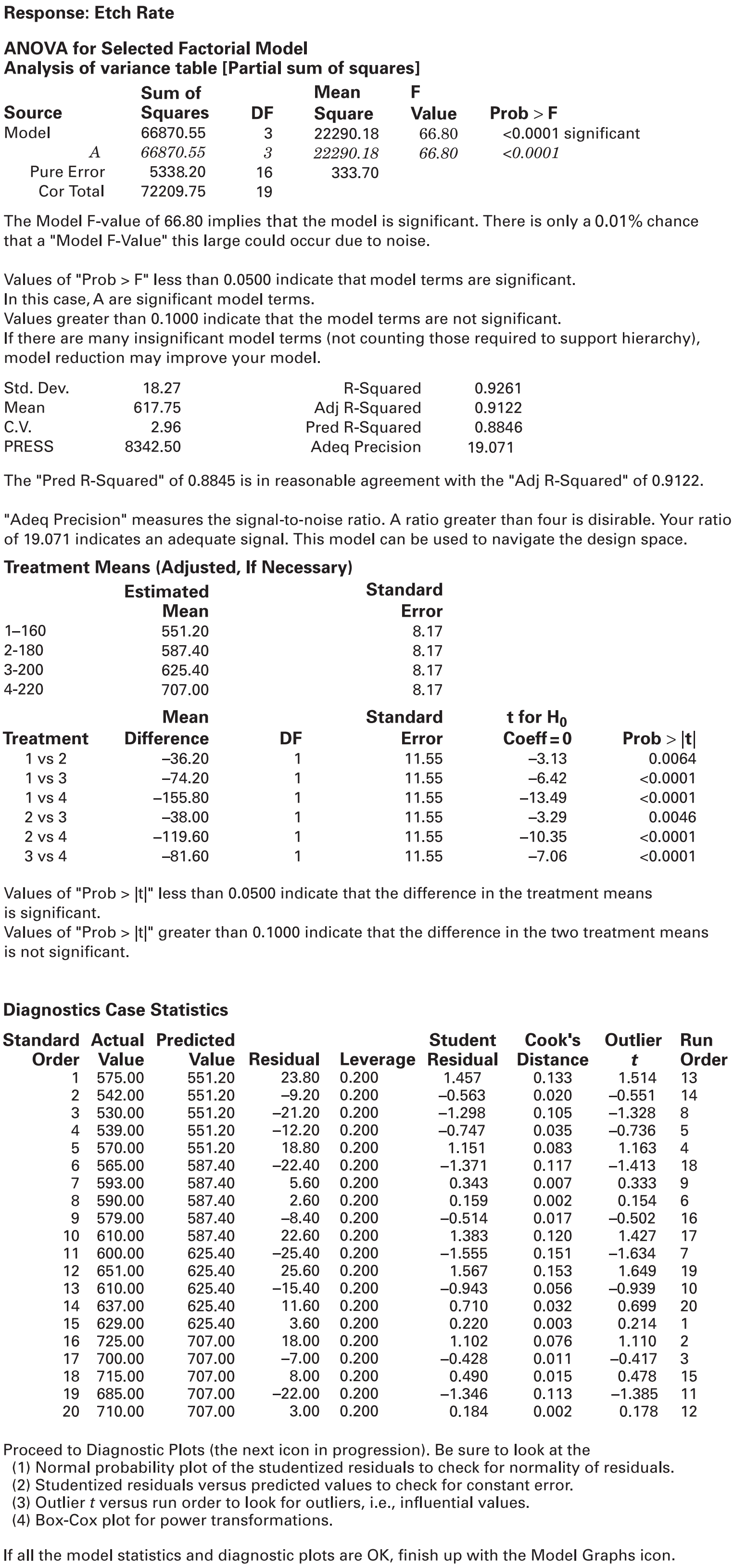
The computer program also calculates and displays the residuals, as defined in Equation 3.16. The program will also produce all of the residual plots that we discussed in Section 3.4. There are also several other residual diagnostics displayed in the output. Some of these will be discussed later. Design-Expert also displays the studentized residual (called "Student Residual" in the output) calculated as

$$r_{ij} = \frac{e_{ij}}{\sqrt{MS_E(1 - \text{Leverage}_{ij})}}$$

where Leverage_{ij} is a measure of the influence of the ijth observation on the model. We will discuss leverage in more detail and show how it is calculated in Chapter 10. Studentized residuals are considered to be more effective in identifying potential outliers rather than either the ordinary residuals or standardized residuals.

Finally, notice that the computer program also has some interpretative guidance embedded in the output. This "advisory" information is fairly standard in many PC-based statistics packages. Remember in reading such guidance that it is written in very general terms and may not exactly suit the report writing requirements of any specific experimenter. This advisory output may be hidden upon request by the user.

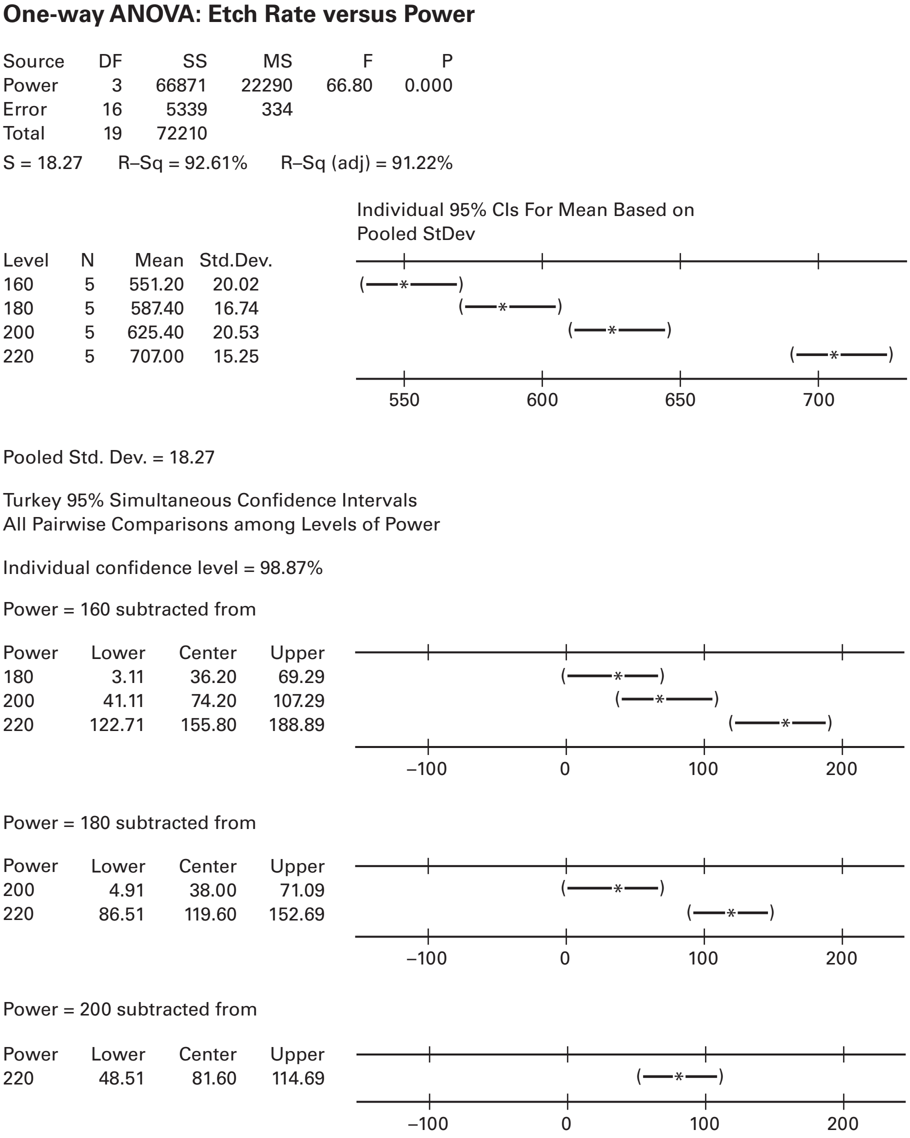
■ **FIGURE 3.12** Design-Expert computer output for Example 3.1



```
Response: Etch Rate

ANOVA for Selected Factorial Model
Analysis of variance table [Partial sum of squares]
                 Sum of               Mean       F
Source           Squares      DF      Square     Value     Prob > F
Model            66870.55      3      22290.18   66.80     <0.0001 significant
          A      66870.55      3      22290.18   66.80     <0.0001
   Pure Error     5338.20     16        333.70
   Cor Total     72209.75     19

The Model F-value of 66.80 implies that the model is significant. There is only a 0.01% chance
that a "Model F-Value" this large could occur due to noise.

Values of "Prob > F" less than 0.0500 indicate that model terms are significant.
In this case, A are significant model terms.
Values greater than 0.1000 indicate that the model terms are not significant.
If there are many insignificant model terms (not counting those required to support hierarchy),
model reduction may improve your model.

Std. Dev.        18.27                  R-Squared      0.9261
Mean            617.75              Adj R-Squared      0.9122
C.V.              2.96             Pred R-Squared      0.8846
PRESS          8342.50             Adeq Precision     19.071

The "Pred R-Squared" of 0.8845 is in reasonable agreement with the "Adj R-Squared" of 0.9122.

"Adeq Precision" measures the signal-to-noise ratio. A ratio greater than four is disirable. Your ratio
of 19.071 indicates an adequate signal. This model can be used to navigate the design space.

Treatment Means (Adjusted, If Necessary)
             Estimated                   Standard
                  Mean                      Error
1–160           551.20                       8.17
2-180           587.40                       8.17
3-200           625.40                       8.17
4-220           707.00                       8.17

                  Mean                   Standard     t for H0
Treatment   Difference        DF            Error     Coeff=0      Prob > |t|
   1 vs 2       –36.20         1            11.55       –3.13          0.0064
   1 vs 3       –74.20         1            11.55       –6.42         <0.0001
   1 vs 4      –155.80         1            11.55      –13.49         <0.0001
   2 vs 3       –38.00         1            11.55       –3.29          0.0046
   2 vs 4      –119.60         1            11.55      –10.35         <0.0001
   3 vs 4       –81.60         1            11.55       –7.06         <0.0001

Values of "Prob > |t|" less than 0.0500 indicate that the difference in the treatment means
is significant.
Values of "Prob > |t|" greater than 0.1000 indicate that the difference in the two treatment means
is not significant.


Diagnostics Case Statistics
Standard  Actual  Predicted                      Student   Cook's    Outlier   Run
   Order   Value      Value  Residual  Leverage  Residual  Distance        t   Order
       1  575.00     551.20     23.80  0.200       1.457     0.133     1.514   13
       2  542.00     551.20     –9.20  0.200      –0.563     0.020    –0.551   14
       3  530.00     551.20    –21.20  0.200      –1.298     0.105    –1.328   8
       4  539.00     551.20    –12.20  0.200      –0.747     0.035    –0.736   5
       5  570.00     551.20     18.80  0.200       1.151     0.083     1.163   4
       6  565.00     587.40    –22.40  0.200      –1.371     0.117    –1.413   18
       7  593.00     587.40      5.60  0.200       0.343     0.007     0.333   9
       8  590.00     587.40      2.60  0.200       0.159     0.002     0.154   6
       9  579.00     587.40     –8.40  0.200      –0.514     0.017    –0.502   16
      10  610.00     587.40     22.60  0.200       1.383     0.120     1.427   17
      11  600.00     625.40    –25.40  0.200      –1.555     0.151    –1.634   7
      12  651.00     625.40     25.60  0.200       1.567     0.153     1.649   19
      13  610.00     625.40    –15.40  0.200      –0.943     0.056    –0.939   10
      14  637.00     625.40     11.60  0.200       0.710     0.032     0.699   20
      15  629.00     625.40      3.60  0.200       0.220     0.003     0.214   1
      16  725.00     707.00     18.00  0.200       1.102     0.076     1.110   2
      17  700.00     707.00     –7.00  0.200      –0.428     0.011    –0.417   3
      18  715.00     707.00      8.00  0.200       0.490     0.015     0.478   15
      19  685.00     707.00    –22.00  0.200      –1.346     0.113    –1.385   11
      20  710.00     707.00      3.00  0.200       0.184     0.002     0.178   12

Proceed to Diagnostic Plots (the next icon in progression). Be sure to look at the
  (1) Normal probability plot of the studentized residuals to check for normality of residuals.
  (2) Studentized residuals versus predicted values to check for constant error.
  (3) Outlier t versus run order to look for outliers, i.e., influential values.
  (4) Box-Cox plot for power transformations.

If all the model statistics and diagnostic plots are OK, finish up with the Model Graphs icon.
```

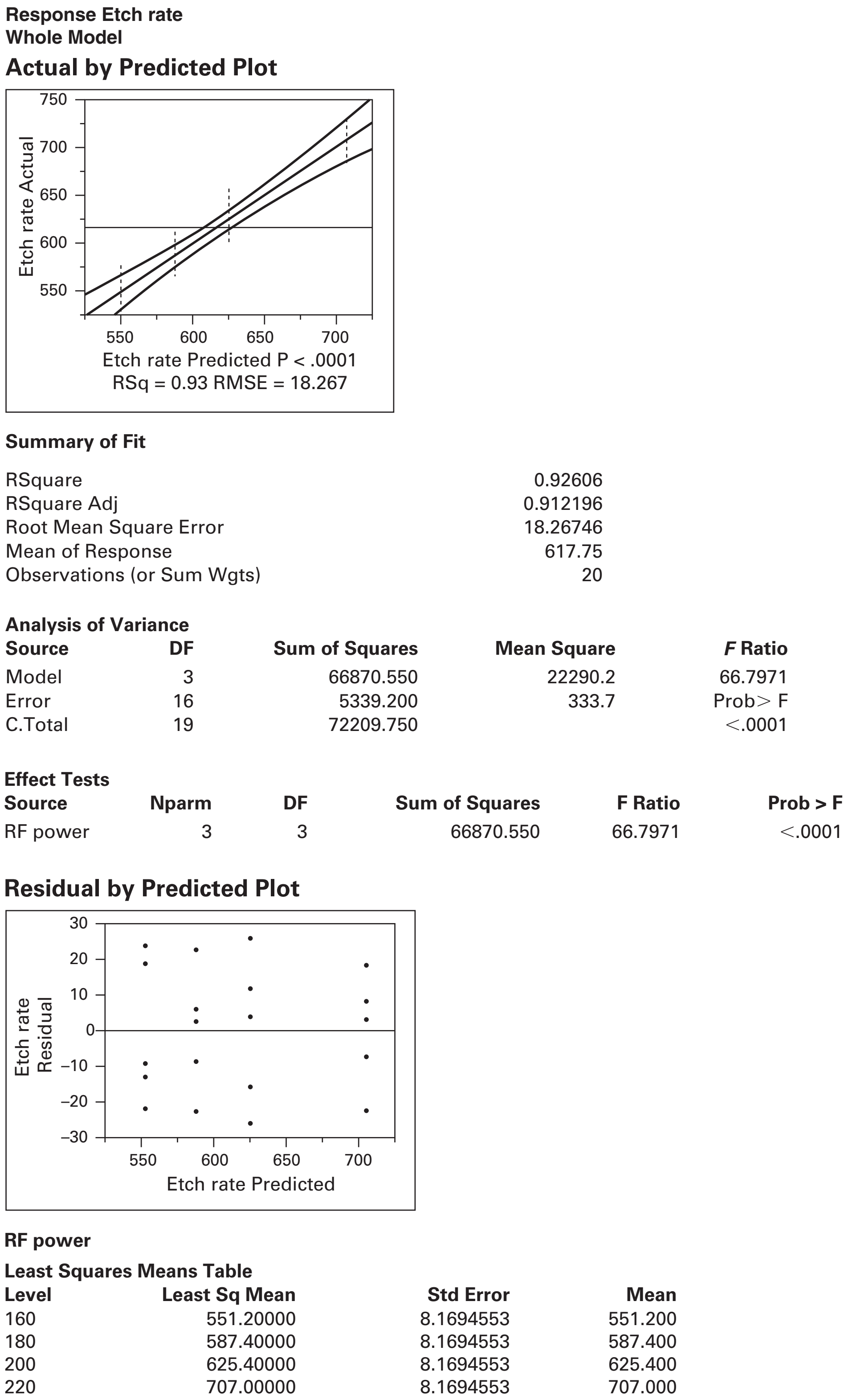
Figure 3.13 presents the output from Minitab for the plasma etching experiment. The output is very similar to the Design-Expert output in Figure 3.12. Note that confidence intervals on each individual treatment mean are provided and that the pairs of means are compared using Tukey's method. However, the Tukey method is presented using the confidence interval format instead of the hypothesis-testing format that we used in Section 3.5.7. None of the Tukey confidence intervals includes zero, so we would conclude that all of the means are different.

Figure 3.14 is the output from JMP for the plasma etch experiment in Example 3.1. The output information is very similar to that from Design-Expert and Minitab. The plots of actual observations versus the predicted values and residuals versus the predicted values are default output. There is an option in JMP to provide the Fisher LSD procedure or Tukey's method to compare all pairs of means.



```
One-way ANOVA: Etch Rate versus Power

Source   DF     SS     MS      F      P
Power     3  66871  22290  66.80  0.000
Error    16   5339    334
Total    19  72210
S = 18.27    R-Sq = 92.61%    R-Sq (adj) = 91.22%

                               Individual 95% CIs For Mean Based on
                               Pooled StDev
Level  N    Mean  Std.Dev.   ----+---------+---------+---------+----
160    5  551.20     20.02   (--*---)
180    5  587.40     16.74          (--*---)
200    5  625.40     20.53                 (--*---)
220    5  707.00     15.25                               (--*---)
                             ----+---------+---------+---------+----
                                550       600       650       700

Pooled Std. Dev. = 18.27

Turkey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of Power

Individual confidence level = 98.87%

Power = 160 subtracted from

Power   Lower  Center   Upper   -----+---------+---------+---------+---
180      3.11   36.20   69.29                  (--*---)
200     41.11   74.20  107.29                     (--*---)
220    122.71  155.80  188.89                            (--*---)
                                -----+---------+---------+---------+---
                                  -100         0       100       200

Power = 180 subtracted from

Power   Lower  Center   Upper   -----+---------+---------+---------+---
200      4.91   38.00   71.09                  (--*---)
220     86.51  119.60  152.69                         (--*--)
                                -----+---------+---------+---------+---
                                  -100         0       100       200

Power = 200 subtracted from

Power   Lower  Center   Upper   -----+---------+---------+---------+---
220     48.51   81.60  114.69                      (--*--)
                                -----+---------+---------+---------+---
                                  -100         0       100       200
```

■ **FIGURE 3.13** Minitab computer output for Example 3.1



```
Response Etch rate
Whole Model
Actual by Predicted Plot
```

```
Summary of Fit

RSquare                                          0.92606
RSquare Adj                                     0.912196
Root Mean Square Error                          18.26746
Mean of Response                                  617.75
Observations (or Sum Wgts)                            20

Analysis of Variance
Source          DF      Sum of Squares      Mean Square        F Ratio
Model            3           66870.550          22290.2        66.7971
Error           16            5339.200            333.7        Prob> F
C.Total         19           72209.750                          <.0001

Effect Tests
Source          Nparm      DF      Sum of Squares      F Ratio       Prob > F
RF power            3       3           66870.550      66.7971         <.0001

Residual by Predicted Plot
```

```
RF power
Least Squares Means Table
Level       Least Sq Mean           Std Error          Mean
160             551.20000           8.1694553       551.200
180             587.40000           8.1694553       587.400
200             625.40000           8.1694553       625.400
220             707.00000           8.1694553       707.000
```

■ FIGURE 3.14 JMP output from Example 3.1

3.7 Determining Sample Size

In any experimental design problem, a critical decision is the choice of sample size—that is, determining the number of replicates to run. Generally, if the experimenter is interested in detecting small effects, more replicates are required than if the experimenter is interested in detecting large effects. In this section, we discuss several approaches to determining sample size. Although our discussion focuses on a single-factor design, most of the methods can be used in more complex experimental situations.

3.7.1 Operating Characteristic and Power Curves

Recall that an **operating characteristic (OC) curve** is a plot of the type II error probability β of a statistical test for a particular sample size versus a parameter that reflects the extent to which the null hypothesis is false. Alternatively a *Power Curve* plots power or 1−β versus this parameter. Power and/or OC curves can be constructed from software and are useful in guiding the experimenter in selecting the number of replicates so that the design will be sensitive to important potential differences in the treatments.

We consider the probability of type II error of the fixed effects model for the case of equal sample sizes per treatment, say

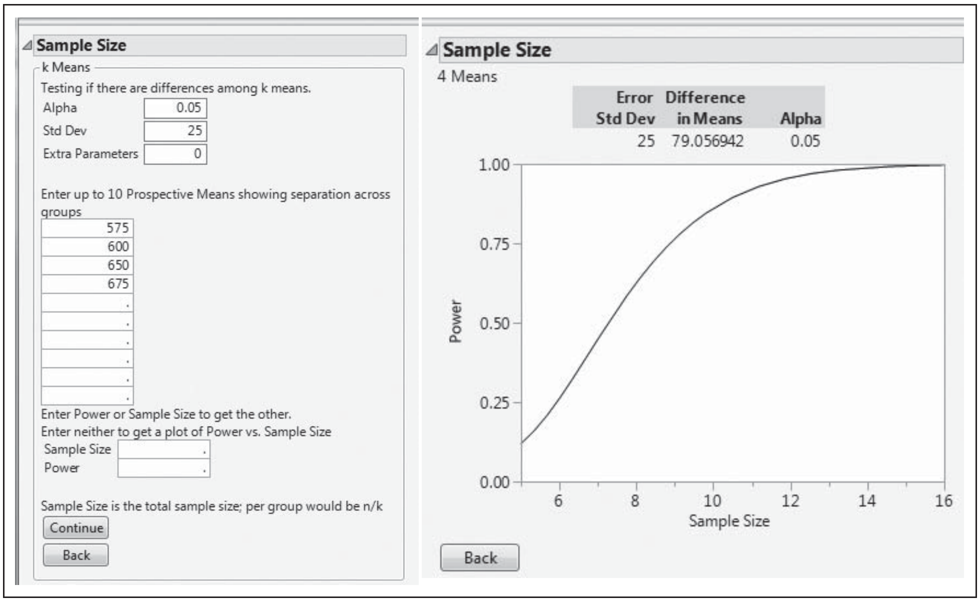
$$\begin{aligned}\beta &= 1 - P\{\text{Reject } H_0 | H_0 \text{ is false}\}\\ &= 1 - P\{F_0 > F_{\alpha,a-1,N-a} | H_0 \text{ is false}\}\end{aligned} \quad \textbf{(3.43)}$$

To evaluate the probability statement in Equation 3.43, we need to know the distribution of the test statistic F_0 if the null hypothesis is false. It can be shown that, if H_0 is false, the statistic $F_0 = MS_{\text{Treatments}}/MS_E$ is distributed as a **noncentral *F*** random variable with $a - 1$ and $N - a$ degrees of freedom and the noncentrality parameter δ. If $\delta = 0$, the noncentral F distribution becomes the usual (central) F distribution.

We will illustrate the sample size determination method implemented in JMP. Consider the plasma etching experiment described in Exampe 3.1 Suppose that the experimenter is interested in rejecting the null hypothesis with a probability of at least 0.9 (power = 0.9) if the true treatment means are

$$\mu_1 = 575,\ \mu_2 = 600,\ \mu_3 = 650, \text{and } \mu_1 = 675$$

The experimenter feels that the standard deviation of etch rate will be no larger than $\sigma = 25$ Å/min. The input and output from the JMP power and sample size platform for comparing several means is shown in the following display:



Sample Size

k Means

Testing if there are differences among k means.

Alpha	0.05
Std Dev	25
Extra Parameters	0

Enter up to 10 Prospective Means showing separation across groups

575
600
650
675
.
.
.
.
.
.

Enter Power or Sample Size to get the other.
Enter neither to get a plot of Power vs. Sample Size

Sample Size	.
Power	.

Sample Size is the total sample size; per group would be n/k

Sample Size

4 Means

Error Std Dev	Difference in Means	Alpha
25	79.056942	0.05

The graph on the right is a plot of power versus the total sample size. This plot indicates that at least 4 replicates are required to obtain a power that exceeds 0.90.

A potential problem with this approach to determining sample size is that it can be difficult to select a set of treatment means on which the sample size decision should be based. An alternate approach is to select a sample size such that if the difference between any two treatment means exceeds a specified value, the null hypothesis should be rejected.

Minitab uses this approach to perform power calculations and find sample sizes for single-factor ANOVAs. Consider the following display:

```
Power and Sample Size

One-way ANOVA

Alpha = 0.01 Assumed standard deviation = 25
Number of Levels = 4

                 Sample                           Maximum
SS Means           Size         Power          Difference
  2812.5              5      0.804838                  75

The sample size is for each level.

Power and Sample Size

One-way ANOVA

Alpha = 0.01 Assumed standard deviation = 25
Number of Levels 5 4

              Sample     Target                           Maximum
SS Means        Size      Power      Actual Power      Difference
  2812.5           6        0.9          0.915384              75

The sample size is for each level.
```

In the upper portion of the display, we asked Minitab to calculate the power for $n = 5$ replicates when the maximum difference in treatment means is 75. The bottom portion of the display is the output when the experimenter requests the sample size to obtain a target power of at least 0.90.

3.7.2 Confidence Interval Estimation Method

This approach assumes that the experimenter wishes to express the final results in terms of confidence intervals and is willing to specify in advance how wide he or she wants these confidence intervals to be. For example, suppose that in the plasma etching experiment from Example 3.1, we wanted a 95 percent confidence interval on the difference in mean etch rate for any two power settings to be ±30 Å/min and a prior estimate of σ is 25. Then, using Equation 3.13, we find that the accuracy of the confidence interval is

$$\pm t_{\alpha/2,N-a}\sqrt{\frac{2MS_E}{n}}$$

Suppose that we try $n = 5$ replicates. Then, using $\sigma^2 = (25)^2 = 625$ as an estimate of MS_E, the accuracy of the confidence interval becomes

$$\pm 2.120\sqrt{\frac{2(625)}{5}} = \pm 33.52$$

which does not meet the requirement. Trying $n = 6$ gives

$$\pm 2.086\sqrt{\frac{2(625)}{6}} = \pm 30.11$$

Trying $n = 7$ gives

$$\pm 2.064\sqrt{\frac{2(625)}{7}} = \pm 27.58$$

Clearly, $n = 7$ is the smallest sample size that will lead to the desired accuracy.

The quoted level of significance in the above illustration applies only to one confidence interval. However, the same general approach can be used if the experimenter wishes to prespecify a *set* of confidence intervals about which a **joint** or **simultaneous confidence statement** is made (see the comments about simultaneous confidence intervals in Section 3.3.3). Furthermore, the confidence intervals could be constructed about more general contrasts in the treatment means than the pairwise comparison illustrated above.

3.8 Other Examples of Single-Factor Experiments

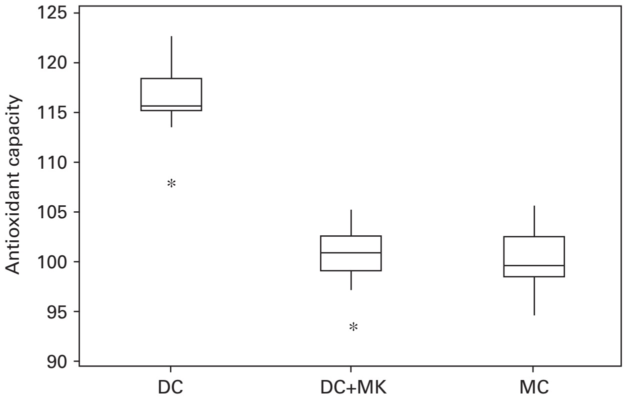
3.8.1 Chocolate and Cardiovascular Health

An article in *Nature* describes an experiment to investigate the effect of consuming chocolate on cardiovascular health (“Plasma Antioxidants from Chocolate,” *Nature*, Vol. 424, 2003, pp. 1013). The experiment consisted of using three different types of chocolates: 100 g of dark chocolate, 100 g of dark chocolate with 200 mL of full-fat milk, and 200 g of milk chocolate. A total of 12 subjects were used, 7 women and 5 men, with an average age range of 32.2 ± 1 years, an average weight of 65.8 ± 3.1 kg, and body-mass index of 21.9 ± 0.4 kg m^{-2}. On different days a subject consumed one of the chocolate-factor levels and 1 hour later the total antioxidant capacity of their blood plasma was measured in an assay. Data similar to that summarized in the article are shown in Table 3.12.

Figure 3.15 presents box plots for the data from this experiment. The result is an indication that the blood antioxidant capacity one hour after eating the dark chocolate is higher than for the other two treatments. The variability in the sample data from all three treatments seems very similar. Table 3.13 is the Minitab ANOVA output. The test statistic is highly significant (Minitab reports a *P*-value of 0.000, which is clearly wrong because *P*-values cannot be zero; this means that the *P*-value is less than 0.001), indicating that some of the treatment means are different. The output also contains the Fisher LSD analysis for this experiment. This indicates that the mean antioxidant capacity after consuming dark chocolate is higher than after consuming dark chocolate plus milk or milk chocolate alone is the mean antioxidant capacity after consuming dark chocolate plus milk or milk chocolate alone is equal. Figure 3.16 is the normal probability plot of the residual and Figure 3.17 is the plot of residuals versus predicted values. These plots do not suggest any problems with model assumptions. We conclude that consuming dark chocolate results in higher mean blood antioxidant capacity after one hour than consuming either dark chocolate plus milk or milk chocolate alone.

■ **TABLE 3.12**
Blood Plasma Levels One Hour Following Chocolate Consumption

	Subjects (Observations)											
Factor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
DC	118.8	122.6	115.6	113.6	119.5	115.9	115.8	115.1	116.9	115.4	115.6	107.9
DC + MK	105.4	101.1	102.7	97.1	101.9	98.9	100.0	99.8	102.6	100.9	104.5	93.5
MC	102.1	105.8	99.6	102.7	98.8	100.9	102.8	98.7	94.7	97.8	99.7	98.6



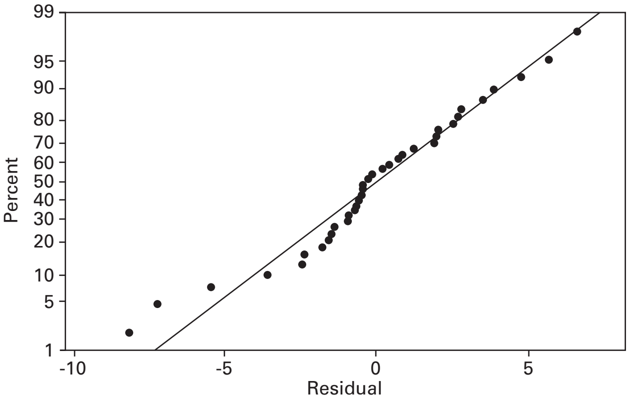
■ **FIGURE 3.15** **Box plots of the blood antioxidant capacity data from the chocolate consumption experiment**

■ **TABLE 3.13**

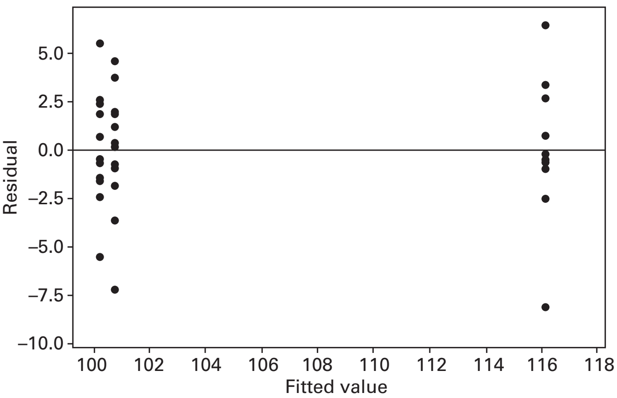
Minitab ANOVA Output, Chocolate Consumption Experiment

One-way ANOVA: DC, DC+MK, MC

```
Source  DF      SS     MS      F      P
Factor   2  1952.6  976.3  93.58  0.000
Error   33   344.3   10.4
Total   35  2296.9

S = 3.230   R-Sq = 85.01%   R-Sq(adj) = 84.10%

                            Individual 95% CIs For Mean Based on
                            Pooled StDev
Level   N    Mean  StDev   ---+---------+---------+---------+------
DC     12  116.06   3.53                                  (---*---)
DC+MK  12  100.70   3.24    (--*---)
MC     12  100.18   2.89   (--*---)
                           ---+---------+---------+---------+------
                          100.0     105.0     110.0     115.0

Pooled StDev = 3.23

Fisher 95% Individual Confidence Intervals
All Pairwise Comparisons
Simultaneous confidence level = 88.02
DC subtracted from:

          Lower   Center    Upper    -+---------+---------+---------+---
DC+MK   -18.041  -15.358  -12.675      (---*----)
MC      -18.558  -15.875  -13.192     (----*---)
                                     -+---------+---------+---------+---
                                  -18.0     -12.0      -6.0       0.0

DC+MK subtracted from:

      Lower  Center  Upper     -+---------+---------+---------+--------
MC   -3.200  -0.517  2.166                               (---*----)
                               -+---------+---------+---------+--------
                            -18.0     -12.0      -6.0       0.0
```



■ **FIGURE 3.16 Normal probability plot of the residuals from the chocolate consumption experiment**



■ **FIGURE 3.17 Plot of residuals versus the predicted values from the chocolate consumption experiment**

3.8.2 A Real Economy Application of a Designed Experiment

Designed experiments have had tremendous impact on manufacturing industries, including the design of new products and the improvement of existing ones, development of new manufacturing processes, and process improvement. In the last 15 years, designed experiments have begun to be widely used outside of this traditional environment. These applications are in financial services, telecommunications, health care, e-commerce, legal services, marketing, logistics and transportation, and many of the nonmanufacturing components of manufacturing businesses. These types of businesses are sometimes referred to as the real economy. It has been estimated that manufacturing accounts for only about 20 percent of the total US economy, so applications of experimental design in the real economy are of growing importance. In this section, we present an example of a designed experiment in marketing.

A soft drink distributor knows that end-aisle displays are an effective way to increase sales of the product. However, there are several ways to design these displays: by varying the text displayed, the colors used, and the visual images. The marketing group has designed three new end-aisle displays and wants to test their effectiveness. They have identified 15 stores of similar size and type to participate in the study. Each store will test one of the displays for a period of one month. The displays are assigned at random to the stores, and each display is tested in five stores. The response variable is the percentage increase in sales activity over the typical sales for that store when the end-aisle display is not in use. The data from this experiment are shown in Table 3.14.

Table 3.15 shows the analysis of the end-aisle display experiment. This analysis was conducted using JMP. The P-value for the model F-statistic in the ANOVA indicates that there is a difference in the mean percentage increase in sales between the three display types. In this application, we had JMP use the Fisher LSD procedure to compare the

■ **TABLE 3.14**
The End-Aisle Display Experimental Design

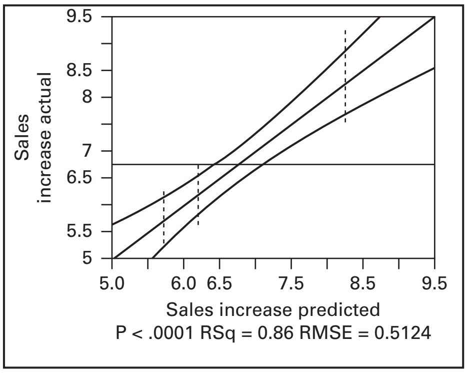
Display Design	Sample Observations, Percent Increase in Sales				
1	5.43	5.71	6.22	6.01	5.29
2	6.24	6.71	5.98	5.66	6.60
3	8.79	9.20	7.90	8.15	7.55

■ **TABLE 3.15**
JMP Output for the End-Aisle Display Experiment

Response Sales Increase

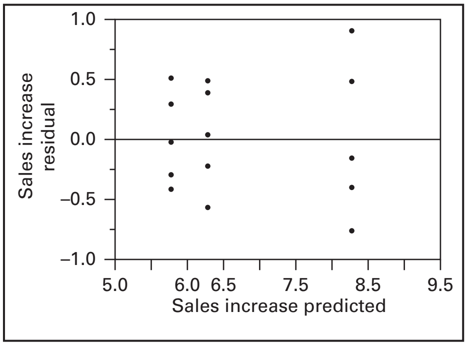
Whole Model

Actual by Predicted Plot



```
Summary of Fit
RSquare                                    0.856364
RSquare Adj                                0.832425
Root Mean Square Error                     0.512383
Mean of Response                           6.762667
Observations (or Sum Wgts)                       15

Analysis of Variance
Source          DF   Sum of Squares   Mean Square    F Ratio
Model            2        18.783053       9.39153    35.7722
Error           12         3.150440       0.26254   Prob > F
C.Total         14        21.933493                  < .0001

Effect Tests
Source                  Nparm   DF   Sum of Squares    F Ratio   Prob > F
Display                     2    2        18.783053    35.7722     < .001
```

Residual by Predicted Plot



(Continued)

■ **TABLE 3.15** (*Continued*)

Least Squares Means Table

Level	**Least Sq Mean**	**Std Error**	**Mean**
1	5.7320000	0.22914479	5.73200
2	6.2380000	0.22914479	6.23800
3	8.3180000	0.22914479	8.31800

LSMeans Differences Student's t

$a = 0.050$ $t = 2.17881$

LSMean[i] By LSMean [i]

Mean[i]-Mean [i] Std Err Dif Lower CL Dif Upper CL Dif	1	2	3
1	0 0 0 0	−0.506 0.32406 −1.2121 0.20007	−2.586 −2.586 −3.2921 −1.8799
2	0.506 0.32406 −0.2001 1.21207	0 0 0 0	−2.08 0.32406 −2.7861 −1.3739
3	2.586 0.32406 1.87993 3.29207	2.08 0.32406 1.37393 2.78607	0 0 0 0

Level		**Least Sq Mean**
3	A	8.3180000
2	B	6.2380000
1	B	5.7320000

Levels not connected by same letter are significantly different.

pairs of treatment means (JMP labels these as the least squares means). The results of this comparison are presented as confidence intervals on the difference in pairs of means. For pairs of means where the confidence interval includes zero, we would not declare that the pairs of means are different. The JMP output indicates that display designs 1 and 2 are similar in that they result in the same mean increase in sales, but that display design 3 is different from both designs 1 and 2 and that the mean increase in sales for display 3 exceeds that of both designs 1 and 2. Notice that JMP automatically includes some useful graphics in the output, a plot of the actual observations versus the predicted values from the model, and a plot of the residuals versus the predicted values. There is some mild indication that display design 3 may exhibit more variability in sales increase than the other two designs.

3.8.3 Discovering Dispersion Effects

We have focused on using the analysis of variance and related methods to determine which factor levels result in differences among treatment or factor level means. It is customary to refer to these effects as **location effects.** If there

■ **TABLE 3.16**

Data for the Smelting Experiment

Ratio Control Algorithm	Observations					
	1	2	3	4	5	6
1	4.93(0.05)	4.86(0.04)	4.75(0.05)	4.95(0.06)	4.79(0.03)	4.88(0.05)
2	4.85(0.04)	4.91(0.02)	4.79(0.03)	4.85(0.05)	4.75(0.03)	4.85(0.02)
3	4.83(0.09)	4.88(0.13)	4.90(0.11)	4.75(0.15)	4.82(0.08)	4.90(0.12)
4	4.89(0.03)	4.77(0.04)	4.94(0.05)	4.86(0.05)	4.79(0.03)	4.76(0.02)

was inequality of variance at the different factor levels, we used transformations to stabilize the variance to improve our inference on the location effects. In some problems, however, we are interested in discovering whether the different factor levels affect **variability**; that is, we are interested in discovering potential **dispersion effects.** This will occur whenever the standard deviation, variance, or some other measure of variability is used as a response variable.

To illustrate these ideas, consider the data in Table 3.16, which resulted from a designed experiment in an aluminum smelter. Aluminum is produced by combining alumina with other ingredients in a reaction cell and applying heat by passing electric current through the cell. Alumina is added continuously to the cell to maintain the proper ratio of alumina to other ingredients. Four different ratio control algorithms were investigated in this experiment. The response variables studied were related to cell voltage. Specifically, a sensor scans cell voltage several times each second, producing thousands of voltage measurements during each run of the experiment. The process engineers decided to use the average voltage and the standard deviation of cell voltage (shown in parentheses) over the run as the response variables. The average voltage is important because it affects cell temperature, and the standard deviation of voltage (called "pot noise" by the process engineers) is important because it affects the overall cell efficiency.

An analysis of variance was performed to determine whether the different ratio control algorithms affect average cell voltage. This revealed that the ratio control algorithm had no **location effect**; that is, changing the ratio control algorithms does not change the average cell voltage. (Refer to Problem 3.38.)

To investigate dispersion effects, it is usually best to use

$$\log(s) \qquad \text{or} \qquad \log(s^2)$$

as a response variable since the log transformation is effective in stabilizing variability in the distribution of the sample standard deviation. Because all sample standard deviations of pot voltage are less than unity, we will use

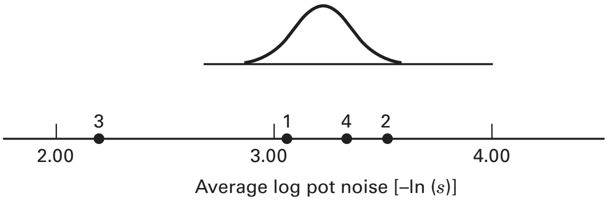
$$y = -\ln(s)$$

as the response variable. Table 3.17 presents the analysis of variance for this response, the natural logarithm of "pot noise." Notice that the choice of a ratio control algorithm affects pot noise; that is, the ratio control algorithm has a **dispersion effect.** Standard tests of model adequacy, including normal probability plots of the residuals, indicate that there are no problems with experimental validity. (Refer to Problem 3.39.)

■ **TABLE 3.17**

Analysis of Variance for the Natural Logarithm of Pot Noise

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	*P*-Value
Ratio control algorithm	6.166	3	2.055	21.96	< 0.001
Error	1.872	20	0.094		
Total	8.038	23			



■ **FIGURE 3.18 Average log pot noise [−ln (*s*)] for four ratio control algorithms relative to a scaled *t* distribution with scale factor $\sqrt{MS_E/n} = \sqrt{0.094/6} = 0.125$**

Figure 3.18 plots the average log pot noise for each ratio control algorithm and also presents a scaled *t* distribution for use as a **reference distribution** in discriminating between ratio control algorithms. This plot clearly reveals that ratio control algorithm 3 produces greater pot noise or greater cell voltage standard deviation than the other algorithms. There does not seem to be much difference between algorithms 1, 2, and 4.

3.9 The Random Effects Model

3.9.1 A Single Random Factor

An experimenter is frequently interested in a factor that has a large number of possible levels. If the experimenter randomly selects *a* of these levels from the population of factor levels, then we say that the factor is **random.** Because the levels of the factor actually used in the experiment were chosen randomly, inferences are made about the entire population of factor levels. We assume that the population of factor levels is either of infinite size or is large enough to be considered infinite. Situations in which the population of factor levels is small enough to employ a finite population approach are not encountered frequently. Refer to Bennett and Franklin (1954) and Searle and Fawcett (1970) for a discussion of the finite population case.

The linear statistical model is

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, n \end{cases} \tag{3.44}$$

where both the treatment effects τ_i and ϵ_{ij} are random variables. We will assume that the treatment effects τ_i are NID $(0, \sigma_\tau^2)$, random variables[3] and that the errors are NID$(0, \sigma^2)$, random variables, and that the τ_i and ϵ_{ij} are independent. Because τ_i is independent of ϵ_{ij}, the variance of any observation is

$$V(y_{ij}) = \sigma_\tau^2 + \sigma^2$$

The variances σ_τ^2 and σ^2 are called **variance components**, and the model (Equation 3.44) is called the **components of variance** or **random effects model.** The observations in the random effects model are normally distributed because they are linear combinations of the two normally and independently distributed random variables τ_i and ϵ_{ij}. However, unlike the fixed effects case in which all of the observations y_{ij} are independent, in the random model the observations y_{ij} are only independent if they come from different factor levels. Specifically, we can show that the covariance of any two observations is

$$\begin{aligned} &Cov\ (y_{ij}, y_{ij'}) = \sigma_\tau^2 \qquad j \neq j' \\ &Cov\ (y_{ij}, y_{i'j'}) = 0 \qquad i \neq i' \end{aligned}$$

Note that the observations within a specific factor level all have the same covariance, because before the experiment is conducted, we expect the observations at that factor level to be similar because they all have the same random component. Once the experiment has been conducted, we can assume that all observations can be assumed to be independent, because the parameter τ_i has been determined and the observations in that treatment differ only because of random error.

[3] The assumption that the $[\tau_i]$ are independent random variables implies that the usual assumption of $\sum_{i=1}^{a} \tau_i = 0$ from the fixed effects model does not apply to the random effects model.

We can express the covariance structure of the observations in the single-factor random effects model through the **covariance matrix** of the observations. To illustrate, suppose that we have $a = 3$ treatments and $n = 2$ replicates. There are $N = 6$ observations, which we can write as a vector

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix}$$

and the 6 × 6 covariance matrix of these observations is

$$Cov(\mathbf{y}) = \begin{bmatrix} \sigma_\tau^2 + \sigma^2 & \sigma_\tau^2 & 0 & 0 & 0 & 0 \\ \sigma_\tau^2 & \sigma_\tau^2 + \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_\tau^2 + \sigma^2 & \sigma_\tau^2 & 0 & 0 \\ 0 & 0 & \sigma_\tau^2 & \sigma_\tau^2 + \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_\tau^2 + \sigma^2 & \sigma^2 \\ 0 & 0 & 0 & 0 & \sigma_\tau^2 & \sigma_\tau^2 + \sigma^2 \end{bmatrix}$$

The main diagonals of this matrix are the variances of each individual observation and every off-diagonal element is the covariance of a pair of observations.

3.9.2 Analysis of Variance for the Random Model

The basic ANOVA sum of squares identity

$$SS_T = SS_{\text{Treatments}} + SS_E \tag{3.45}$$

is still valid. That is, we partition the total variability in the observations into a component that measures the variation between treatments ($SS_{\text{Treatments}}$) and a component that measures the variation within treatments (SS_E). Testing hypotheses about individual treatment effects is not very meaningful because they were selected randomly, we are more interested in the **population** of treatments, so we test hypotheses about the variance component σ_τ^2.

$$\begin{aligned} H_0 &: \sigma_\tau^2 = 0 \\ H_1 &: \sigma_\tau^2 > 0 \end{aligned} \tag{3.46}$$

If $\sigma_\tau^2 = 0$, all treatments are identical; but if $\sigma_\tau^2 = 0$, variability exists between treatments. As before, SS_E/σ^2 is distributed as chi-square with $N - a$ degrees of freedom and, under the null hypothesis, $SS_{\text{Treatments}}/\sigma^2$ is distributed as chi-square with $a - 1$ degrees of freedom. Both random variables are independent. Thus, under the null hypothesis $\sigma_\tau^2 = 0$, the ratio

$$F_0 = \frac{\frac{SS_{\text{Treatments}}}{a-1}}{\frac{SS_E}{N-a}} = \frac{MS_{\text{Treatments}}}{MS_E} \tag{3.47}$$

is distributed as F with $a - 1$ and $N - a$ degrees of freedom. However, we need to examine the expected mean squares to fully describe the test procedure.

Consider

$$\begin{aligned} E(MS_{\text{Treatments}}) &= \frac{1}{a-1}E(SS_{\text{Treatments}}) = \frac{1}{a-1}E\left[\sum_{i=1}^{a}\frac{y_{i.}^2}{n} - \frac{y_{..}^2}{N}\right] \\ &= \frac{1}{a-1}E\left[\frac{1}{n}\sum_{i=1}^{a}\left(\sum_{j=1}^{n}\mu + \tau_i + \epsilon_{ij}\right)^2 - \frac{1}{N}\left(\sum_{i=1}^{a}\sum_{j=1}^{n}\mu + \tau_i + \epsilon_{ij}\right)^2\right] \end{aligned}$$

When squaring and taking expectation of the quantities in brackets, we see that terms involving τ_i^2 are replaced by σ_τ^2 as $E(\tau_i) = 0$. Also, terms involving $\epsilon_{i.}^2$, $\epsilon_{..}^2$, and $\sum_{i=1}^{a} \sum_{j=1}^{n} \tau_i^2$ are replaced by $n\sigma^2$, $an\sigma^2$, and an^2, respectively. Furthermore, all cross-product terms involving τ_i and ϵ_{ij} have zero expectation. This leads to

$$E(MS_{\text{Treatments}}) = \frac{1}{a-1}[N\mu^2 + N\sigma_\tau^2 + a\sigma^2 - N\mu^2 - n\sigma_\tau^2 - \sigma^2]$$

or

$$E(MS_{\text{Treatments}}) = \sigma^2 + n\sigma_\tau^2 \tag{3.48}$$

Similarly, we may show that

$$E(MS_E) = \sigma^2 \tag{3.49}$$

From the expected mean squares, we see that under H_0 both the numerator and denominator of the test statistic (Equation 3.47) are unbiased estimators of σ^2, whereas under H_1 the expected value of the numerator is greater than the expected value of the denominator. Therefore, we should reject H_0 for values of F_0 that are too large. This implies an upper-tail, one-tail critical region, so we reject H_0 if $F_0 > F_{\alpha,a-1,N-a}$.

The computational procedure and ANOVA for the random effects model are identical to those for the fixed effects case. The conclusions, however, are quite different because they apply to the entire population of treatments.

3.9.3 Estimating the Model Parameters

We are usually interested in estimating the variance components (σ^2 and σ_τ^2) in the model. One very simple procedure that we can use to estimate σ^2 and σ_τ^2 is called the **analysis of variance method** because it makes use of the lines in the analysis of variance table. The procedure consists of equating the expected mean squares to their observed values in the ANOVA table and solving for the variance components. In equating observed and expected mean squares in the single-factor random effects model, we obtain

$$MS_{\text{Treatments}} = \sigma^2 + n\sigma_\tau^2$$

and

$$MS_E = \sigma^2$$

Therefore, the estimators of the variance components are

$$\hat{\sigma}^2 = MS_E \tag{3.50}$$

and

$$\hat{\sigma}_\tau^2 = \frac{MS_{\text{Treatments}} - MS_E}{n} \tag{3.51}$$

For unequal sample sizes, replace n in Equation 3.51 by

$$n_0 = \frac{1}{a-1}\left[\sum_{i=1}^{a} n_i - \frac{\sum_{i=1}^{a} n_i^2}{\sum_{i=1}^{a} n_i}\right] \tag{3.52}$$

The analysis of variance method of variance component estimation is a **method of moments procedure.** It does not require the normality assumption. It does yield estimators of σ^2 and σ_τ^2 that are best quadratic unbiased (i.e., of all unbiased quadratic functions of the observations, these estimators have minimum variance). There is a different method based on maximum likelihood that can be used to estimate the variance components that will be introduced later.

Occasionally, the analysis of variance method produces a negative estimate of a variance component. Clearly, variance components are by definition nonnegative, so a negative estimate of a variance component is viewed with some concern. One course of action is to accept the estimate and use it as evidence that the true value of the variance component is zero, assuming that sampling variation led to the negative estimate. This has intuitive appeal, but it suffers from some theoretical difficulties. For instance, using zero in place of the negative estimate can disturb the statistical properties of other estimates. Another alternative is to reestimate the negative variance component using a method that always yields nonnegative estimates. Still another alternative is to consider the negative estimate as evidence that the assumed linear model is incorrect and reexamine the problem. Comprehensive treatment of variance component estimation is given by Searle (1971a, 1971b), Searle, Casella, and McCullogh (1992), and Burdick and Graybill (1992).

EXAMPLE 3.10

A textile company weaves a fabric on a large number of looms. It would like the looms to be homogeneous so that it obtains a fabric of uniform strength. The process engineer suspects that, in addition to the usual variation in strength within samples of fabric from the same loom, there may also be significant variations in strength between looms. To investigate this, she selects four looms at random and makes four strength determinations on the fabric manufactured on each loom. This experiment is run in random order, and the data obtained are shown in Table 3.18. The ANOVA is conducted and is shown in Table 3.19. from the ANOVA, we conclude that the looms in the plant differ significantly.

The variance components are estimated by $\hat{\sigma}^2 = 1.90$ and

$$\hat{\sigma}_\tau^2 = \frac{29.73 - 1.90}{4} = 6.96$$

Therefore, the variance of any observation on strength is estimated by

$$\hat{\sigma}_y = \hat{\sigma}^2 + \hat{\sigma}_\tau^2 = 1.90 + 6.96 = 8.86.$$

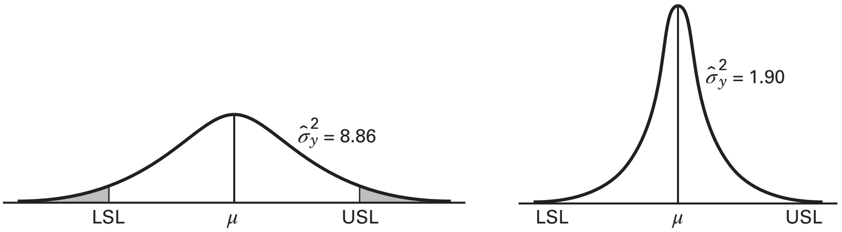
Most of this variability is attributable to differences *between* looms.

■ **TABLE 3.18**
Strength Data for Example 3.10

	Observations				
Looms	**1**	**2**	**3**	**4**	$y_{i.}$
1	98	97	99	96	390
2	91	90	93	92	366
3	96	95	97	95	383
4	95	96	99	98	388
					$1527 = y_{..}$

■ **TABLE 3.19**
Analysis of Variance for the Strength Data

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	*P*-Value
Looms	89.19	3	29.73	15.68	<0.001
Error	22.75	12	1.90		
Total	111.94	15			



(a) Variability of process output

(b) Variability of process output if $\sigma_\tau^2 = 0$

■ **FIGURE 3.19 Process output in the fiber strength problem**

This example illustrates an important use of variance components—isolating different sources of variability that affect a product or system. The problem of product variability frequently arises in quality assurance, and it is often difficult to isolate the sources of variability. For example, this study may have been motivated by an observation that there is too much variability in the strength of the fabric, as illustrated in Figure 3.19a. This graph displays the process output (fiber strength) modeled as a normal distribution with variance $\hat{\sigma}_y^2 = 8.86$. (This is the estimate of the variance of any observation on strength from Example 3.10.) Upper and lower specifications on strength are also shown in Figure 3.19a, and it is relatively easy to see that a fairly large proportion of the process output is outside the specifications (the shaded tail areas in Figure 3.19a). The process engineer has asked why so much fabric is defective and must be scrapped, reworked, or downgraded to a lower quality product. The answer is that most of the product strength variability is the result of differences between looms. Different loom performance could be the result of faulty setup, poor maintenance, ineffective supervision, poorly trained operators, defective input fiber, and so forth.

The process engineer must now try to isolate the specific causes of the differences in loom performance. If she could identify and eliminate these sources of between-loom variability, the variance of the process output could be reduced considerably, perhaps to as low as $\hat{\sigma}_y^2 = 1.90$, the estimate of the within-loom (error) variance component in Example 3.10. Figure 3.19b shows a normal distribution of fiber strength with $\hat{\sigma}_y^2 = 1.90$. Note that the proportion of defective product in the output has been dramatically reduced. Although it is unlikely that *all* of the between-loom variability can be eliminated, it is clear that a significant reduction in this variance component would greatly increase the quality of the fiber produced.

We may easily find a confidence interval for the variance component σ^2. If the observations are normally and independently distributed, then $(N-a)MS_E/\sigma^2$ is distributed as χ^2_{N-a}. Thus,

$$P\left[\chi^2_{1-(\alpha/2),N-a} \leq \frac{(N-a)MS_E}{\sigma^2} \leq \chi^2_{\alpha/2,N-a}\right] = 1-\alpha$$

and a $100(1-\alpha)$ percent confidence interval for σ^2 is

$$\frac{(N-a)MS_E}{\chi^2_{\alpha/2,N-a}} \leq \sigma^2 \leq \frac{(N-a)MS_E}{\chi^2_{1-(\alpha/2),N-a}} \tag{3.53}$$

Since $MS_E = 190, N = 16, a = 4, \chi^2_{0.025,12} = 23{,}3367$ and $\chi^2_{0.975,12} = 4.4038$, the 95% CI on σ^2 is $0.9770 \leq \sigma^2 \leq 5.1775$.

Now consider the variance component σ_τ^2. The point estimator of σ_τ^2 is

$$\hat{\sigma}_\tau^2 = \frac{MS_{\text{Treatments}} - MS_E}{n}$$

The random variable $(a-1)MS_{\text{Treatments}}/(\sigma^2 + n\sigma_\tau^2)$ is distributed as χ^2_{a-1}, and $(N-a)MS_E/\sigma^2$ is distributed as χ^2_{N-a}. Thus, the probability distribution of $\hat{\sigma}_\tau^2$ is a linear combination of two chi-square random variables, say

$$u_1\chi^2_{a-1} - u_2\chi^2_{N-a}$$

where

$$u_1 = \frac{\sigma^2 + n\sigma_\tau^2}{n(a-1)} \quad \text{and} \quad u_2 = \frac{\sigma^2}{n(N-a)}$$

Unfortunately, a closed-form expression for the distribution of this linear combination of chi-square random variables cannot be obtained. Thus, an exact confidence interval for σ_τ^2 cannot be constructed. Approximate procedures are given in Graybill (1961) and Searle (1971a). Also see Section 13.6 of Chapter 13.

It is easy to find an exact expression for a confidence interval on the ratio $\sigma_\tau^2/(\sigma_\tau^2 + \sigma^2)$. This ratio is called the **intraclass correlation coefficient**, and it reflects the *proportion* of the variance of an observation [recall that $V(y_{ij}) = \sigma_\tau^2 + \sigma^2$] that is the result of differences between treatments. To develop this confidence interval for the case of a balanced design, note that $MS_{\text{Treatments}}$ and MS_E are independent random variables and, furthermore, it can be shown that

$$\frac{MS_{\text{Treatments}}/(n\sigma_\tau^2 + \sigma^2)}{MS_E/\sigma^2} \sim F_{a-1,N-a}$$

Thus,

$$\mathrm{P}\left(F_{1-\alpha/2,a-1,N-a} \leq \frac{MS_{\text{Treatments}}}{MS_E}\,\frac{\sigma^2}{n\sigma_\tau^2 + \sigma^2} \leq F_{\alpha/2,a-1,N-a}\right) = 1 - \alpha \tag{3.54}$$

By rearranging Equation 3.54, we may obtain the following:

$$P\left(L \leq \frac{\sigma_\tau^2}{\sigma^2} \leq U\right) = 1 - \alpha \tag{3.55}$$

where

$$L = \frac{1}{n}\left(\frac{MS_{\text{Treatments}}}{MS_E}\,\frac{1}{F_{\alpha/2,a-1,N-a}} - 1\right) \tag{3.56a}$$

and

$$U = \frac{1}{n}\left(\frac{MS_{\text{Treatments}}}{MS_E}\,\frac{1}{F_{1-\alpha/2,a-1,N-a}} - 1\right) \tag{3.56b}$$

Note that L and U are $100(1-\alpha)$ percent lower and upper confidence limits, respectively, for the ratio σ_τ^2/σ^2. Therefore, a $100(1-\alpha)$ percent confidence interval for $\sigma_\tau^2/(\sigma_\tau^2 + \sigma^2)$ is

$$\frac{L}{1+L} \leq \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma^2} \leq \frac{U}{1+U} \tag{3.57}$$

To illustrate this procedure, we find a 95 percent confidence interval on $\sigma_\tau^2/(\sigma_\tau^2 + \sigma^2)$ for the strength data in Example 3.10. Recall that $MS_{\text{Treatments}} = 29.73, MS_E = 1.90, a = 4, n = 4, F_{0.025,3,12} = 4.47$, and $F_{0.975,3,12} = 1/F_{0.025,12,3} = 1/14.34 = 0.070$. Therefore, from Equation 3.56a and *b*,

$$L = \frac{1}{4}\left[\left(\frac{29.73}{1.90}\right)\left(\frac{1}{4.47}\right) - 1\right] = 0.625$$

$$U = \frac{1}{4}\left[\left(\frac{29.73}{1.90}\right)\left(\frac{1}{0.070}\right) - 1\right] = 55.633$$

and from Equation 3.57, the 95 percent confidence interval on $\sigma_\tau^2/(\sigma_\tau^2 + \sigma^2)$ is

$$\frac{0.625}{1.625} \leq \frac{\sigma^2}{\sigma_\tau^2 + \sigma^2} \leq \frac{55.633}{56.633}$$

or

$$0.38 \leq \frac{\sigma^2}{\sigma_\tau^2 + \sigma^2} \leq 0.98$$

We conclude that variability between looms accounts for between 38 and 98 percent of the variability in the observed strength of the fabric produced. This confidence interval is relatively wide because of the small number of looms used in the experiment. Clearly, however, the variability between looms (σ_τ^2) is not negligible.

Estimation of the Overall Mean μ. In many random effects experiments, the experimenter is interested in estimating the overall mean μ. From the basic model assumptions, it is easy to see that the expected value of any observation is just the overall mean. Consequently, an unbiased estimator of the overall mean is

$$\hat{\mu} = \bar{y}_{..}$$

So for Example 3.10 the estimate of the overall mean strength is

$$\hat{\mu} = \bar{y}_{..} = \frac{y_{..}}{N} = \frac{1527}{16} = 95.44$$

It is also possible to find a 100(1 − α)% confidence interval on the overall mean. The variance of $\bar{y}$ is

$$V(\bar{y}_{..}) = V\left(\frac{\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}}{an}\right) = \frac{n\sigma_\tau^2 + \sigma^2}{an}$$

The numerator of this ratio is estimated by the treatment mean square, so an unbiased estimator of $V(\bar{y})$ is

$$\hat{V}(\bar{y}_{..}) = \frac{MS_{\text{Treatments}}}{an}$$

Therefore, the 100(1 − α)% CI on the overall mean is

$$\bar{y}_{..} - t_{\alpha/2,a(n-1)}\sqrt{\frac{MS_{\text{Treatments}}}{an}} \leq \mu \leq \bar{y}_{..} + t_{\alpha/2,a(n-1)}\sqrt{\frac{MS_{\text{Treatments}}}{an}} \tag{3.58}$$

To find a 95% CI on the overall mean in the fabric strength experiment from Example 3.10, we need $MS_{\text{Treatments}} =$ 29.73 and $t_{0.025,12} = 2.18$. The CI is computed from Equation 3.58 as follows:

$$\bar{y}_{..} - t_{\alpha/2,a(n-1)}\sqrt{\frac{MS_{\text{Treatments}}}{an}} \leq \mu \leq \bar{y}_{..} + t_{\alpha/2,a(n-1)}\sqrt{\frac{MS_{\text{Treatments}}}{an}}$$

$$95.44 - 2.18\sqrt{\frac{29.73}{16}} \leq \mu \leq 95.44 + 2.18\sqrt{\frac{29.73}{16}}$$

$$92.47 \leq \mu \leq 98.41$$

So, at 95 percent confidence the mean strength of the fabric produced by the looms in this facility is between 92.47 and 98.41. This is a relatively wide confidence interval because a small number of looms were sampled and there is a large difference between looms as reflected by the large portion of total variability that is accounted for by the differences between looms.

Maximum Likelihood Estimation of the Variance Components. Earlier in this section we presented the analysis of variance method of variance component estimation. This method is relatively straightforward to apply and makes use of familiar quantities—the mean squares in the analysis of variance table. However, the method has some disadvantages. As we pointed out previously, it is a **method of moments estimator**, a technique that mathematical statisticians generally do not prefer to use for parameter estimation because it often results in parameter estimates that do not have good statistical properties. One obvious problem is that it does not always lead to an easy way to construct confidence intervals on the variance components of interest. For example, in the single-factor random model, there is not a simple way to construct confidence intervals on σ_τ^2, which is certainly a parameter of primary interest to the experimenter. The preferred parameter estimation technique is called the **method of maximum likelihood.** The implementation of this method can be somewhat involved, particularly for an experimental design model, but it has been incorporated in some modern computer software packages that support designed experiments, including JMP.

A complete presentation of the method of maximum likelihood is beyond the scope of this book, but the general idea can be illustrated very easily. Suppose that x is a random variable with probability distribution $f(x, \theta)$, where θ is an unknown parameter. Let $x_1, x_2, \ldots, x_n$ be a random sample of n observations. The joint probability distribution of the sample is $\prod_{i=1}^{n} f(x_i, \theta)$. The **likelihood function** is just this joint probability distribution with the sample observations consider fixed and the parameter θ unknown. Note that the likelihood function, say

$$L(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

is now a function of only the unknown parameter θ. The **maximum likelihood estimator** of θ is the value of θ that maximizes the likelihood function $L(x_1, x_2, \ldots, x_n; \theta)$. To illustrate how this applies to an experimental design model with random effects, let $\mathbf{y}$ be the $an \times 1$ vector of observations for a single-factor random effects model with a treatments and n replicates and let $\sum$ be the $an \times an$ covariance matrix of the observations. Refer to Section 3.9.1 where we developed this covariance matrix for the special case where $a = 3$ and $n = 2$. The likelihood function is

$$L(x_{11}, x_{12}, \ldots, x_{a,n}; \mu, \sigma_\tau^2, \sigma^2) = \frac{1}{(2\pi)^{N/2} \left[\sum\right]^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{j}_N\mu)' \sum^{-1} (\mathbf{y} - \mathbf{j}_N\mu)\right]$$

where $N = an$ is the total number of observations, $\mathbf{j}_N$ is an $N \times 1$ vector of 1s, and μ is the overall mean in the model. The maximum likelihood estimates of the parameters μ, σ_τ^2, and σ^2 are the values of these quantities that maximize the likelihood function.

Maximum likelihood estimators (MLEs) have some very useful properties. For large samples, they are unbiased, and they have a normal distribution. Furthermore, the inverse of the matrix of second derivatives of the likelihood function (multiplied by -1) is the covariance matrix of the MLEs. This makes it relatively easy to obtain approximate confidence intervals on the MLEs.

The standard variant of maximum likelihood estimation that is used for estimating variance components is known as the **residual maximum likelihood (REML) method.** It is popular because it produces unbiased estimators and like all MLEs, it is easy to find CIs. The basic characteristic of REML is that it takes the location parameters in the model into account when estimating the random effects. As a simple example, suppose that we want to estimate the mean and variance of a normal distribution using the method of maximum likelihood. It is easy to show that the MLEs are

$$\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n} = \overline{y}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n}$$

Notice that the MLE $\hat{\sigma}^2$ is not the familiar sample standard deviation. It does not take the estimation of the location parameter μ into account. The REML estimator would be

$$S^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n - 1}$$

The REML estimator is unbiased.

■ **TABLE 3.20**
JMP Output for the Loom Experiment in Example 3.10

Response Y

Summary of Fit

RSquare	0.793521
RSquare Adj	0.793521
Root Mean Square Error	1.376893
Mean of Response	95.4375
Observations (or Sum Wgts)	16

Parameter Estimates

Term	Estimate	Std Error	DFDen	t Ratio	Prob > |t|
Intercept	95.4375	1.363111	3	70.01	< .0001*

REML Variance Component Estimates

Random Effect	Var Ratio	Var Component	Std Error	95% Lower	95% Upper	Pct of Total
X1	3.6703297	6.9583333	6.0715247	−4.941636	18.858303	78.588
Residual		1.8958333	0.7739707	0.9748608	5.1660065	21.412
Total		8.8541667				100.000

Covariance Matrix of Variance Component Estimates

Random Effect	X1	Residual
X1	36.863412	−0.149758
Residual	−0.149758	0.5990307

To illustrate the REML method, Table 3.20 presents the JMP output for the loom experiment in Example 3.10. The REML estimates of the model parameters μ, σ_τ^2, and σ^2 are shown in the output. Note that the REML estimates of the variance components are identical to those found earlier by the ANOVA method. These two methods will agree for balanced designs. However, the REML output also contains the covariance matrix of the variance components. The square roots of the main diagonal elements of this matrix are the standard errors of the variance components. If $\hat{\theta}$ is the MLE of θ and $\hat{\sigma}(\hat{\theta})$ is its estimated standard error, then the approximate $100(1-\alpha)$ percent confidence interval on θ is

$$\hat{\theta} - Z_{\alpha/2}\hat{\sigma}(\hat{\theta}) \leq \theta \leq \hat{\theta} + Z_{\alpha/2}\hat{\sigma}(\hat{\theta})$$

JMP uses this approach to find the approximate CIs of σ_τ^2 and σ^2 shown in the output. The 95 percent CI from REML for σ^2 is very similar to the chi-square-based interval computed earlier in Section 3.9.

3.10 The Regression Approach to the Analysis of Variance

We have given an intuitive or heuristic development of the analysis of variance. However, it is possible to give a more formal development. The method will be useful later in understanding the basis for the statistical analysis of more complex designs. Called the **general regression significance test**, the procedure essentially consists of finding the reduction in the total sum of squares for fitting the model with all parameters included and the reduction in sum of squares when the model is restricted to the null hypotheses. The difference between these two sums of squares is the treatment sum of squares with which a test of the null hypothesis can be conducted. The procedure requires the least squares estimators of the parameters in the analysis of variance model. We have given these parameter estimates previously (in Section 3.3.3); however, we now give a formal development.

3.10.1 Least Squares Estimation of the Model Parameters

We now develop estimators for the parameter in the single-factor ANOVA fixed-effects model

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

using the method of least squares. To find the least squares estimators of μ and τ_i, we first form the sum of squares of the errors

$$L = \sum_{i=1}^{a} \sum_{j=1}^{n} \epsilon_{ij}^2 = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \mu - \tau_i)^2 \tag{3.59}$$

and then choose values of μ and τ_i, say $\hat{\mu}$ and $\hat{\tau}_i$, that minimize L. The appropriate values would be the solutions to the $a + 1$ simultaneous equations

$$\begin{aligned} \left.\frac{\partial L}{\partial \mu}\right|_{\hat{\mu},\hat{\tau}_i} &= 0 \\ \left.\frac{\partial L}{\partial \tau_i}\right|_{\hat{\mu},\hat{\tau}_i} &= 0 \quad i = 1, 2, \ldots, a \end{aligned}$$

Differentiating Equation 3.59 with respect to μ and τ_i and equating to zero, we obtain

$$-2 \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \hat{\mu} - \hat{\tau}_i) = 0$$

and

$$-2 \sum_{j=1}^{n} (y_{ij} + \hat{\mu} - \hat{\tau}_i) = 0 \quad i = 1, 2, \ldots, a$$

which, after simplification, yield

$$\begin{array}{llllll} N\hat{\mu} & + n\hat{\tau}_1 & + n\hat{\tau}_2 & + \cdots & + n\hat{\tau}_a & = y_{..} \\ n\hat{\mu} & + n\hat{\tau}_1 & & & & = y_{1.} \\ n\hat{\mu} & & + n\hat{\tau}_2 & & & = y_{2.} \\ & & \vdots & & & \vdots \\ n\hat{\mu} & & & & + n\hat{\tau}_a & = y_{a.} \end{array} \tag{3.60}$$

The $a + 1$ equations (Equation 3.60) in $a + 1$ unknowns are called the **least squares normal equations.** Notice that if we add the last a normal equations, we obtain the first normal equation. Therefore, the normal equations are not linearly independent, and no unique solution for $\mu, \tau_i, \ldots, \tau_a$ exists. This has happened because the effects model is **overparameterized.** This difficulty can be overcome by several methods. Because we have defined the treatment effects as deviations from the overall mean, it seems reasonable to apply the **constraint**

$$\sum_{i=1}^{a} \hat{\tau}_i = 0 \tag{3.61}$$

Using this constraint, we obtain as the solution to the normal equations

$$\begin{aligned} \hat{\mu} &= \overline{y}_{..} \\ \hat{\tau}_i &= \overline{y}_{i.} - \overline{y}_{..} \quad i = 1, 2, \ldots, a \end{aligned} \tag{3.62}$$

This solution is obviously not unique and depends on the constraint (Equation 3.61) that we have chosen. At first this may seem unfortunate because two different experimenters could analyze the same data and obtain different results if they apply different constraints. However, certain **functions** of the model parameters *are* uniquely estimated, regardless of the constraint. Some examples are $\tau_i - \tau_j$, which would be estimated by $\hat{\tau}_i - \hat{\tau}_j = \overline{y}_{i.} - \overline{y}_{j.}$, and the *i*th treatment mean $\mu_i = \mu + \tau_i$, which would be estimated by $\hat{\mu}_i = \hat{\mu} + \hat{\tau}_i = \overline{y}_{i.}$.

Because we are usually interested in differences among the treatment effects rather than their actual values, it causes no concern that the τ_i cannot be uniquely estimated. In general, any function of the model parameters that is a linear combination of the left-hand side of the normal equations (Equations 3.60) can be uniquely estimated. Functions that are uniquely estimated regardless of which constraint is used are called **estimable functions.** For more information, see the **supplemental material** for this chapter. We are now ready to use these parameter estimates in a general development of the analysis of variance.

3.10.2 The General Regression Significance Test

A fundamental part of this procedure is writing the normal equations for the model. These equations may always be obtained by forming the least squares function and differentiating it with respect to each unknown parameter, as we did in Section 3.9.1. However, an easier method is available. The following rules allow the normal equations for *any* experimental design model to be written directly:

RULE 1. There is one normal equation for each parameter in the model to be estimated.

RULE 2. The right-hand side of any normal equation is just the sum of all observations that contain the parameter associated with that particular normal equation.

To illustrate this rule, consider the single-factor model. The first normal equation is for the parameter μ; therefore, the right-hand side is $y_{..}$ because *all* observations contain μ.

RULE 3. The left-hand side of any normal equation is the sum of all model parameters, where each parameter is multiplied by the number of times it appears in the total on the right-hand side. The parameters are written with a circumflex (ˆ) to indicate that they are **estimators** and not the true parameter values.

For example, consider the first normal equation in a single-factor experiment. According to the aforementioned rules, it would be

$$N\hat{\mu} + n\hat{\tau}_1 + n\hat{\tau}_2 + \cdots + n\hat{\tau}_a = y_{..}$$

because μ appears in all N observations, τ_1 appears only in the n observations taken under the first treatment, τ_2 appears only in the n observations taken under the second treatment, and so on. From Equation 3.60, we verify that the equation shown above is correct. The second normal equation would correspond to τ_1 and is

$$n\hat{\mu} + n\hat{\tau}_1 = y_{1.}$$

because only the observations in the first treatment contain τ_1 (this gives $y_{1.}$ as the right-hand side), μ and τ_1 appear exactly n times in $y_{1.}$, and all other τ_i appear zero times. In general, the left-hand side of any normal equation is the expected value of the right-hand side.

Now, consider finding the reduction in the sum of squares by fitting a particular model to the data. By fitting a model to the data, we "explain" some of the variability; that is, we reduce the unexplained variability by some amount. The reduction in the unexplained variability is always the sum of the parameter estimates, each multiplied by the right-hand side of the normal equation that corresponds to that parameter. For example, in a single-factor experiment, the reduction due to fitting the **full model** $y_{ij} = \mu + \tau_i + \epsilon_{ij}$ is

$$\begin{aligned} R(\mu, \tau) &= \hat{\mu}y_{..} + \hat{\tau}_1 y_{1.} + \hat{\tau}_2 y_{2.} + \cdots + \hat{\tau}_a y_{a.} \\ &= \hat{\mu}y_{..} + \sum_{i=1}^{a} \hat{\tau}_i y_{i.} \end{aligned} \tag{3.63}$$

The notation $R(\mu, \tau)$ means that reduction in the sum of squares from fitting the model containing μ and $\{\tau_i\}$. $R(\mu, \tau)$ is also sometimes called the "regression" sum of squares for the full model $y_{ij} = \mu + \tau_i + \epsilon_{ij}$. The number of degrees of freedom associated with a reduction in the sum of squares, such as $R(\mu, \tau)$, is always equal to the number of linearly independent normal equations. The remaining variability unaccounted for by the model is found from

$$SS_E = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}^2 - R(\mu, \tau) \tag{3.64}$$

This quantity is used in the denominator of the test statistic for $H_0 : \tau_1 = \tau_2 = \ldots = \tau_a = 0$.

We now illustrate the general regression significance test for a single-factor experiment and show that it yields the usual one-way analysis of variance. The model is $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, and the normal equations are found from the above rules as

$$\begin{array}{llllll} N\hat{\mu} & + n\hat{\tau}_1 & + n\hat{\tau}_2 & + \cdots & + n\hat{\tau}_a & = y_{..} \\ n\hat{\mu} & + n\hat{\tau}_1 & & & & = y_{1.} \\ n\hat{\mu} & & + n\hat{\tau}_2 & & & = y_{2.} \\ & & \vdots & & & \vdots \\ n\hat{\mu} & & & & + n\hat{\tau}_a & = y_{a.} \end{array}$$

Compare these normal equations with those obtained in Equation 3.60.

Applying the constraint $\sum_{i=1}^{a} \hat{\tau}_i = 0$, we find that the estimators for μ and τ_i are

$$\hat{\mu} = \overline{y}_{..} \qquad \hat{\tau}_i = \overline{y}_{i.} - \overline{y}_{..}\ i = 1, 2, \ldots, a$$

The reduction in the sum of squares due to fitting this full model is found from Equation 3.48 as

$$\begin{aligned} R(\mu, \tau) &= \hat{\mu} y_{..} + \sum_{i=1}^{a} \hat{\tau}_i y_{i.} \\ &= (\overline{y}_{..}) y_{..} + \sum_{i=1}^{a} (\overline{y}_{i.} - \overline{y}_{..}) y_{i.} \\ &= \frac{y_{..}^2}{N} + \sum_{i=1}^{a} \overline{y}_{i.} y_{i.} - \overline{y}_{..} \sum_{i=1}^{a} y_{i.} \\ &= \sum_{i=1}^{a} \frac{y_{i.}^2}{n} \end{aligned}$$

which has a degrees of freedom because there are a linearly independent normal equations. The error sum of squares is, from Equation 3.64,

$$\begin{aligned} SS_E &= \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}^2 - R(\mu, \tau) \\ &= \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}^2 - \sum_{i=1}^{a} \frac{y_{i.}^2}{n} \end{aligned}$$

and has $N - a$ degrees of freedom.

To find the sum of squares resulting from the treatment effects (the $\{\tau_i\}$), we consider a **reduced model**; that is, the model to be restricted to the null hypothesis ($\tau_i = 0$ for all i). The reduced model is $y_{ij} = \mu + \epsilon_{ij}$. There is only one normal equation for this model:

$$N\hat{\mu} = y_{..}$$

and the estimator of μ is $\hat{\mu} = \overline{y}_{..}$. Thus, the reduction in the sum of squares that results from fitting the reduced model containing only μ is

$$R(\mu) = (\overline{y}_{..})(y_{..}) = \frac{y_{..}^2}{N}$$

Because there is only one normal equation for this reduced model, $R(\mu)$ has one degree of freedom. The sum of squares due to the $\{\tau_i\}$, given that μ is already in the model, is the difference between $R(\mu, \tau)$ and $R(\mu)$, which is

$$\begin{aligned} R(\tau|\mu) &= R(\mu, \tau) - R(\mu) \\ &= R(\text{Full Model}) - R(\text{Reduced Model}) \\ &= \frac{1}{n}\sum_{i=1}^{a} y_{i.}^2 - \frac{y_{..}^2}{N} \end{aligned}$$

with $a - 1$ degrees of freedom, which we recognize from Equation 3.9 as $SS_{\text{Treatments}}$. Making the usual normality assumption, we obtain the appropriate statistic for testing $H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$

$$F_0 = \frac{R(\tau|\mu)(/(a-1)}{\left[\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - R(\mu, \tau)\right]/(N-a)}$$

which is distributed as $F_{a-1,N-a}$ under the null hypothesis. This is, of course, the test statistic for the single-factor analysis of variance.

3.11 Nonparametric Methods in the Analysis of Variance

3.11.1 The Kruskal–Wallis Test

In situations where the normality assumption is unjustified, the experimenter may wish to use an alternative procedure to the F-test analysis of variance that does not depend on this assumption. Such a procedure has been developed by Kruskal and Wallis (1952). The Kruskal–Wallis test is used to test the null hypothesis that the a treatments are identical against the alternative hypothesis that some of the treatments generate observations that are larger than others. Because the procedure is designed to be sensitive for testing differences in means, it is sometimes convenient to think of the Kruskal–Wallis test as a test for equality of treatment means. The Kruskal–Wallis test is a **nonparametric alternative** to the usual analysis of variance.

To perform a Kruskal–Wallis test, first rank the observations y_{ij} in ascending order and replace each observation by its rank, say R_{ij}, with the smallest observation having rank 1. In the case of ties (observations having the same value), assign the average rank to each of the tied observations. Let $R_{i.}$ be the sum of the ranks in the ith treatment. The test statistic is

$$H = \frac{1}{S^2}\left[\sum_{i=1}^{a} \frac{R_{i.}^2}{n_i} - \frac{N(N+1)^2}{4}\right] \tag{3.65}$$

where n_i is the number of observations in the ith treatment, N is the total number of observations, and

$$S^2 = \frac{1}{N-1}\left[\sum_{i=1}^{a}\sum_{j=1}^{n_i} R_{ij}^2 - \frac{N(N+1)^2}{4}\right] \tag{3.66}$$

Note that S^2 is just the variance of the ranks. If there are no ties, $S^2 = N(N+1)/12$ and the test statistic simplifies to

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{a} \frac{R_{i.}^2}{n_i} - 3(N+1) \tag{3.67}$$

When the number of ties is moderate, there will be little difference between Equations 3.66 and 3.67, and the simpler form (Equation 3.67) may be used. If the n_i are reasonably large, say $n_i \geq 5, H$ is distributed approximately as χ_{a-1}^2 under the null hypothesis. Therefore, if

$$H > \chi_{\alpha,a-1}^2$$

the null hypothesis is rejected. The P-value approach could also be used.

EXAMPLE 3.11

The data from Example 3.1 and their corresponding ranks are shown in Table 3.21. There are ties, so we use Equation 3.65 as the test statistic. From Equation 3.65

$$S^2 = \frac{1}{19}\left[2869.50 - \frac{20(21)^2}{4}\right] = 34.97$$

and the test statistic is

$$\begin{aligned} H &= \frac{1}{S^2}\left[\sum_{i=1}^{a} \frac{R_{i.}^2}{n_i} - \frac{N(N+1)^2}{4}\right] \\ &= \frac{1}{34.97}[2796.30 - 2205] \\ &= 16.91 \end{aligned}$$

■ **TABLE 3.21**
Data and Ranks for the Plasma Etching Experiment in Example 3.1

Power							
160		**180**		**200**		**220**	
y_{1j}	R_{1j}	y_{2j}	R_{2j}	y_{3j}	R_{3j}	y_{4j}	R_{4j}
575	6	565	4	600	10	725	20
542	3	593	9	651	15	700	17
530	1	590	8	610	11.5	715	19
539	2	579	7	637	14	685	16
570	5	610	11.5	629	13	710	18
$R_{i.}$	17		39.5		63.5		90

Because $H > \chi^2_{0.01,3} = 11.34$, we would reject the null hypothesis and conclude that the treatments differ. (The P-value for $H = 16.91$ is $P = 7.38 \times 10^{-4}$.) This is the same conclusion as given by the usual analysis of variance F-test.

3.11.2 General Comments on the Rank Transformation

The procedure used in the previous section of replacing the observations by their ranks is called the **rank transformation.** It is a very powerful and widely useful technique. If we were to apply the ordinary F-test to the ranks rather than to the original data, we would obtain

$$F_0 = \frac{H/(a-1)}{(N-1-H)/(N-a)} \tag{3.68}$$

as the test statistic [see Conover (1980), p. 337]. Note that as the Kruskal–Wallis statistic H increases or decreases, F_0 also increases or decreases, so the Kruskal–Wallis test is equivalent to applying the usual analysis of variance to the ranks.

The rank transformation has wide applicability in experimental design problems for which no nonparametric alternative to the analysis of variance exists. This includes many of the designs in subsequent chapters of this book. If the data are ranked and the ordinary F-test is applied, an approximate procedure that has good statistical properties results [see Conover and Iman (1976, 1981)]. When we are concerned about the normality assumption or the effect of outliers or "wild" values, we recommend that the usual analysis of variance be performed on both the original data and the ranks. When both procedures give similar results, the analysis of variance assumptions are probably satisfied reasonably well, and the standard analysis is satisfactory. When the two procedures differ, the rank transformation should be preferred because it is less likely to be distorted by nonnormality and unusual observations. In such cases, the experimenter may want to investigate the use of transformations for nonnormality and examine the data and the experimental procedure to determine whether outliers are present and why they have occurred.

3.12 Problems

3.1 An experimenter has conducted a single-factor experiment with four levels of the factor, and each factor level has been replicated six times. The computed value of the F-statistic is $F_0 = 3.26$. Find bounds on the P-value.

3.2 An experimenter has conducted a single-factor experiment with six levels of the factor, and each factor level has been replicated three times. The computed value of the F-statistic is $F_0 = 5.81$. Find bounds on the P-value.

3.3 An experimenter has conducted a single-factor completely randomized design with five levels of the factor and three replicates. The computed value of the F-statistic is 4.87. Find bounds on the P-value.

3.4 An experimenter has conducted a single-factor completely randomized design with three levels of the factor and five replicates. The computed value of the F-statistic is 2.91. Find bounds on the P-value.

3.5 The mean square for error in the ANOVA provides an estimate of

(a) The variance of the random error

(b) The variance of an individual treatment average

(c) The standard deviation of an individual observation

(d) None of the above

3.6 It is always a good idea to check the normality assumption in the ANOVA by applying a test for normality such as the Anderson–Darling test to the residuals.

(a) True

(b) False

3.7 A computer ANOVA output is shown below. Fill in the blanks. You may give bounds on the P-value.

One-way ANOVA

```
Source     DF       SS     MS    F    P
Factor      3    36.15      ?    ?    ?
Error       ?        ?      ?
Total      19   196.04
```

3.8 A computer ANOVA output is shown below. Fill in the blanks. You may give bounds on the P-value.

One-way ANOVA

```
Source    DF        SS        MS    F    P
Factor     ?         ?    246.93    ?    ?
Error     25    186.53         ?
Total     29   1174.24
```

3.9 An article appeared in *The Wall Street Journal* on Tuesday, April 27, 2010, with the title "Eating Chocolate Is Linked to Depression." The article reported on a study funded by the National Heart, Lung and Blood Institute (part of the National Institutes of Health) and conducted by faculty at the University of California, San Diego, and the University of California, Davis. The research was also published in the *Archives of Internal Medicine* (2010, pp. 699–703). The study examined 931 adults who were not taking antidepressants and did not have known cardiovascular disease or diabetes. The group was about 70% men and the average age of the group was reported to be about 58. The participants were asked about chocolate consumption and then screened for depression using a questionnaire. People who score less than 16 on the questionnaire were not considered depressed, while those with scores above 16 and less than or equal to 22 were considered possibly depressed, while those with scores above 22 were considered likely to be depressed. The survey found that people who were not depressed ate an average 5.4 servings of chocolate per month, possibly depressed individuals ate an average of 8.4 servings of chocolate per month, while those individuals who scored above 22 and were likely to be depressed ate the most chocolate, an average of 11.8 servings per month. No differentiation was made between dark and milk chocolate. Other foods were also examined, but no pattern emerged between other foods and depression. Is this study really a designed experiment? Does it establish a cause-and-effect link between chocolate consumption and depression? How would the study have to be conducted to establish such a cause-and effect link?

3.10 An article in *Bioelectromagnetics* ("Electromagnetic Effects on Forearm Disuse Osteopenia: A Randomized, Double-Blind, Sham-Controlled Study," Vol. 32, 2011, pp. 273–282) described a randomized, double-blind, sham-controlled, feasibility and dosing study to determine if a common pulsing electromagnetic field (PEMF) treatment could moderate the substantial osteopenia that occurs after forearm disuse. Subjects were randomized into four groups after a distal radius fracture, or carpal surgery requiring immobilization in a cast. Active or identical sham PEMF transducers were worn on the distal forearm for 1, 2, or 4 h/day for 8 weeks starting after cast removal ("baseline") when bone density continues to decline. Bone mineral density (BMD) and bone geometry were measured in the distal forearm by dual energy X-ray absorptiometry (DXA) and peripheral quantitative computed tomography (pQCT). The data below are the percent losses in BMD measurements on the radius after 16 weeks for patients wearing the active or sham PEMF transducers for 1, 2, or 4 h/day (data were constructed to match the means and standard deviations read from a graph in the paper).

(a) Is there evidence to support a claim that PEMF usage affects BMD loss? If so, analyze the data to determine which specific treatments produce the differences.

(b) Analyze the residuals from this experiment and comment on the underlying assumptions and model adequacy.

Sham	PEMF 1 h/day	PEMF 2 h/day	PEMF 4 h/day
4.51	5.32	4.73	7.03
7.95	6.00	5.81	4.65
4.97	5.12	5.69	6.65
3.00	7.08	3.86	5.49
7.97	5.48	4.06	6.98
2.23	6.52	6.56	4.85
3.95	4.09	8.34	7.26
5.64	6.28	3.01	5.92
9.35	7.77	6.71	5.58
6.52	5.68	6.51	7.91
4.96	8.47	1.70	4.90
6.10	4.58	5.89	4.54
7.19	4.11	6.55	8.18
4.03	5.72	5.34	5.42
2.72	5.91	5.88	6.03
9.19	6.89	7.50	7.04
5.17	6.99	3.28	5.17
5.70	4.98	5.38	7.60
5.85	9.94	7.30	7.90
6.45	6.38	5.46	7.91

3.11 The tensile strength of Portland cement is being studied. Four different mixing techniques can be used economically. A completely randomized experiment was conducted and the following data were collected:

Mixing Technique	Tensile Strength (lb/in^2)			
1	3129	3000	2865	2890
2	3200	3300	2975	3150
3	2800	2900	2985	3050
4	2600	2700	2600	2765

(a) Test the hypothesis that mixing techniques affect the strength of the cement. Use $\alpha = 0.05$.

(b) Construct a graphical display as described in Section 3.5.3 to compare the mean tensile strengths for the four mixing techniques. What are your conclusions?

(c) Use the Fisher LSD method with $\alpha = 0.05$ to make comparisons between pairs of means.

(d) Construct a normal probability plot of the residuals. What conclusion would you draw about the validity of the normality assumption?

(e) Plot the residuals versus the predicted tensile strength. Comment on the plot.

(f) Prepare a scatter plot of the results to aid the interpretation of the results of this experiment.

3.12 **(a)** Rework part (c) of Problem 3.11 using Tukey's test with $\alpha = 0.05$. Do you get the same conclusions from Tukey's test that you did from the graphical procedure and/or the Fisher LSD method?

(b) Explain the difference between the Tukey and Fisher procedures.

3.13 Reconsider the experiment in Problem 3.11. Find a 95 percent confidence interval on the mean tensile strength of the Portland cement produced by each of the four mixing techniques. Also find a 95 percent confidence interval on the difference in means for techniques 1 and 3. Does this aid you in interpreting the results of the experiment?

3.14 A product developer is investigating the tensile strength of a new synthetic fiber that will be used to make cloth for men's shirts. Strength is usually affected by the percentage of cotton used in the blend of materials for the fiber. The engineer conducts a completely randomized experiment with five levels of cotton content and replicates the experiment five times. The data are shown in the following table.

Cotton Weight Percent	Observations				
15	7	7	15	11	9
20	12	17	12	18	18
25	14	19	19	18	18
30	19	25	22	19	23
35	7	10	11	15	11

(a) Is there evidence to support the claim that cotton content affects the mean tensile strength? Use $\alpha = 0.05$.

(b) Use the Fisher LSD method to make comparisons between the pairs of means. What conclusions can you draw?

(c) Analyze the residuals from this experiment and comment on model adequacy.

3.15 Reconsider the experiment described in Problem 3.14. Suppose that 30 percent cotton content is a control. Use Dunnett's test with $\alpha = 0.05$ to compare all of the other means with the control.

3.16 A pharmaceutical manufacturer wants to investigate the bioactivity of a new drug. A completely randomized single-factor experiment was conducted with three dosage levels, and the following results were obtained.

Dosage	Observations			
20 g	24	28	37	30
30 g	37	44	31	35
40 g	42	47	52	38

(a) Is there evidence to indicate that dosage level affects bioactivity? Use $\alpha = 0.05$.

(b) If it is appropriate to do so, make comparisons between the pairs of means. What conclusions can you draw?

(c) Analyze the residuals from this experiment and comment on model adequacy.

3.17 A rental car company wants to investigate whether the type of car rented affects the length of the rental period. An experiment is run for one week at a particular location, and 10 rental contracts are selected at random for each car type. The results are shown in the following table.

Type of Car	Observations									
Subcompact	3	5	3	7	6	5	3	2	1	6
Compact	1	3	4	7	5	6	3	2	1	7
Midsize	4	1	3	5	7	1	2	4	2	7
Full size	3	5	7	5	10	3	4	7	2	7

(a) Is there evidence to support a claim that the type of car rented affects the length of the rental contract? Use $\alpha = 0.05$. If so, which types of cars are responsible for the difference?

(b) Analyze the residuals from this experiment and comment on model adequacy.

(c) Notice that the response variable in this experiment is a count. Should this cause any potential concerns about the validity of the analysis of variance?

3.18 I belong to a golf club in my neighborhood. I divide the year into three golf seasons: summer (June–September), winter (November–March), and shoulder (October, April, and May). I believe that I play my best golf during the summer (because I have more time and the course isn't crowded) and shoulder (because the course isn't crowded) seasons, and my worst golf is during the winter (because when all of the part-year residents show up, the course is crowded, play is slow, and I get frustrated). Data from the last year are shown in the following table.

Season	Observations									
Summer	83	85	85	87	90	88	88	84	91	90
Shoulder	91	87	84	87	85	86	83			
Winter	94	91	87	85	87	91	92	86		

(a) Do the data indicate that my opinion is correct? Use $\alpha = 0.05$.

(b) Analyze the residuals from this experiment and comment on model adequacy.

3.19 A regional opera company has tried three approaches to solicit donations from 24 potential sponsors. The 24 potential sponsors were randomly divided into three groups of eight, and one approach was used for each group. The dollar amounts of the resulting contributions are shown in the following table.

Approach	Contributions (in $)							
1	1000	1500	1200	1800	1600	1100	1000	1250
2	1500	1800	2000	1200	2000	1700	1800	1900
3	900	1000	1200	1500	1200	1550	1000	1100

(a) Do the data indicate that there is a difference in results obtained from the three different approaches? Use $\alpha = 0.05$.

(b) Analyze the residuals from this experiment and comment on model adequacy.

3.20 An experiment was run to determine whether four specific firing temperatures affect the density of a certain type of brick. A completely randomized experiment led to the following data:

Temperature	Density				
100	21.8	21.9	21.7	21.6	21.7
125	21.7	21.4	21.5	21.4	
150	21.9	21.8	21.8	21.6	21.5
175	21.9	21.7	21.8	21.4	

(a) Does the firing temperature affect the density of the bricks? Use $\alpha = 0.05$.

(b) Is it appropriate to compare the means using the Fisher LSD method (for example) in this experiment?

(c) Analyze the residuals from this experiment. Are the analysis of variance assumptions satisfied?

(d) Construct a graphical display of the treatment as described in Section 3.5.3. Does this graph adequately summarize the results of the analysis of variance in part (a)?

3.21 Rework part (d) of Problem 3.20 using the Tukey method. What conclusions can you draw? Explain carefully how you modified the technique to account for unequal sample sizes.

3.22 A manufacturer of television sets is interested in the effect on tube conductivity of four different types of coating for color picture tubes. A completely randomized experiment is conducted and the following conductivity data are obtained:

Coating Type	Conductivity			
1	143	141	150	146
2	152	149	137	143
3	134	136	132	127
4	129	127	132	129

(a) Is there a difference in conductivity due to coating type? Use $\alpha = 0.05$.

(b) Estimate the overall mean and the treatment effects.

(c) Compute a 95 percent confidence interval estimate of the mean of coating type 4. Compute a 99 percent confidence interval estimate of the mean difference between coating types 1 and 4.

(d) Test all pairs of means using the Fisher LSD method with $\alpha = 0.05$.

(e) Use the graphical method discussed in Section 3.5.3 to compare the means. Which coating type produces the highest conductivity?

(f) Assuming that coating type 4 is currently in use, what are your recommendations to the manufacturer? We wish to minimize conductivity.

3.23 Reconsider the experiment from Problem 3.22. Analyze the residuals and draw conclusions about model adequacy.

3.24 An article in the *ACI Materials Journal* (Vol. 84, 1987, pp. 213–216) describes several experiments investigating the rodding of concrete to remove entrapped air. A 3-inch × 6-inch cylinder was used, and the number of times this rod was used is the design variable. The resulting compressive strength of the concrete specimen is the response. The data are shown in the following table:

Rodding Level	Compressive Strength		
10	1530	1530	1440
15	1610	1650	1500
20	1560	1730	1530
25	1500	1490	1510

(a) Is there any difference in compressive strength due to the rodding level? Use $\alpha = 0.05$.

(b) Find the P-value for the F-statistic in part (a).

(c) Analyze the residuals from this experiment. What conclusions can you draw about the underlying model assumptions?

(d) Construct a graphical display to compare the treatment means as described in Section 3.5.3.

3.25 An article in *Environment International* (Vol. 18, No. 4, 1992) describes an experiment in which the amount of radon released in showers was investigated. Radon-enriched water was used in the experiment, and six different orifice diameters were tested in shower heads. The data from the experiment are shown in the following table:

Orifice Diameter	**Radon Released (%)**			
0.37	80	83	83	85
0.51	75	75	79	79
0.71	74	73	76	77
1.02	67	72	74	74
1.40	62	62	67	69
1.99	60	61	64	66

(a) Does the size of the orifice affect the mean percentage of radon released? Use $\alpha = 0.05$.

(b) Find the P-value for the F-statistic in part (a).

(c) Analyze the residuals from this experiment.

(d) Find a 95 percent confidence interval on the mean percent of radon released when the orifice diameter is 1.40.

(e) Construct a graphical display to compare the treatment means as described in Section 3.5.3. What conclusions can you draw?

3.26 The response time in milliseconds was determined for three different types of circuits that could be used in an automatic valve shutoff mechanism. The results from a completely randomized experiment are shown in the following table:

Circuit Type	**Response Time**				
1	9	12	10	8	15
2	20	21	23	17	30
3	6	5	8	16	7

(a) Test the hypothesis that the three circuit types have the same response time. Use $\alpha = 0.01$.

(b) Use Tukey's test to compare pairs of treatment means. Use $\alpha = 0.01$.

(c) Use the graphical procedure in Section 3.5.3 to compare the treatment means. What conclusions can you draw? How do they compare with the conclusions from part (b)?

(d) Construct a set of orthogonal contrasts, assuming that at the outset of the experiment you suspected the response time of circuit type 2 to be different from the other two.

(e) If you were the design engineer and you wished to minimize the response time, which circuit type would you select?

(f) Analyze the residuals from this experiment. Are the basic analysis of variance assumptions satisfied?

3.27 The effective life of insulating fluids at an accelerated load of 35 kV is being studied. Test data have been obtained for four types of fluids. The results from a completely randomized experiment are as follows:

Fluid Type	**Life (in h) at 35 kV Load**					
1	17.6	18.9	16.3	17.4	20.1	21.6
2	16.9	15.3	18.6	17.1	19.5	20.3
3	21.4	23.6	19.4	18.5	20.5	22.3
4	19.3	21.1	16.9	17.5	18.3	19.8

(a) Is there any indication that the fluids differ? Use $\alpha = 0.05$.

(b) Which fluid would you select, given that the objective is long life?

(c) Analyze the residuals from this experiment. Are the basic analysis of variance assumptions satisfied?

3.28 Four different designs for a digital computer circuit are being studied to compare the amount of noise present. The following data have been obtained:

Circuit Design	**Noise Observed**				
1	19	20	19	30	8
2	80	61	73	56	80
3	47	26	25	35	50
4	95	46	83	78	97

(a) Is the same amount of noise present for all four designs? Use $\alpha = 0.05$.

(b) Analyze the residuals from this experiment. Are the analysis of variance assumptions satisfied?

(c) Which circuit design would you select for use? Low noise is best.

3.29 Four chemists are asked to determine the percentage of methyl alcohol in a certain chemical compound. Each chemist makes three determinations, and the results are the following:

Chemist	Percentage of Methyl Alcohol		
1	84.99	84.04	84.38
2	85.15	85.13	84.88
3	84.72	84.48	85.16
4	84.20	84.10	84.55

(a) Do chemists differ significantly? Use $\alpha = 0.05$.

(b) Analyze the residuals from this experiment.

(c) If chemist 2 is a new employee, construct a meaningful set of orthogonal contrasts that might have been useful at the start of the experiment.

3.30 Three brands of batteries are under study. It is suspected that the lives (in weeks) of the three brands are different. Five randomly selected batteries of each brand are tested with the following results:

Weeks of Life		
Brand 1	**Brand 2**	**Brand 3**
100	76	108
96	80	100
92	75	96
96	84	98
92	82	100

(a) Are the lives of these brands of batteries different?

(b) Analyze the residuals from this experiment.

(c) Construct a 95 percent confidence interval estimate on the mean life of battery brand 2. Construct a 99 percent confidence interval estimate on the mean difference between the lives of battery brands 2 and 3.

(d) Which brand would you select for use? If the manufacturer will replace without charge any battery that fails in less than 85 weeks, what percentage would the company expect to replace?

3.31 Four catalysts that may affect the concentration of one component in a three-component liquid mixture are being investigated. The following concentrations are obtained from a completely randomized experiment:

Catalyst			
1	**2**	**3**	**4**
58.2	56.3	50.1	52.9
57.2	54.5	54.2	49.9
58.4	57.0	55.4	50.0
55.8	55.3		51.7
54.9			

(a) Do the four catalysts have the same effect on the concentration?

(b) Analyze the residuals from this experiment.

(c) Construct a 99 percent confidence interval estimate of the mean response for catalyst 1.

3.32 An experiment was performed to investigate the effectiveness of five insulating materials. Four samples of each material were tested at an elevated voltage level to accelerate the time to failure. The failure times (in minutes) are shown below:

Material	Failure Time (minutes)			
1	110	157	194	178
2	1	2	4	18
3	880	1256	5276	4355
4	495	7040	5307	10,050
5	7	5	29	2

(a) Do all five materials have the same effect on mean failure time?

(b) Plot the residuals versus the predicted response. Construct a normal probability plot of the residuals. What information is conveyed by these plots?

(c) Based on your answer to part (b), conduct another analysis of the failure time data and draw appropriate conclusions.

3.33 A semiconductor manufacturer has developed three different methods for reducing particle counts on wafers. All three methods are tested on five different wafers and the after treatment particle count obtained. The data are shown below:

Method	Count				
1	31	10	21	4	1
2	62	40	24	30	35
3	53	27	120	97	68

(a) Do all methods have the same effect on mean particle count?

(b) Plot the residuals versus the predicted response. Construct a normal probability plot of the residuals. Are there potential concerns about the validity of the assumptions?

(c) Based on your answer to part (b), conduct another analysis of the particle count data and draw appropriate conclusions.

3.34 A manufacturer suspects that the batches of raw material furnished by his supplier differ significantly in calcium content. There are a large number of batches currently in the warehouse. Five of these are randomly selected for study. A chemist makes five determinations on each batch and obtains the following data:

Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
23.46	23.59	23.51	23.28	23.29
23.48	23.46	23.64	23.40	23.46
23.56	23.42	23.46	23.37	23.37
23.39	23.49	23.52	23.46	23.32
23.40	23.50	23.49	23.39	23.38

(a) Is there significant variation in calcium content from batch to batch? Use $\alpha = 0.05$.

(b) Estimate the components of variance.

(c) Find a 95 percent confidence interval for $\sigma_\tau^2/(\sigma_\tau^2 + \sigma^2)$.

(d) Analyze the residuals from this experiment. Are the analysis of variance assumptions satisfied?

(e) Use the REML method to analyze this data. Compare the 95 percent confidence interval on the error variance from REML with the exact chi-square confidence interval.

3.35 Several ovens in a metal working shop are used to heat metal specimens. All the ovens are supposed to operate at the same temperature, although it is suspected that this may not be true. Three ovens are selected at random, and their temperatures on successive heats are noted. The data collected are as follows:

Oven	Temperature					
1	491.50	498.30	498.10	493.50	493.60	
2	488.50	484.65	479.90	477.35		
3	490.10	484.80	488.25	473.00	471.85	478.65

(a) Is there significant variation in temperature between ovens? Use $\alpha = 0.05$.

(b) Estimate the components of variance for this model.

(c) Analyze the residuals from this experiment and draw conclusions about model adequacy.

3.36 An article in the *Journal of the Electrochemical Society* (Vol. 139, No. 2, 1992, pp. 524–532) describes an experiment to investigate the low-pressure vapor deposition of polysilicon. The experiment was carried out in a large-capacity reactor at Sematech in Austin, Texas. The reactor has several wafer positions, and four of these positions are selected at random. The response variable is film thickness uniformity. Three replicates of the experiment were run, and the data are as follows:

Wafer Position	Uniformity		
1	2.76	5.67	4.49
2	1.43	1.70	2.19
3	2.34	1.97	1.47
4	0.94	1.36	1.65

(a) Is there a difference in the wafer positions? Use $\alpha = 0.05$.

(b) Estimate the variability due to wafer positions.

(c) Estimate the random error component.

(d) Analyze the residuals from this experiment and comment on model adequacy.

3.37 Consider the vapor-deposition experiment described in Problem 3.36.

(a) Estimate the total variability in the uniformity response.

(b) How much of the total variability in the uniformity response is due to the difference between positions in the reactor?

(c) To what level could the variability in the uniformity response be reduced if the position-to-position variability in the reactor could be eliminated? Do you believe this is a significant reduction?

3.38 A single-factor completely randomized design has four levels of the factor. There are three replicates and the total sum of squares is 330.56. The treatment sum of squares is 250.65.

(a) What is the estimate of the error variance σ^2?

(b) What proportion of the variability in the response variable is explained by the treatment effect?

3.39 A single-factor completely randomized design has six levels of the factor. There are five replicates and the total sum of squares is 900.25. The treatment sum of squares is 750.50.

(a) What is the estimate of the error variance σ^2?

(b) What proportion of the variability in the response variable is explained by the treatment effect?

3.40 Find a 95% confidence interval on the intraclass correlation coefficient for the experiment in Problem 3.38.

3.41 Find a 95% confidence interval on the intraclass correlation coefficient for the experiment in Problem 3.39.

3.42 An article in the *Journal of Quality Technology* (Vol. 13, No. 2, 1981, pp. 111–114) describes an experiment that investigates the effects of four bleaching chemicals on pulp brightness. These four chemicals were selected at random from a large population of potential bleaching agents. The data are as follows:

Oven	**Temperature**				
1	77.199	74.466	92.746	76.208	82.876
2	80.522	79.306	81.914	80.346	73.385
3	79.417	78.017	91.596	80.802	80.626
4	78.001	78.358	77.544	77.364	77.386

(a) Is there a difference in the chemical types? Use $\alpha = 0.05$.

(b) Estimate the variability due to chemical types.

(c) Estimate the variability due to random error.

(d) Analyze the residuals from this experiment and comment on model adequacy.

3.43 Consider the single-factor random effects model discussed in this chapter. Develop a procedure for finding a $100(1 - \alpha)$ percent confidence interval on the ratio $\sigma^2/(\sigma_\tau^2 + \sigma^2)$. Assume that the experiment is balanced.

3.44 Consider testing the equality of the means of two normal populations, where the variances are unknown but are assumed to be equal. The appropriate test procedure is the pooled *t*-test. Show that the pooled *t*-test is equivalent to the single-factor analysis of variance.

3.45 Show that the variance of the linear combination $\sum_{i=1}^{a} c_i y_{i.}$ is $\sigma^2 \sum_{i=1}^{a} n_i c_i^2$.

3.46 In a fixed effects experiment, suppose that there are n observations for each of the four treatments. Let Q_1^2, Q_2^2, Q_3^2 be single-degree-of-freedom components for the orthogonal contrasts. Prove that $SS_{\text{Treatments}} = Q_1^2 + Q_2^2 + Q_3^2$.

3.47 Use Bartlett's test to determine if the assumption of equal variances is satisfied in Problem 3.30. Use $\alpha = 0.05$. Did you reach the same conclusion regarding equality of variances by examining residual plots?

3.48 Use the modified Levene test to determine if the assumption of equal variances is satisfied in Problem 3.30. Use $\alpha = 0.05$. Did you reach the same conclusion regarding the equality of variances by examining residual plots?

3.49 Refer to Problem 3.26. If we wish to detect a maximum difference in mean response times of 10 milliseconds with a probability of at least 0.90, what sample size should be used? How would you obtain a preliminary estimate of σ^2?

3.50 Refer to Problem 3.30.

(a) If we wish to detect a maximum difference in battery life of 10 hours with a probability of at least 0.90, what sample size should be used? Discuss how you would obtain a preliminary estimate of σ^2 for answering this question.

(b) If the maximum difference between brands is 8 hours, what sample size should be used if we wish to detect this with a probability of at least 0.90?

3.51 Consider the experiment in Problem 3.30. If we wish to construct a 95 percent confidence interval on the difference in two mean battery lives that has an accuracy of ±2 weeks, how many batteries of each brand must be tested?

3.52 Suppose that four normal populations have means of $\mu_1 = 50$, $\mu_2 = 60$, $\mu_3 = 50$, and $\mu_4 = 60$. How many observations should be taken from each population so that the probability of rejecting the null hypothesis of equal population means is at least 0.90? Assume that $\alpha = 0.05$ and that a reasonable estimate of the error variance is $\sigma^2 = 25$.

3.53 Refer to Problem 3.52.

(a) How would your answer change if a reasonable estimate of the experimental error variance were $\sigma^2 = 36$?

(b) How would your answer change if a reasonable estimate of the experimental error variance were $\sigma^2 = 49$?

(c) Can you draw any conclusions about the sensitivity of your answer in this particular situation about how your estimate of σ affects the decision about sample size?

(d) Can you make any recommendations about how we should use this general approach to choosin g n in practice?

3.54 Refer to the aluminum smelting experiment described in Section 3.8.3. Verify that ratio control methods do not affect average cell voltage. Construct a normal probability plot of the residuals. Plot the residuals versus the predicted values. Is there an indication that any underlying assumptions are violated?

3.55 Refer to the aluminum smelting experiment in Section 3.8.3. Verify the ANOVA for pot noise summarized in Table 3.17. Examine the usual residual plots and comment on the experimental validity.

3.56 Four different feed rates were investigated in an experiment on a CNC machine producing a component part used in an aircraft auxiliary power unit. The manufacturing engineer in charge of the experiment knows that a critical part dimension of interest may be affected by the feed rate. However, prior experience has indicated that only dispersion effects are likely to be present. That is, changing the feed rate does not affect the *average* dimension, but it could affect dimensional variability. The engineer makes five production runs at each feed rate and obtains the standard deviation of the critical dimension (in 10^{-3} mm). The data are shown below. Assume that all runs were made in random order.

Feed Rate	Production Run				
(in/min)	1	2	3	4	5
10	0.09	0.10	0.13	0.08	0.07
12	0.06	0.09	0.12	0.07	0.12
14	0.11	0.08	0.08	0.05	0.06
16	0.19	0.13	0.15	0.20	0.11

(a) Does feed rate have any effect on the standard deviation of this critical dimension?

(b) Use the residuals from this experiment to investigate model adequacy. Are there any problems with experimental validity?

3.57 Consider the data shown in Problem 3.26.

(a) Write out the least squares normal equations for this problem and solve them for $\hat{\mu}$ and $\hat{\tau}_i$, using the usual constraint $\left(\sum_{i=1}^{3} \hat{\tau}_i = 0\right)$. Estimate $\tau_1 - \tau_2$.

(b) Solve the equations in (a) using the constraint $\hat{\tau}_3 = 0$. Are the estimators $\hat{\tau}_i$ and $\hat{\mu}$ the same as you found in (a)? Why? Now estimate $\tau_1 - \tau_2$ and compare your answer with that for (a). What statement can you make about estimating contrasts in the τ_i?

(c) Estimate $\mu + \tau_1$, $2\tau_1 - \tau_2 - \tau_3$, and $\mu + \tau_1 + \tau_2$ using the two solutions to the normal equations. Compare the results obtained in each case.

3.58 Apply the general regression significance test to the experiment in Example 3.6. Show that the procedure yields the same results as the usual analysis of variance.

3.59 Use the Kruskal–Wallis test for the experiment in Problem 3.27. Compare the conclusions obtained with those from the usual analysis of variance.

3.60 Use the Kruskal–Wallis test for the experiment in Problem 3.28. Are the results comparable to those found by the usual analysis of variance?

3.61 Consider the experiment in Example 3.6. Suppose that the largest observation on etch rate is incorrectly recorded as 250 Å/min. What effect does this have on the usual analysis of variance? What effect does it have on the Kruskal–Wallis test?

3.62 A textile mill has a large number of looms. Each loom is supposed to provide the same output of cloth per minute. To investigate this assumption, five looms are chosen at random, and their output is noted at different times. The following data are obtained:

Loom	Output (lb/min)				
1	14.0	14.1	14.2	14.0	14.1
2	13.9	13.8	13.9	14.0	14.0
3	14.1	14.2	14.1	14.0	13.9
4	13.6	13.8	14.0	13.9	13.7
5	13.8	13.6	13.9	13.8	14.0

(a) Explain why this is a random effects experiment. Are the looms equal in output? Use $\alpha = 0.05$.

(b) Estimate the variability between looms.

(c) Estimate the experimental error variance.

(d) Find a 95 percent confidence interval for $\sigma_\tau^2/(\sigma_\tau^2 + \sigma^2)$.

(e) Analyze the residuals from this experiment. Do you think that the analysis of variance assumptions are satisfied?

(f) Use the REML method to analyze this data. Compare the 95 percent confidence interval on the error variance from REML with the exact chi-square confidence interval.

3.63 The normality assumption is extremely important in the analysis of variance.

(a) True

(b) False

3.64 The analysis of variance treats both quantitative and qualitative factors alike so far as the basic computations for sums of squares are concerned.

(a) True

(b) False

3.65 If a single-factor experiment has a levels of the factor and a polynomial of degree a – 1 is fit to the experimental data, the error sum of squares for the polynomial model will be exactly the same as the error sum of squares for the standard ANOVA.

(a) True

(b) False

3.66 Fisher's LSD procedure is an extremely conservative method for comparing pairs of treatment means following an ANOVA.

(a) True

(b) False

3.67 The REML method of estimating variance components is a technique based on maximum likelihood, while the ANOVA method is a method-of-moments procedure.

(a) True

(b) False

3.68 One advantage of the REML method of estimating variance components is that it automatically produces confidence intervals on the variance components.

(a) True

(b) False

3.69 The Tukey method is used to compare all treatment means to a control.

(a) True

(b) False

3.70 An experiment with a single factor has been conducted as a completely randomized design and analyzed using computer software. A portion of the output is shown below.

```
Source  DF      SS     MS     F
Factor   ?       ?  25.69  3.65
Error   12   84.35      ?
Total   15  161.42
```

(a) Fill in the missing information.

(b) How many levels of the factor were used in this experiment?

(c) How many replicates were used in this experiment?

(d) Find bounds on the *P*-value.

3.71 The estimate of the standard deviation of any observation in the experiment in Problem 3.70 is

(a) 7.03 **(b)** 2.65 **(c)** 5.91

(d) 1.95 **(e)** none of the above

CHAPTER 4

Randomized Blocks, Latin Squares, and Related Designs

CHAPTER OUTLINE

The supplemental material is on the textbook website www.wiley.com/college/montgomery.

CHAPTER LEARNING OBJECTIVES

1. Learn about how the blocking principle can be effective in reducing the variability arising from controllable nuisance factors.
2. Learn about the randomized complete block design.
3. Understand how the analysis of variance can be extended to the randomized complete block design.
4. Know how to do model adequacy checking for the randomized complete block design.
5. Understand how a Latin square design can be used to control two sources of nuisance variability in an experiment.

4.1 The Randomized Complete Block Design

In any experiment, variability arising from a nuisance factor can affect the results. Generally, we define a **nuisance factor** as a design factor that probably has an effect on the response, but we are not interested in that effect. Sometimes a nuisance factor is **unknown and uncontrolled**; that is, we don't know that the factor exists, and it may even be changing levels while we are conducting the experiment. **Randomization** is the design technique used to guard against such a "lurking" nuisance factor. In other cases, the nuisance factor is **known but uncontrollable**. If we can at least observe the value that the nuisance factor takes on at each run of the experiment, we can compensate for it in the statistical analysis by using the **analysis of covariance**, a technique we will discuss in Chapter 15. When the nuisance source of variability is **known and controllable**, a design technique called **blocking** can be used to systematically eliminate its effect on the statistical comparisons among treatments. Blocking is an extremely important design technique used extensively in industrial experimentation and is the subject of this chapter.

To illustrate the general idea, reconsider the hardness testing experiment first described in Section 2.5.1. Suppose now that we wish to determine whether or not four different tips produce different readings on a hardness testing machine. An experiment such as this might be part of a gauge capability study. The machine operates by pressing the tip into a metal test coupon, and from the depth of the resulting depression, the hardness of the coupon can be determined. The experimenter has decided to obtain four observations on Rockwell C-scale hardness for each tip. There is only one factor—tip type—and a completely randomized single-factor design would consist of randomly assigning each one of the $4 \times 4 = 16$ runs to an **experimental unit**, that is, a metal coupon, and observing the hardness reading that results. Thus, 16 different metal test coupons would be required in this experiment, one for each run in the design.

There is a potentially serious problem with a completely randomized experiment in this design situation. If the metal coupons differ slightly in their hardness, as might happen if they are taken from ingots that are produced in different heats, the experimental units (the coupons) will contribute to the variability observed in the hardness data. As a result, the experimental error will reflect *both* random error *and* variability between coupons.

We would like to make the experimental error as small as possible; that is, we would like to remove the variability between coupons from the experimental error. A design that would accomplish this requires the experimenter to test each tip once on each of four coupons. This design, shown in Table 4.1, is called a **randomized complete block design (RCBD)**. The word "complete" indicates that each block (coupon) contains all the treatments (tips). By using this design, the blocks, or coupons, form a more homogeneous experimental unit on which to compare the tips. Effectively, this design strategy improves the accuracy of the comparisons among tips by eliminating the variability among the coupons. Within a block, the order in which the four tips are tested is randomly determined. Notice the similarity of this design problem to the paired *t*-test of Section 2.5.1. The randomized complete block design is a generalization of that concept.

The RCBD is one of the most widely used experimental designs. Situations for which the RCBD is appropriate are numerous. Units of test equipment or machinery are often different in their operating characteristics and would be a typical blocking factor. Batches of raw material, people, and time are also common nuisance sources of variability in an experiment that can be systematically controlled through blocking.[1]

Blocking may also be useful in situations that do not necessarily involve nuisance factors. For example, suppose that a chemical engineer is interested in the effect of catalyst feed rate on the viscosity of a polymer. She knows that there are several factors, such as raw material source, temperature, operator, and raw material purity that are very difficult to control in the full-scale process. Therefore, she decides to test the catalyst feed rate factor in blocks, where

■ **TABLE 4.1**
Randomized Complete Block Design for the Hardness Testing Experiment

Test Coupon (Block)			
1	**2**	**3**	**4**
Tip 3	Tip 3	Tip 2	Tip 1
Tip 1	Tip 4	Tip 1	Tip 4
Tip 4	Tip 2	Tip 3	Tip 2
Tip 2	Tip 1	Tip 4	Tip 3

[1] A special case of blocking occurs where the blocks are experimental units such as people, and each block receives the treatments over time or the treatment effects are measured at different times. These are called **repeated measures** designs. They are discussed in Chapter 15.

each block consists of some combination of these uncontrollable factors. In effect, she is using the blocks to test the **robustness** of her process variable (feed rate) to conditions she cannot easily control. For more discussion of this, see Coleman and Montgomery (1993).

4.1.1 Statistical Analysis of the RCBD

Suppose we have, in general, *a* treatments that are to be compared and *b* blocks. The randomized complete block design is shown in Figure 4.1. There is one observation per treatment in each block, and the order in which the treatments are run within each block is determined randomly. Because the only randomization of treatments is within the blocks, we often say that the blocks represent a **restriction on randomization**.

The **statistical model** for the RCBD can be written in several ways. The traditional model is an **effects model**:

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \qquad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \end{cases} \tag{4.1}$$

where μ is an overall mean, τ_i is the effect of the *i*th treatment, β_j is the effect of the *j*th block, and ϵ_{ij} is the usual NID $(0, \sigma^2)$ random error term. We will initially consider treatments and blocks to be fixed factors. The case of random blocks, which is very important, is considered in Section 4.1.3. Just as in the single-factor experimental design model in Chapter 3, the effects model for the RCBD is an overspecified model. Consequently, we usually think of the treatment and block effects as deviations from the overall mean so that

$$\sum_{i=1}^{a} \tau_i = 0 \quad \text{and} \quad \sum_{j=1}^{b} \beta_j = 0$$

It is also possible to use a **means model** for the RCBD, say

$$y_{ij} = \mu_{ij} + \epsilon_{ij} \qquad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \end{cases}$$

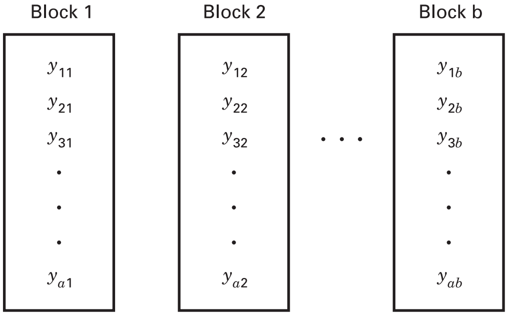
where $\mu_{ij} = \mu + \tau_i + \beta_j$. However, we will use the effects model in Equation 4.1 throughout this chapter.

In an experiment involving the RCBD, we are interested in testing the equality of the treatment means. Thus, the hypotheses of interest are

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 = \cdots = \mu_a \\ &H_1 : \text{at least one } \mu_i \neq \mu_j \end{aligned}$$

Because the *i*th treatment mean $\mu_i = (1/b) \sum_{j=1}^{b} (\mu + \tau_i + \beta_j) = \mu + \tau_i$, an equivalent way to write the above hypotheses is in terms of the treatment effects, say

$$\begin{aligned} &H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0 \\ &H_1 : \tau_i \neq 0 \text{ at least one } i \end{aligned}$$



■ **FIGURE 4.1 The randomized complete block design**

The analysis of variance can be easily extended to the RCBD. Let $y_i.$ be the total of all observations taken under treatment $i, y_{.j}$ be the total of all observations in block $j, y_{..}$ be the grand total of all observations, and $N = ab$ be the total number of observations. Expressed mathematically,

$$y_{i.} = \sum_{j=1}^{b} y_{ij} \qquad i = 1, 2, \ldots, a \tag{4.2}$$

$$y_{.j} = \sum_{i=1}^{a} y_{ij} \qquad j = 1, 2, \ldots, b \tag{4.3}$$

and

$$y_{..} = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij} = \sum_{i=1}^{a} y_{i.} = \sum_{j=1}^{b} y_{.j} \tag{4.4}$$

Similarly, $\bar{y}_{i.}$ is the average of the observations taken under treatment $i, \bar{y}_{.j}$ is the average of the observations in block j, and $\bar{y}_{..}$ is the grand average of all observations. That is,

$$\bar{y}_{i.} = y_{i.}/b \quad \bar{y}_{.j} = y_{.j}/a \quad \bar{y}_{..} = y_{..}/N \tag{4.5}$$

We may express the total corrected sum of squares as

$$\sum_{i=1}^{a}\sum_{j=1}^{b} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} [(\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}]^2 \tag{4.6}$$

By expanding the right-hand side of Equation 4.6, we obtain

$$\begin{aligned}\sum_{i=1}^{a}\sum_{j=1}^{b} (y_{ij} - \bar{y}_{..})^2 &= b\sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2 + a\sum_{j=1}^{b} (\bar{y}_{.j} - \bar{y}_{..})^2 \\ &+ \sum_{i=1}^{a}\sum_{j=1}^{b} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 + 2\sum_{i=1}^{a}\sum_{j=1}^{b} (\bar{y}_{i.} - \bar{y}_{..})(\bar{y}_{.j} - \bar{y}_{..}) \\ &+ 2\sum_{i=1}^{a}\sum_{j=1}^{b} (\bar{y}_{.j} - \bar{y}_{..})(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}) \\ &+ 2\sum_{i=1}^{a}\sum_{j=1}^{b} (\bar{y}_{i.} - \bar{y}_{..})(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})\end{aligned}$$

Simple but tedious algebra proves that the three cross products are zero. Therefore,

$$\begin{aligned}\sum_{i=1}^{a}\sum_{j=1}^{b} (y_{ij} - \bar{y}_{..})^2 &= b\sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2 + a\sum_{j=1}^{b} (\bar{y}_{.j} - \bar{y}_{..})^2 \\ &+ \sum_{i=1}^{a}\sum_{j=1}^{b} (y_{ij} - \bar{y}_{.j} - \bar{y}_{i.} + \bar{y}_{..})^2\end{aligned} \tag{4.7}$$

represents a partition of the total sum of squares. This is the fundamental ANOVA equation for the RCBD. Expressing the sums of squares in Equation 4.7 symbolically, we have

$$SS_T = SS_{\text{Treatments}} + SS_{\text{Blocks}} + SS_E \tag{4.8}$$

Because there are N observations, SS_T has $N - 1$ degrees of freedom. There are a treatments and b blocks, so $SS_{\text{Treatments}}$ and SS_{Blocks} have $a - 1$ and $b - 1$ degrees of freedom, respectively. The error sum of squares is just a sum of squares between cells minus the sum of squares for treatments and blocks. There are ab cells with $ab - 1$ degrees of freedom between them, so SS_E has $ab - 1 - (a - 1) - (b - 1) = (a - 1)(b - 1)$ degrees of freedom. Furthermore, the degrees of freedom on the right-hand side of Equation 4.8 add to the total on the left; therefore, making the usual normality assumptions on the errors, one may use Theorem 3-1 to show that $SS_{\text{Treatments}}/\sigma^2, SS_{\text{Blocks}}/\sigma^2$, and SS_E/σ^2 are independently distributed chi-square random variables. Each sum of squares divided by its degrees of freedom is a mean square. The expected value of the mean squares, if treatments and blocks are fixed, can be shown to be

$$E(MS_{\text{Treatments}}) = \sigma^2 + \frac{b\sum_{i=1}^{a}\tau_i^2}{a-1}$$

$$E(MS_{\text{Blocks}}) = \sigma^2 + \frac{a\sum_{j=1}^{b}\beta_j^2}{b-1}$$

$$E(MS_E) = \sigma^2$$

Therefore, to test the equality of treatment means, we would use the test statistic

$$F_0 = \frac{MS_{\text{Treatments}}}{MS_E}$$

which is distributed as $F_{\alpha-1,(a-1)(b-1)}$ if the null hypothesis is true. The critical region is the upper tail of the F distribution, and we would reject H_0 if $F_0 > F_{\alpha,a-1,(a-1)(b-1)}$. A P-value approach can also be used.

We may also be interested in comparing block means because, if these means do not differ greatly, blocking may not be necessary in future experiments. From the expected mean squares, it seems that the hypothesis $H_0 : \beta_j = 0$ may be tested by comparing the statistic $F_0 = MS_{\text{Blocls}}/MS_E$ to $F_{\alpha,b-1,(a-1)(b-1)}$. However, recall that randomization has been applied only to treatments *within* blocks; that is, the blocks represent a **restriction on randomization**. What effect does this have on the statistic $F_0 = MS_{\text{Blocks}}/MS_E$? Some differences in treatment of this question exist. For example, Box, Hunter, and Hunter (2005) point out that the usual analysis of variance F-test can be justified on the basis of randomization only,[2] without direct use of the normality assumption. They further observe that the test to compare block means cannot appeal to such a justification because of the randomization restriction; but if the errors are NID$(0, \sigma^2)$, the statistic $F_0 = MS_{\text{Blocks}}/MS_E$ can be used to compare block means. On the other hand, Anderson and McLean (1974) argue that the randomization restriction prevents this statistic from being a meaningful test for comparing block means and that this F ratio really is a test for the equality of the block means plus the randomization restriction [which they call a restriction error; see Anderson and McLean (1974) for further details].

In practice, then, what do we do? Because the normality assumption is often questionable, to view $F_0 = MS_{\text{Blocks}}/MS_E$ as an exact F-test on the equality of block means is not a good general practice. For that reason, we exclude this F-test from the analysis of variance table. However, as an approximate procedure to investigate the effect of the blocking variable, examining the ratio of MS_{Blocks} to MS_E is certainly reasonable. If this ratio is large, it implies that the blocking factor has a large effect and that the noise reduction obtained by blocking was probably helpful in improving the precision of the comparison of treatment means.

The procedure is usually summarized in an ANOVA table, such as the one shown in Table 4.2. The computing would usually be done with a statistical software package. However, computing formulas for the sums of squares may be obtained for the elements in Equation 4.7 by working directly with the identity

$$y_{ij} - \bar{y}_{..} = (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})$$

[2] Actually, the normal-theory F distribution is an approximation to the randomization distribution generated by calculating F_0 from every possible assignment of the responses to the treatments.

■ **TABLE 4.2**
Analysis of Variance for a Randomized Complete Block Design

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$SS_{\text{Treatments}}$	$a-1$	$\frac{SS_{\text{Treatments}}}{a-1}$	$\frac{MS_{\text{Treatments}}}{MS_E}$
Blocks	SS_{Blocks}	$b-1$	$\frac{SS_{\text{Blocks}}}{b-1}$	
Error	SS_E	$(a-1)(b-1)$	$\frac{SS_E}{(a-1)(b-1)}$	
Total	SS_T	$N-1$		

These quantities can be computed in the columns of a spreadsheet (Excel). Then each column can be squared and summed to produce the sum of squares. Alternatively, computing formulas can be expressed in terms of treatment and block totals. These formulas are

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2 - \frac{y_{..}^2}{N} \tag{4.9}$$

$$SS_{\text{Treatments}} = \frac{1}{b}\sum_{i=1}^{a} y_{i.}^2 - \frac{y_{..}^2}{N} \tag{4.10}$$

$$SS_{\text{Blocks}} = \frac{1}{a}\sum_{j=1}^{b} y_{.j}^2 - \frac{y_{..}^2}{N} \tag{4.11}$$

and the error sum of squares is obtained by subtraction as

$$SS_E = SS_T - SS_{\text{Treatments}} - SS_{\text{Blocks}} \tag{4.12}$$

EXAMPLE 4.1

A medical device manufacturer produces vascular grafts (artificial veins). These grafts are produced by extruding billets of polytetrafluoroethylene (PTFE) resin combined with a lubricant into tubes. Frequently, some of the tubes in a production run contain small, hard protrusions on the external surface. These defects are known as “flicks.” The defect is cause for rejection of the unit.

The product developer responsible for the vascular grafts suspects that the extrusion pressure affects the occurrence of flicks and therefore intends to conduct an experiment to investigate this hypothesis. However, the resin is manufactured by an external supplier and is delivered to the medical device manufacturer in batches. The engineer also suspects that there may be significant batch-to-batch variation, because while the material should be consistent with respect to parameters such as molecular weight, mean particle size, retention, and peak height ratio, it probably isn’t due to manufacturing variation at the resin supplier and natural variation in the material. Therefore, the product developer decides to investigate the effect of four different levels of extrusion pressure on flicks using a randomized complete block design considering batches of resin as blocks. The RCBD is shown in Table 4.3. Note that there are four levels of extrusion pressure (treatments) and six batches of resin (blocks). Remember that the order in which the extrusion pressures are tested within each block is random. The response variable is yield, or the percentage of tubes in the production run that did not contain any flicks.

■ **TABLE 4.3**
Randomized Complete Block Design for the Vascular Graft Experiment

Extrusion Pressure (PSI)	Batch of Resin (Block)						Treatment Total
	1	2	3	4	5	6	
8500	90.3	89.2	98.2	93.9	87.4	97.9	556.9
8700	92.5	89.5	90.6	94.7	87.0	95.8	550.1
8900	85.5	90.8	89.6	86.2	88.0	93.4	533.5
9100	82.5	89.5	85.6	87.4	78.9	90.7	514.6
Block totals	350.8	359.0	364.0	362.2	341.3	377.8	$y_{..} = 2155.1$

To perform the analysis of variance, we need the following sums of squares:

$$SS_T = \sum_{i=1}^{4}\sum_{j=1}^{6} y_{ij}^2 - \frac{y_{..}^2}{N}$$

$$= 193{,}999.31 - \frac{(2155.1)^2}{24} = 480.31$$

$$SS_{\text{Treatments}} = \frac{1}{b}\sum_{i=1}^{4} y_{i.}^2 - \frac{y_{..}^2}{N}$$

$$= \frac{1}{6}[(556.9)^2 + (550.1)^2 + (533.5)^2$$

$$+(514.6)^2] - \frac{(2155.1)^2}{24} = 178.17$$

$$SS_{\text{Blocks}} = \frac{1}{a}\sum_{j=1}^{6} y_{.j}^2 - \frac{y_{..}^2}{N}$$

$$= \frac{1}{4}[(350.8)^2 + (359.0)^2 + \cdots + (377.8)^2]$$

$$-\frac{(2155.1)^2}{24} = 192.25$$

$$SS_E = SS_T - SS_{\text{Treatments}} - SS_{\text{Blocks}}$$

$$= 480.31 - 178.17 - 192.25 = 109.89$$

The ANOVA is shown in Table 4.4. Using $\alpha = 0.05$, the critical value of F is $F_{0.05,3,15} = 3.29$. Because $8.11 > 3.29$, we conclude that extrusion pressure affects the mean yield. The P-value for the test is also quite small. Also, the resin batches (blocks) seem to differ significantly, because the mean square for blocks is large relative to error.

■ **TABLE 4.4**
Analysis of Variance for the Vascular Graft Experiment

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-Value
Treatments (extrusion pressure)	178.17	3	59.39	8.11	0.0019
Blocks (batches)	192.25	5	38.45		
Error	109.89	15	7.33		
Total	480.31	23			

It is interesting to observe the results we would have obtained from this experiment had we not been aware of randomized block designs. Suppose that this experiment had been run as a completely randomized design, and (by chance) the same design resulted as in Table 4.3. The incorrect analysis of these data as a completely randomized single-factor design is shown in Table 4.5.

Because the P-value is less than 0.05, we would still reject the null hypothesis and conclude that extrusion pressure significantly affects the mean yield. However, note that the mean square for error has more than doubled, increasing from 7.33 in the RCBD to 15.11. All of the variability due to blocks is now in the error term. This makes it easy to see why we sometimes call the RCBD a noise-reducing design technique; it effectively increases the signal-to-noise ratio

■ **TABLE 4.5**
Incorrect Analysis of the Vascular Graft Experiment as a Completely Randomized Design

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	*P*-Value
Extrusion pressure	178.17	3	59.39	3.95	0.0235
Error	302.14	20	15.11		
Total	480.31	23			

in the data, or it improves the precision with which treatment means are compared. This example also illustrates an important point. If an experimenter fails to block when he or she should have, the effect may be to inflate the experimental error, and it would be possible to inflate the error so much that important differences among the treatment means could not be identified.

Sample Computer Output. Condensed computer output for the vascular graft experiment in Example 4.1, obtained from Design-Expert and JMP, is shown in Figure 4.2. The Design-Expert output is in Figure 4.2*a* and the JMP output is in Figure 4.2*b*. Both outputs are very similar and match the manual computation given earlier. Note that JMP computes an *F*-statistic for blocks (the batches). The sample means for each treatment are shown in the output. At 8500 psi, the mean yield is $\overline{y}_{1.} = 92.82$, at 8700 psi the mean yield is $\overline{y}_{2.} = 91.68$, at 8900 psi the mean yield is $\overline{y}_{3.} = 88.92$, and at 9100 psi the mean yield is $\overline{y}_{4.} = 85.77$. Remember that these sample mean yields estimate the treatment means μ_1, μ_2, μ_3, and μ_4. The model residuals are shown at the bottom of the Design-Expert output. The residuals are calculated from

$$e_{ij} = y_{ij} - \hat{y}_{ij}$$

and, as we will later show, the fitted values are $\hat{y}_{ij} = \overline{y}_{i.} + \overline{y}_{.j} - \overline{y}_{..}$, so

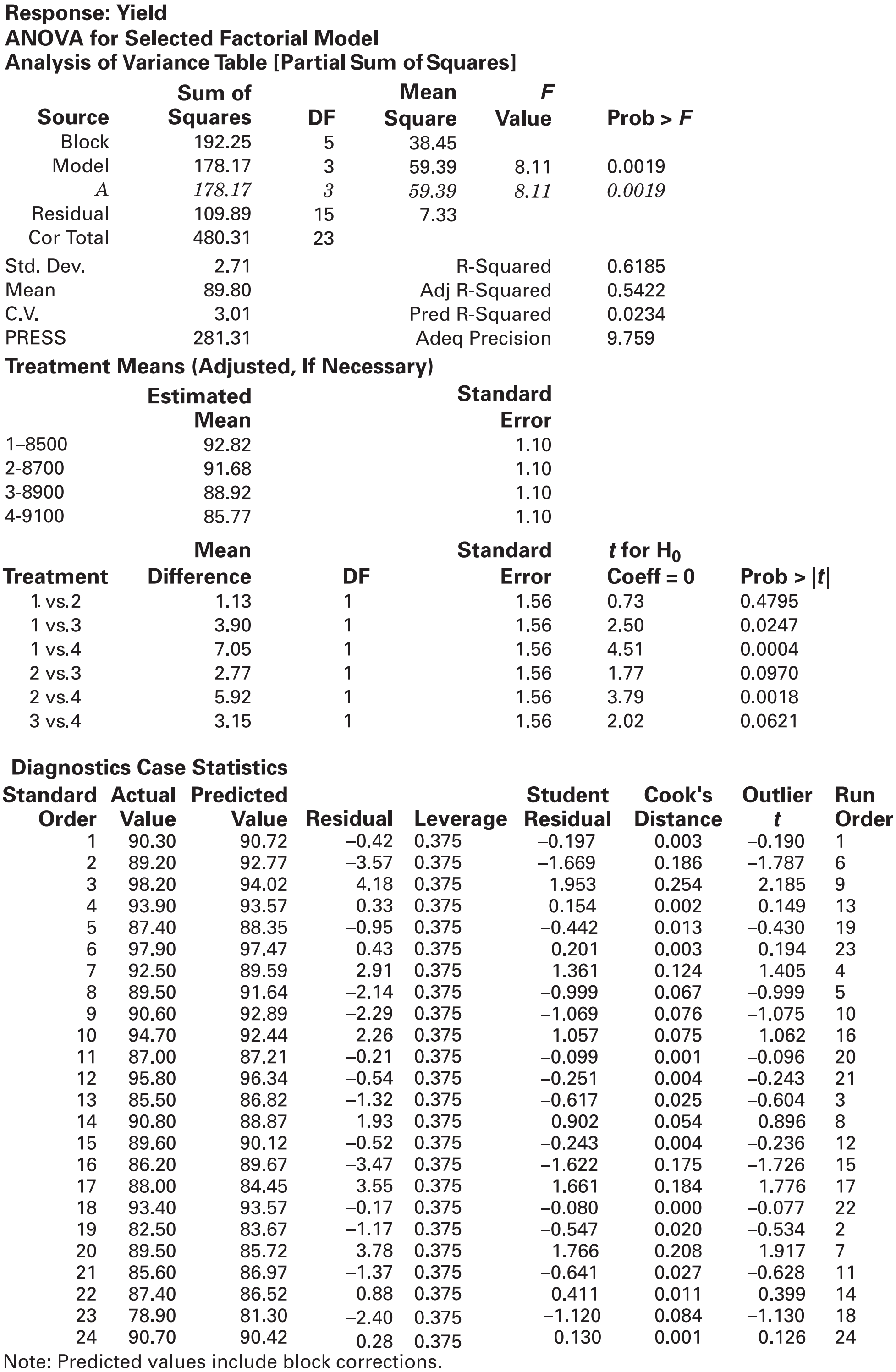
$$e_{ij} = y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..} \tag{4.13}$$

In the next section, we will show how the residuals are used in **model adequacy checking**.

Multiple Comparisons. If the treatments in an RCBD are fixed, and the analysis indicates a significant difference in treatment means, the experimenter is usually interested in multiple comparisons to discover *which* treatment means differ. Any of the multiple comparison procedures discussed in Section 3.5 may be used for this purpose. In the formulas of Section 3.5, simply replace the number of replicates in the single-factor completely randomized design (n) by the number of blocks (b). Also, remember to use the number of error degrees of freedom for the randomized block $[(a-1)(b-1)]$ instead of those for the completely randomized design $[a(n-1)]$.

The Design-Expert output in Figure 4.2 illustrates the Fisher LSD procedure. Notice that we would conclude that $\mu_1 = \mu_2$, because the *P*-value is very large. Furthermore, μ_1 differs from all other means. Now the *P*-value for $H_0 : \mu_2 = \mu_3$ is 0.097, so there is some evidence to conclude that $\mu_2 \neq \mu_3$, and $\mu_2 \neq \mu_4$ because the *P*-value is 0.0018. Overall, we would conclude that lower extrusion pressures (8500 psi and 8700 psi) lead to fewer defects.

We can also use the graphical procedure of Section 3.5.1 to compare mean yield at the four extrusion pressures. Figure 4.3 plots the four means from Example 4.1 relative to a scaled *t* distribution with a scale factor $\sqrt{MS_E/b} = \sqrt{7.33/6} = 1.10$. This plot indicates that the two lowest pressures result in the same mean yield, but that the mean yields for 8700 psi and 8900 psi (μ_2 and μ_3) are also similar. The highest pressure (9100 psi) results in a mean yield that is much lower than all other means. This figure is a useful aid in interpreting the results of the experiment and the Fisher LSD calculations in the Design-Expert output in Figure 4.2.



```
Response: Yield
ANOVA for Selected Factorial Model
Analysis of Variance Table [Partial Sum of Squares]
                 Sum of               Mean        F
      Source    Squares     DF      Square    Value       Prob > F
       Block     192.25      5       38.45
       Model     178.17      3       59.39     8.11       0.0019
           A     178.17      3       59.39     8.11       0.0019
    Residual     109.89     15        7.33
   Cor Total     480.31     23
Std. Dev.          2.71                 R-Squared         0.6185
Mean              89.80             Adj R-Squared         0.5422
C.V.               3.01            Pred R-Squared         0.0234
PRESS            281.31            Adeq Precision         9.759
Treatment Means (Adjusted, If Necessary)
              Estimated                  Standard
                   Mean                     Error
1–8500            92.82                      1.10
2-8700            91.68                      1.10
3-8900            88.92                      1.10
4-9100            85.77                      1.10
                   Mean                  Standard       t for H0
Treatment    Difference       DF            Error       Coeff = 0      Prob > |t|
   1 vs.2          1.13        1             1.56       0.73           0.4795
   1 vs.3          3.90        1             1.56       2.50           0.0247
   1 vs.4          7.05        1             1.56       4.51           0.0004
   2 vs.3          2.77        1             1.56       1.77           0.0970
   2 vs.4          5.92        1             1.56       3.79           0.0018
   3 vs.4          3.15        1             1.56       2.02           0.0621

 Diagnostics Case Statistics
Standard  Actual  Predicted                        Student    Cook's    Outlier  Run
   Order   Value      Value  Residual  Leverage   Residual  Distance         t   Order
       1   90.30      90.72     –0.42  0.375        –0.197     0.003    –0.190   1
       2   89.20      92.77     –3.57  0.375        –1.669     0.186    –1.787   6
       3   98.20      94.02      4.18  0.375         1.953     0.254     2.185   9
       4   93.90      93.57      0.33  0.375         0.154     0.002     0.149   13
       5   87.40      88.35     –0.95  0.375        –0.442     0.013    –0.430   19
       6   97.90      97.47      0.43  0.375         0.201     0.003     0.194   23
       7   92.50      89.59      2.91  0.375         1.361     0.124     1.405   4
       8   89.50      91.64     –2.14  0.375        –0.999     0.067    –0.999   5
       9   90.60      92.89     –2.29  0.375        –1.069     0.076    –1.075   10
      10   94.70      92.44      2.26  0.375         1.057     0.075     1.062   16
      11   87.00      87.21     –0.21  0.375        –0.099     0.001    –0.096   20
      12   95.80      96.34     –0.54  0.375        –0.251     0.004    –0.243   21
      13   85.50      86.82     –1.32  0.375        –0.617     0.025    –0.604   3
      14   90.80      88.87      1.93  0.375         0.902     0.054     0.896   8
      15   89.60      90.12     –0.52  0.375        –0.243     0.004    –0.236   12
      16   86.20      89.67     –3.47  0.375        –1.622     0.175    –1.726   15
      17   88.00      84.45      3.55  0.375         1.661     0.184     1.776   17
      18   93.40      93.57     –0.17  0.375        –0.080     0.000    –0.077   22
      19   82.50      83.67     –1.17  0.375        –0.547     0.020    –0.534   2
      20   89.50      85.72      3.78  0.375         1.766     0.208     1.917   7
      21   85.60      86.97     –1.37  0.375        –0.641     0.027    –0.628   11
      22   87.40      86.52      0.88  0.375         0.411     0.011     0.399   14
      23   78.90      81.30     –2.40  0.375        –1.120     0.084    –1.130   18
      24   90.70      90.42      0.28  0.375         0.130     0.001     0.126   24
Note: Predicted values include block corrections.
```

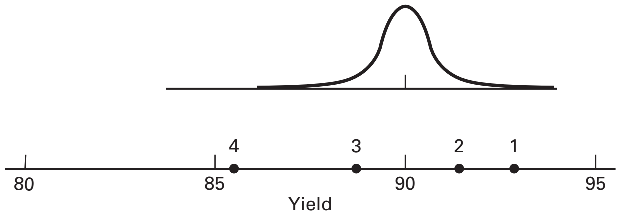
(*a*)

■ **FIGURE 4.2** Computer output for Example 4.1. (*a*) Design-Expert; (*b*) JMP

```
Oneway Analysis of Yield by Pressure
Block
Batch

Oneway Anova
Summary of Fit

Rsquare                        0.771218
Adj Rsquare                    0.649201
Root Mean Square Error         2.706612
Mean of Response               89.79583
Observations (or Sum Wgts)           24

Analysis of Variance
Source      DF   Sum of Squares    Mean Square   F Ratio   Prob > F
Pressure     3        178.17125        59.3904    8.1071     0.0019
Batch        5        192.25208        38.4504    5.2487     0.0055
Error       15        109.88625         7.3257
C.Total     23        480.30958

Means for Oneway Anova
Level     Number      Mean    Std. Error    Lower 95%    Upper 95%
8500           6   92.8167        1.1050       90.461       95.172
8700           6   91.6833        1.1050       89.328       94.039
8900           6   88.9167        1.1050       86.561       91.272
9100           6   85.7667        1.1050       83.411       88.122

Std. Error uses a pooled estimate of error variance

Block Means
Batch         Mean        Number
1          87.7000             4
2          89.7500             4
3          91.0000             4
4          90.5500             4
5          85.3250             4
6          94.4500             4

            (b)
```

■ **FIGURE 4.2** (*Continued*)

■ **FIGURE 4.3** **Mean yields for the four extrusion pressures relative to a scaled t distribution with a scale factor $\sqrt{MS_E/b} = \sqrt{7.33/6} = 1.10$**



4.1.2 Model Adequacy Checking

We have previously discussed the importance of checking the adequacy of the assumed model. Generally, we should be alert for potential problems with the normality assumption, unequal error variance by treatment or block, and block–treatment interaction. As in the completely randomized design, residual analysis is the major tool used in this diagnostic checking. The residuals for the randomized block design in Example 4.1 are listed at the bottom of the Design-Expert output in Figure 4.2.

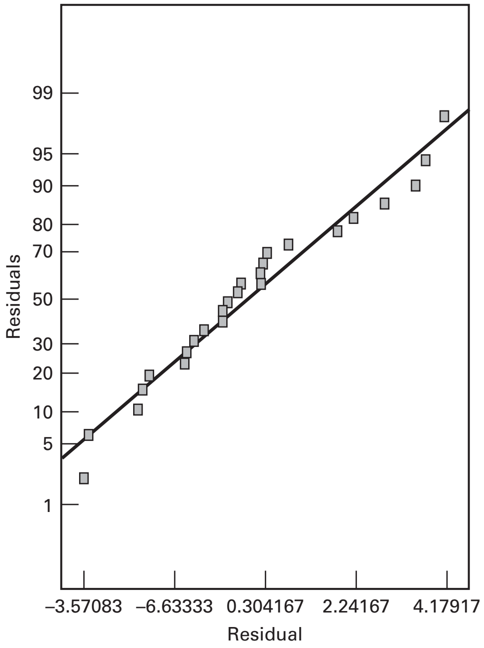
A normal probability plot of these residuals is shown in Figure 4.4. There is no severe indication of nonnormality, nor is there any evidence pointing to possible outliers. Figure 4.5 plots the residuals versus the fitted values $\hat{y}_{ij}$. There should be no relationship between the size of the residuals and the fitted values $\hat{y}_{ij}$. This plot reveals nothing of unusual interest. Figure 4.6 shows plots of the residuals by treatment (extrusion pressure) and by batch of resin or block. These plots are potentially very informative. If there is more scatter in the residuals for a particular treatment, it could indicate that this treatment produces more erratic response readings than the others. More scatter in the residuals for a particular block could indicate that the block is not homogeneous. However, in our example, Figure 4.6 gives no indication of inequality of variance by treatment, but there is an indication that there is less variability in the yield for batch 6. However, since all of the other residual plots are satisfactory, we will ignore this.

Sometimes the plot of residuals versus $\hat{y}_{ij}$ has a curvilinear shape; for example, there may be a tendency for negative residuals to occur with low $\hat{y}_{ij}$ values, positive residuals with intermediate $\hat{y}_{ij}$ values, and negative residuals with high $\hat{y}_{ij}$ values. This type of pattern is suggestive of **interaction** between blocks and treatments. If this pattern occurs, a transformation should be used in an effort to eliminate or minimize the interaction. In Section 5.3.7, we describe a statistical test that can be used to detect the presence of interaction in a randomized block design.

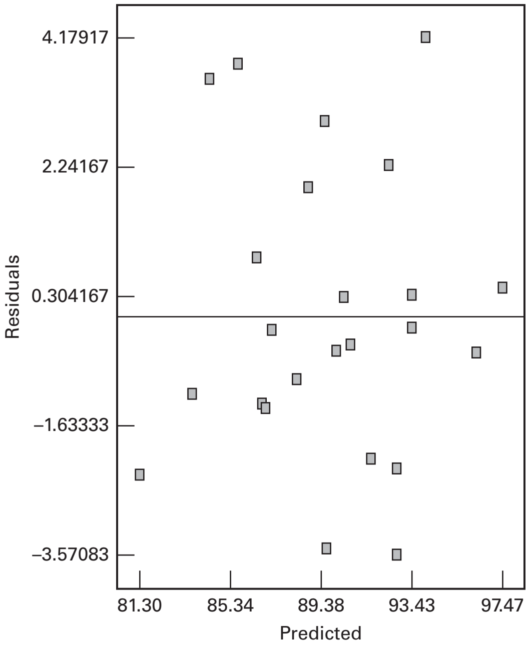
4.1.3 Some Other Aspects of the Randomized Complete Block Design

Additivity of the Randomized Block Model. The linear statistical model that we have used for the randomized block design

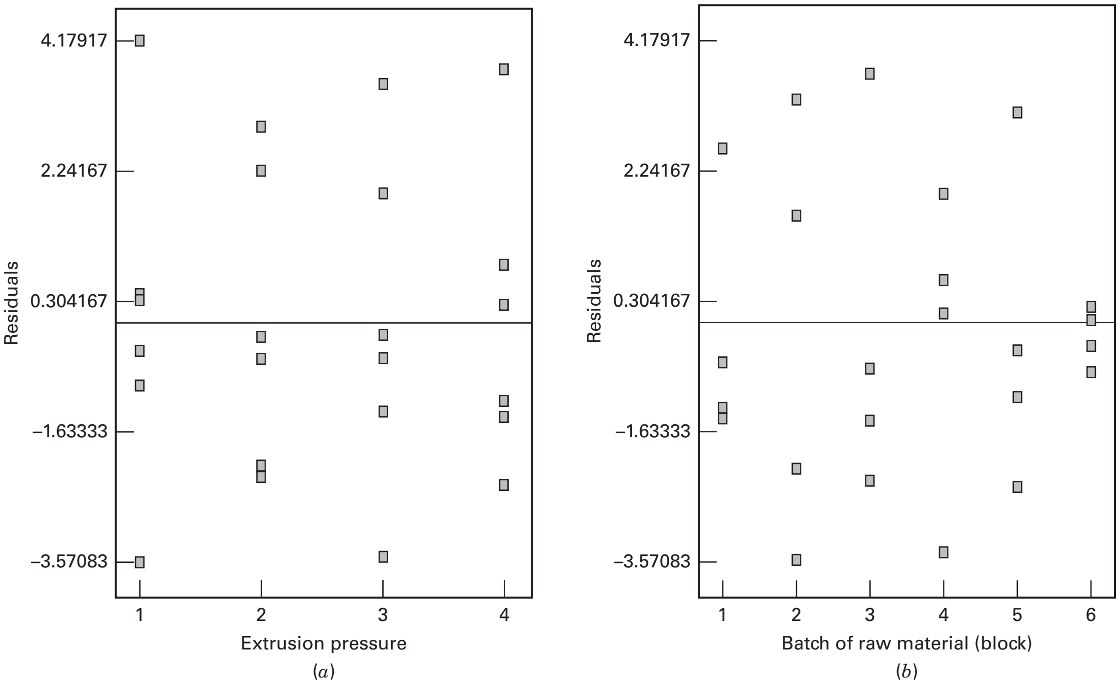
$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij}$$



■ **FIGURE 4.4 Normal probability plot of residuals for Example 4.1**



■ **FIGURE 4.5 Plot of residuals versus $\hat{y}_{ij}$ for Example 4.1**



■ **FIGURE 4.6** **Plot of residuals by extrusion pressure (treatment) and by batches of resin (block) for Example 4.1**

is completely **additive**. This says that, for example, if the first treatment causes the expected response to increase by five units ($\tau_1 = 5$) and if the first block increases the expected response by 2 units ($\beta_1 = 2$), the expected increase in response of *both* treatment 1 *and* block 1 together is $E(y_{11}) = \mu + \tau_1 + \beta_1 = \mu + 5 + 2 = \mu + 7$. In general, treatment 1 *always* increases the expected response by 5 units over the sum of the overall mean and the block effect.

Although this simple additive model is often useful, in some situations it is inadequate. Suppose, for example, that we are comparing four formulations of a chemical product using six batches of raw material; the raw material batches are considered blocks. If an impurity in batch 2 affects formulation 2 adversely, resulting in an unusually low yield, but does not affect the other formulations, an **interaction** between formulations (or treatments) and batches (or blocks) has occurred. Similarly, interactions between treatments and blocks can occur when the response is measured on the wrong scale. Thus, a relationship that is multiplicative in the original units, say

$$E(y_{ij}) = \mu\tau_i\beta_j$$

is linear or additive in a log scale since, for example,

$$\ln E(y_{ij}) = \ln \mu + \ln \tau_i + \ln \beta_j$$

or

$$E(y^*_{ij}) = \mu^* + \tau^*_i + \beta^*_j$$

Although this type of interaction can be eliminated by a transformation, not all interactions are so easily treated. For example, transformations do not eliminate the formulation–batch interaction discussed previously. Residual analysis and other diagnostic checking procedures can be helpful in detecting nonadditivity.

If interaction is present, it can seriously affect and possibly invalidate the analysis of variance. In general, the presence of interaction inflates the error mean square and may adversely affect the comparison of treatment means. In situations where both factors, as well as their possible interaction, are of interest, **factorial designs** must be used. These designs are discussed extensively in Chapters 5 through 9.

Random Treatments and Blocks. Our presentation of the randomized complete block design thus far has focused on the case when both the treatments and blocks were considered as fixed factors. There are many situations where either treatments or blocks (or both) are random factors. It is very common to find that the blocks are random. This is usually what the experimenter would like to do, because we would like for the conclusions from the experiment to be valid across the population of blocks that the ones selected for the experiments were sampled from. First, we consider the case where the treatments are fixed and the blocks are random. Equation 4.1 is still the appropriate statistical model, but now the block effects are random, that is, we assume that the β_j, $j = 1, 2, \ldots, b$ are $NID(0, \sigma_\beta^2)$ random variables. This is a special case of a mixed model (because it contains both fixed and random factors). In Chapters 13 and 14 we will discuss mixed models in more detail and provide several examples of situations where they occur. Our discussion here is limited to the RCBD.

Assuming that the RCBD model Equation 4.1 is appropriate, if the blocks are random and the treatments are fixed we can show that

$$\begin{aligned} E(y_{ij}) &= \mu + \tau_i, \qquad i = 1, 2, \ldots, a \\ V(y_{ij}) &= \sigma_\beta^2 + \sigma^2 \\ Cov(y_{ij}, y_{i'j'}) &= 0, \; j \neq j' \\ Cov(y_{ij}, y_{i'j}) &= \sigma_\beta^2 \; i \neq i' \end{aligned} \tag{4.14}$$

Thus, the variance of the observations is constant, the covariance between any two observations in different blocks is zero, but the covariance between two observations from the same block is σ_β^2. The expected mean squares from the usual ANOVA partitioning of the total sum of squares are

$$\begin{aligned} E(MS_{\text{Treatments}}) &= \sigma^2 + \frac{b\sum_{i=1}^{a}\tau_i^2}{a-1} \\ E(MS_{\text{Blocks}}) &= \sigma^2 + a\sigma_\beta^2 \\ E(MS_E) &= \sigma^2 \end{aligned} \tag{4.15}$$

The appropriate statistic for testing the null hypothesis of no treatment effects (all $\tau_i = 0$) is

$$F_0 = \frac{MS_{\text{Treatment}}}{MS_E}$$

which is exactly the same test statistic we used in the case where the blocks were fixed. Based on the expected mean squares, we can obtain an ANOVA-type estimator of the variance component for blocks as

$$\hat{\sigma}_\beta^2 = \frac{MS_{\text{Blocks}} - MS_E}{a} \tag{4.16}$$

For example, for the vascular graft experiment in Example 4.1 the estimate of σ_β^2 is

$$\hat{\sigma}_\beta^2 = \frac{MS_{\text{Blocks}} - MS_E}{a} = \frac{38.45 - 7.33}{4} = 7.78$$

This is a method-of-moments estimate and there is no simple way to find a confidence interval on the block variance component σ_β^2. The REML method would be preferred here. Table 4.6 is the JMP output for Example 4.1 assuming that blocks are random. The REML estimate of σ_β^2 is exactly the same as the ANOVA estimate, but REML automatically produces the standard error of the estimate (6.116215) and the approximate 95 percent confidence interval. JMP gives the test for the fixed effect (pressure), and the results are in agreement with those originally reported in Example 4.1. REML also produces the point estimate and confidence interval for the error variance σ^2. The ease with which confidence intervals can be constructed is a major reason why REML has been so widely adopted.

■ **TABLE 4.6**

JMP Output for Example 4.1 with Blocks Assumed Random

Response Y

Summary of Fit

RSquare	0.756688
RSquare Adj	0.720192
Root Mean Square Error	2.706612
Mean of Response	89.79583
Observations (or Sum Wgts)	24

REML Variance Component Estimates

Random Effect	Var Ratio	Var Component	Std Error	95% Lower	95% Upper	Pct of Total
Block	1.0621666	7.7811667	6.116215	−4.206394	19.768728	51.507
Residual		7.32575	2.6749857	3.9975509	17.547721	48.493
Total		15.106917				100.000

Covariance Matrix of Variance Component Estimates

Random Effect	Block	Residual
Block	37.408085	−1.788887
Residual	−1.788887	7.1555484

Fixed Effect Tests

Source	Nparm	DF	DFDen	F Ratio	Prob > F
Pressure	3	3	15	8.1071	0.0019*

*Significant at the 0.01 level.

Now consider a situation where there is an interaction between treatments and blocks. This could be accounted for by adding an interaction term to the original statistical model Equation 4.1. Let $(\tau\beta)_{ij}$ be the interaction effect of treatment I in block j. Then the model is

$$y_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ij} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \end{cases} \tag{4.17}$$

The interaction effect is assumed to be random because it involves the random block effects. If $\sigma^2_{\tau\beta}$ is the variance component for the block treatment interaction, then we can show that the expected mean squares are

$$\begin{aligned} E(MS_{\text{Treatments}}) &= \sigma^2 + \sigma^2_{\tau\beta} + \frac{b\sum_{i=1}^{a}\tau_i^2}{a-1} \\ E(MS_{\text{Blocks}}) &= \sigma^2 + a\sigma^2_{\beta} \\ E(MS_E) &= \sigma^2 + \sigma^2_{\tau\beta} \end{aligned} \tag{4.18}$$

From the expected mean squares, we see that the usual F-statistic $F = MS_{\text{Treatments}}/MS_E$ would be used to test for no treatment effects. So another advantage of the random block model is that the assumption of no interaction in the RCBD is not important. However, if blocks are fixed and there is an interaction, then the interaction effect is not in the expected mean square for treatments but it is in the error expected mean square, so there would not be a statistical test for the treatment effects.

Estimating Missing Values. When using the RCBD, sometimes an observation in one of the blocks is missing. This may happen because of carelessness or error or for reasons beyond our control, such as unavoidable damage to an experimental unit. A missing observation introduces a new problem into the analysis because treatments are no longer **orthogonal to blocks**; that is, every treatment does not occur in every block. There are two general approaches to the missing value problem. The first is an **approximate analysis** in which the missing observation is estimated and the usual analysis of variance is performed just as if the estimated observations were real data, with the error degrees of freedom reduced by 1. This approximate analysis is the subject of this section. The second is an **exact analysis**, which is discussed in Section 4.1.4.

Suppose the observation y_{ij} for treatment i in block j is missing. Denote the missing observation by x. As an illustration, suppose that in the vascular graft experiment of Example 4.1 there was a problem with the extrusion machine when the 8700 psi run was conducted in the fourth batch of material, and the observation y_{24} could not be obtained. The data might appear as in Table 4.7.

In general, we will let y'_{ij} represent the grand total with one missing observation, $y'_{i.}$ represent the total for the treatment with one missing observation, and $y'_{.j}$ be the total for the block with one missing observation. Suppose we wish to estimate the missing observation x so that x will have a minimum contribution to the error sum of squares. Because $SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$, this is equivalent to choosing x to minimize

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}^2 - \frac{1}{b}\sum_{i=1}^{a}\left(\sum_{j=1}^{b} y_{ij}\right)^2 - \frac{1}{a}\sum_{j=1}^{b}\left(\sum_{i=1}^{a} y_{ij}\right)^2 + \frac{1}{ab}\left(\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}\right)^2$$

or

$$SS_E = x^2 - \frac{1}{b}(y'_{i.} + x)^2 - \frac{1}{a}(y'_{.j} + x)^2 + \frac{1}{ab}(y'_{..} + x)^2 + R \qquad \textbf{(4.19)}$$

where R includes all terms not involving x. From $dSS_E/dx = 0$, we obtain

$$x = \frac{ay'_{i.} + by'_{.j} + y'_{..}}{(a-1)(b-1)} \qquad \textbf{(4.20)}$$

as the estimate of the missing observation.

For the data in Table 4.7, we find that $y'_{2.} = 455.4$, $y'_{.4} = 267.5$, and $y'_{..} = 2060.4$. Therefore, from Equation 4.16,

$$x \equiv y_{24} = \frac{4(455.4) + 6(267.5) - 2060.4}{(3)(5)} = 91.08$$

■ **TABLE 4.7**
Randomized Complete Block Design for the Vascular Graft Experiment with One Missing Value

Extrusion Pressures (PSI)	Batch of Resin (Block) 1	2	3	4	5	6	
8500	90.3	89.2	98.2	93.9	87.4	97.9	556.9
8700	92.5	89.5	90.6	x	87.0	95.8	455.4
8900	85.5	90.8	89.6	86.2	88.0	93.4	533.5
9100	82.5	89.5	85.6	87.4	78.9	90.7	514.6
Block totals	350.8	359.0	364.0	267.5	341.3	377.8	$y'_{..} = 2060.4$

■ **TABLE 4.8**

Approximate Analysis of Variance for Example 4.1 with One Missing Value

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	*P*-Value
Extrusion pressure	166.14	3	55.38	7.63	0.0029
Batches of raw material	189.52	5	37.90		
Error	101.70	14	7.26		
Total	457.36	23			

The usual analysis of variance may now be performed using $y_{24} = 91.08$ and reducing the error degrees of freedom by 1. The analysis of variance is shown in Table 4.8. Compare the results of this approximate analysis with the results obtained for the full data set (Table 4.4).

If several observations are missing, they may be estimated by writing the error sum of squares as a function of the missing values, differentiating with respect to each missing value, equating the results to zero, and solving the resulting equations. Alternatively, we may use Equation 4.20 iteratively to estimate the missing values. To illustrate the iterative approach, suppose that two values are missing. Arbitrarily estimate the first missing value, and then use this value along with the real data and Equation 4.20 to estimate the second. Now Equation 4.20 can be used to reestimate the first missing value, and following this, the second can be reestimated. This process is continued until convergence is obtained. In any missing value problem, the error degrees of freedom are reduced by one for each missing observation.

4.1.4 Estimating Model Parameters and the General Regression Significance Test

If both treatments and blocks are fixed, we may estimate the parameters in the RCBD model by least squares. Recall that the linear statistical model is

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \qquad \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \end{cases} \tag{4.21}$$

Applying the rules in Section 3.9.2 for finding the normal equations for an experimental design model, we obtain

$$\begin{array}{llllllllllllllllll}
\mu: & ab\hat{\mu} & + & b\hat{\tau}_1 & + & b\hat{\tau}_2 & + & \cdots & + & b\hat{\tau}_a & + & a\hat{\beta}_1 & + & a\hat{\beta}_2 & + \cdots + & a\hat{\beta}_b & = & y_{..} \\
\tau_1: & b\hat{\mu} & + & b\hat{\tau}_1 & & & & & & & + & \hat{\beta}_1 & + & \hat{\beta}_2 & + \cdots + & \hat{\beta}_b & = & y_{1.} \\
\tau_2: & b\hat{\mu} & & & + & b\hat{\tau}_2 & & & & & + & \hat{\beta}_1 & + & \hat{\beta}_2 & + \cdots + & \hat{\beta}_b & = & y_{2.} \\
\vdots & & & & & & & & & & & \vdots & & & & & & \vdots \\
\tau_a: & b\hat{\mu} & & & & & & & & b\hat{\tau}_a & + & \hat{\beta}_1 & + & \hat{\beta}_2 & + \cdots + & \hat{\beta}_b & = & y_{a.} \\
\beta_1: & a\hat{\mu} & + & \hat{\tau}_1 & + & \hat{\tau}_2 & + & \cdots & + & \hat{\tau}_a & + & a\hat{\beta}_1 & & & & & = & y_{.1} \\
\beta_2: & a\hat{\mu} & + & \hat{\tau}_1 & + & \hat{\tau}_2 & + & \cdots & + & \hat{\tau}_a & & & + & a\hat{\beta}_2 & & & = & y_{.2} \\
\vdots & & & & & & & & & & & \vdots & & & & & & \vdots \\
\beta_b: & a\hat{\mu} & + & \hat{\tau}_1 & + & \hat{\tau}_2 & + & \cdots & + & \hat{\tau}_a & & & & & + & a\hat{\beta}_b & = & y_{.b}
\end{array} \tag{4.22}$$

Notice that the second through the $(a + 1)$st equations in Equation 4.22 sum to the first normal equation, as do the last b equations. Thus, there are two linear dependencies in the normal equations, implying that two constraints must be imposed to solve Equation 4.22. The usual constraints are

$$\sum_{i=1}^{a} \hat{\tau}_i = 0 \qquad \sum_{j=1}^{b} \hat{\beta}_j = 0 \tag{4.23}$$

Using these constraints helps simplify the normal equations considerably. In fact, they become

$$\begin{aligned} ab\,\hat{\mu} &= y_{..} \\ b\,\hat{\mu} + b\hat{\tau}_i &= y_{i.} \quad i = 1, 2, \ldots, a \\ a\,\hat{\mu} + a\hat{\beta}_j &= y_{.j} \quad j = 1, 2, \ldots, b \end{aligned} \tag{4.24}$$

whose solution is

$$\begin{aligned} \hat{\mu} &= \bar{y}_{..} \\ \hat{\tau}_i &= \bar{y}_{i.} - \bar{y}_{..} \quad i = 1, 2, \ldots, a \\ \hat{\beta}_j &= \bar{y}_{.j} - \bar{y}_{..} \quad j = 1, 2, \ldots, b \end{aligned} \tag{4.25}$$

Using the solution to the normal equation in Equation 4.25, we may find the estimated or fitted values of y_{ij} as

$$\begin{aligned} \hat{y}_{ij} &= \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j \\ &= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) \\ &= \bar{y}_{i.} + \bar{y}_{.j} - \bar{y}_{..} \end{aligned}$$

This result was used previously in Equation 4.13 for computing the residuals from a randomized block design.

The general regression significance test can be used to develop the analysis of variance for the randomized complete block design. Using the solution to the normal equations given by Equation 4.25, the reduction in the sum of squares for fitting the **full model** is

$$\begin{aligned} R(\mu, \tau, \beta) &= \hat{\mu} y_{..} + \sum_{i=1}^{a} \hat{\tau}_i y_{i.} + \sum_{j=1}^{b} \hat{\beta}_j y_{.j} \\ &= \bar{y}_{..} y_{..} + \sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..}) y_{i.} + \sum_{j=1}^{b} (\bar{y}_{.j} - \bar{y}_{..}) y_{.j} \\ &= \frac{y_{..}^2}{ab} + \sum_{i=1}^{a} \bar{y}_{i.} y_{i.} - \frac{y_{..}^2}{ab} + \sum_{j=1}^{b} \bar{y}_{.j} y_{.j} - \frac{y_{..}^2}{ab} \\ &= \sum_{i=1}^{a} \frac{y_{i.}^2}{b} + \sum_{j=1}^{b} \frac{y_{.j}^2}{a} - \frac{y_{..}^2}{ab} \end{aligned}$$

with $a + b - 1$ degrees of freedom, and the error sum of squares is

$$\begin{aligned} SS_E &= \sum_{i=1}^{a} \sum_{j=1}^{b} y_{ij}^2 - R(\mu, \tau, \beta) \\ &= \sum_{i=1}^{a} \sum_{j=1}^{b} y_{ij}^2 - \sum_{i=1}^{a} \frac{y_{i.}^2}{b} - \sum_{j=1}^{b} \frac{y_{.j}^2}{a} + \frac{y_{..}^2}{ab} \\ &= \sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 \end{aligned}$$

with $(a - 1)(b - 1)$ degrees of freedom. Compare this last equation with SS_E in Equation 4.7.

To test the hypothesis $H_0 : \tau_i = 0$, the **reduced model** is

$$y_{ij} = \mu + \beta_j + \in_{ij}$$

which is just a single-factor analysis of variance. By analogy with Equation 3.5, the reduction in the sum of squares for fitting the reduced model is

$$R(\mu, \beta) = \sum_{j=1}^{b} \frac{y_{.j}^2}{a}$$

which has b degrees of freedom. Therefore, the sum of squares due to $\{\tau_i\}$ after fitting μ and $\{\beta_j\}$ is

$$\begin{aligned} R(\tau|\mu, \beta) &= R(\mu, \tau, \beta) - R(\mu, \beta) \\ &= R(\text{full model}) - R(\text{reduced model}) \\ &= \sum_{i=1}^{a} \frac{y_{i.}^2}{b} + \sum_{j=1}^{b} \frac{y_{.j}^2}{a} - \frac{y_{..}^2}{ab} - \sum_{j=1}^{b} \frac{y_{.j}^2}{a} \\ &= \sum_{i=1}^{a} \frac{y_{i.}^2}{b} - \frac{y_{..}^2}{ab} \end{aligned}$$

which we recognize as the treatment sum of squares with $a - 1$ degrees of freedom (Equation 4.10).

The block sum of squares is obtained by fitting the **reduced model**

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

which is also a single-factor analysis. Again, by analogy with Equation 3.5, the reduction in the sum of squares for fitting this model is

$$R(\mu, \tau) = \sum_{i=1}^{a} \frac{y_{i.}^2}{b}$$

with a degrees of freedom. The sum of squares for blocks $\{\beta_j\}$ after fitting μ and $\{\tau_i\}$ is

$$\begin{aligned} R(\beta|\mu, \tau) &= R(\mu, \tau, \beta) - R(\mu, \tau) \\ &= \sum_{i=1}^{a} \frac{y_{i.}^2}{b} + \sum_{j=1}^{b} \frac{y_{.j}^2}{a} - \frac{y_{..}^2}{ab} - \sum_{i=1}^{a} \frac{y_{i.}^2}{b} \\ &= \sum_{j=1}^{b} \frac{y_{.j}^2}{a} - \frac{y_{..}^2}{ab} \end{aligned}$$

with $b - 1$ degrees of freedom, which we have given previously as Equation 4.11.

We have developed the sums of squares for treatments, blocks, and error in the randomized complete block design using the general regression significance test. Although we would not ordinarily use the general regression significance test to actually analyze data in a randomized complete block, the procedure occasionally proves useful in more general randomized block designs, such as those discussed in Section 4.4.

Exact Analysis of the Missing Value Problem. In Section 4.1.3 an approximate procedure for dealing with missing observations in the RCBD was presented. This approximate analysis consists of estimating the missing value so that the error mean square is minimized. It can be shown that the approximate analysis produces a biased mean square for treatments in the sense that $E(MS_{\text{Treatments}})$ is larger than $E(MS_E)$ if the null hypothesis is true. Consequently, too many significant results are reported.

The missing value problem may be analyzed exactly by using the general regression significance test. The missing value causes the design to be **unbalanced**, and because all the treatments do not occur in all blocks,

■ **TABLE 4.9**
Latin Square Design for the Rocket Propellant Problem

Batches of Raw Material	Operators				
	1	2	3	4	5
1	$A = 24$	$B = 20$	$C = 19$	$D = 24$	$E = 24$
2	$B = 17$	$C = 24$	$D = 30$	$E = 27$	$A = 36$
3	$C = 18$	$D = 38$	$E = 26$	$A = 27$	$B = 21$
4	$D = 26$	$E = 31$	$A = 26$	$B = 23$	$C = 22$
5	$E = 22$	$A = 30$	$B = 20$	$C = 29$	$D = 31$

we say that the treatments and blocks are not **orthogonal**. This method of analysis is also used in more general types of randomized block designs; it is discussed further in Section 4.4. Many computer packages will perform this analysis.

4.2 The Latin Square Design

In Section 4.1 we introduced the randomized complete block design as a design to reduce the residual error in an experiment by removing variability due to a known and controllable nuisance variable. There are several other types of designs that utilize the blocking principle. For example, suppose that an experimenter is studying the effects of five different formulations of a rocket propellant used in aircrew escape systems on the observed burning rate. Each formulation is mixed from a batch of raw material that is only large enough for five formulations to be tested. Furthermore, the formulations are prepared by several operators, and there may be substantial differences in the skills and experience of the operators. Thus, it would seem that there are two nuisance factors to be "averaged out" in the design: batches of raw material and operators. The appropriate design for this problem consists of testing each formulation exactly once in each batch of raw material and for each formulation to be prepared exactly once by each of five operators. The resulting design, shown in Table 4.9, is called a **Latin square design**. Notice that the design is a square arrangement and that the five formulations (or treatments) are denoted by the Latin letters A, B, C, D, and E; hence the name Latin square. We see that both batches of raw material (rows) and operators (columns) are orthogonal to treatments.

The Latin square design is used to eliminate two nuisance sources of variability; that is, it systematically allows blocking in two directions. Thus, the rows and columns actually represent **two restrictions on randomization**. In general, a Latin square for p factors, or a $p \times p$ Latin square, is a square containing p rows and p columns. Each of the resulting p^2 cells contains one of the p letters that corresponds to the treatments, and each letter occurs once and only once in each row and column. Some examples of Latin squares are

4 × 4	**5 × 5**	**6 × 6**
$A\ B\ D\ C$	$A\ D\ B\ E\ C$	$A\ D\ C\ E\ B\ F$
$B\ C\ A\ D$	$D\ A\ C\ B\ E$	$B\ A\ E\ C\ F\ D$
$C\ D\ B\ A$	$C\ B\ E\ D\ A$	$C\ E\ D\ F\ A\ B$
$D\ A\ C\ B$	$B\ E\ A\ C\ D$	$D\ C\ F\ B\ E\ A$
	$E\ C\ D\ A\ B$	$F\ B\ A\ D\ C\ E$
		$E\ F\ B\ A\ D\ C$

Latin squares are closely related to a popular puzzle called a sudoku puzzle that originated in Japan (sudoku means "single number" in Japanese). The puzzle typically consists of a 9 × 9 grid, with nine additional 3 × 3 blocks contained within. A few of the squares contain numbers and the others are blank. The goal is to fill the blanks with the integers from 1 to 9 so that each row, each column, and each of the nine 3 × 3 blocks making up the grid contains just one of each of the nine integers. The additional constraint that a standard 9 × 9 sudoku puzzle have 3 × 3 blocks that also contain each of the nine integers reduces the large number of possible 9 × 9 Latin squares to a smaller but still quite large number, approximately 6×10^{21}.

Depending on the number of clues and the size of the grid, sudoku puzzles can be extremely difficult to solve. Solving an $n \times n$ sudoku puzzle belongs to a class of computational problems called *NP*-complete (the *NP* refers to nonpolynomial computing time). An *NP*-complete problem is one for which it's relatively easy to check whether a particular answer is correct but may require an impossibly long time to solve by any simple algorithm as n gets larger.

Solving a sudoku puzzle is also equivalent to "coloring" a graph—an array of points (vertices) and lines (edges) in a particular way. In this case, the graph has 81 vertices, one for each cell of the grid. Depending on the puzzle, only certain pairs of vertices are joined by an edge. Given that some vertices have already been assigned a "color" (chosen from the nine number possibilities), the problem is to "color" the remaining vertices so that any two vertices joined by an edge don't have the same "color."

The **statistical model** for a Latin square is

$$y_{ijk} = \mu + \alpha_i + \tau_j + \beta_k + \epsilon_{ijk} \begin{cases} i = 1, 2, \ldots, p \\ j = 1, 2, \ldots, p \\ k = 1, 2, \ldots, p \end{cases} \tag{4.26}$$

where y_{ijk} is the observation in the ith row and kth column for the jth treatment, μ is the overall mean, α_i is the ith row effect, τ_j is the jth treatment effect, β_k is the kth column effect, and ϵ_{ijk} is the random error. Note that this is an **effects model**. The model is completely **additive**; that is, there is no interaction between rows, columns, and treatments. Because there is only one observation in each cell, only two of the three subscripts $i, j,$ and k are needed to denote a particular observation. For example, referring to the rocket propellant problem in Table 4.8, if $i = 2$ and $k = 3$, we automatically find $j = 4$ (formulation D), and if $i = 1$ and $j = 3$ (formulation C), we find $k = 3$. This is a consequence of each treatment appearing exactly once in each row and column.

The analysis of variance consists of partitioning the total sum of squares of the $N = p^2$ observations into components for rows, columns, treatments, and error, for example,

$$SS_T = SS_{\text{Rows}} + SS_{\text{Columns}} + SS_{\text{Treatments}} + SS_E \tag{4.27}$$

with respective degrees of freedom

$$p^2 - 1 = p - 1 + p - 1 + p - 1 + (p - 2)(p - 1)$$

Under the usual assumption that $\in_{ijk}$ is NID $(0, \sigma^2)$, each sum of squares on the right-hand side of Equation 4.27 is, upon division by σ^2, an independently distributed chi-square random variable. The appropriate statistic for testing for no differences in treatment means is

$$F_0 = \frac{MS_{\text{Treatments}}}{MS_E}$$

which is distributed as $F_{p-1,(p-2)(p-1)}$ under the null hypothesis. We may also test for no row effect and no column effect by forming the ratio of MS_{Rows} or MS_{Columns} to MS_E. However, because the rows and columns represent restrictions on randomization, these tests may not be appropriate.

The computational procedure for the ANOVA in terms of treatment, row, and column totals is shown in Table 4.10. From the computational formulas for the sums of squares, we see that the analysis is a simple extension of the RCBD, with the sum of squares resulting from rows obtained from the row totals.

■ **TABLE 4.10**

Analysis of Variance for the Latin Square Design

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$SS_{\text{Treatments}} = \frac{1}{p}\sum_{j=1}^{p} y_{.j.}^2 - \frac{y_{..}^2}{N}$	$p - 1$	$\frac{SS_{\text{Treatments}}}{p-1}$	$F_0 = \frac{MS_{\text{Treatments}}}{MS_E}$
Rows	$SS_{\text{Rows}} = \frac{1}{p}\sum_{i=1}^{p} y_{i..}^2 - \frac{y_{...}^2}{N}$	$p - 1$	$\frac{SS_{\text{Rows}}}{p-1}$	
Columns	$SS_{\text{Columns}} = \frac{1}{p}\sum_{k=1}^{p} y_{..k}^2 - \frac{y_{...}^2}{N}$	$p - 1$	$\frac{SS_{\text{Columns}}}{p-1}$	
Error	SS_E (by subtraction)	$(p - 2)(p - 1)$	$\frac{SS_E}{(p-2)(p-1)}$	
Total	$SS_T = \sum_i \sum_j \sum_k y_{ijk}^2 - \frac{y_{...}^2}{N}$	$p^2 - 1$		

EXAMPLE 4.2

Consider the rocket propellant problem previously described, where both batches of raw material and operators represent randomization restrictions. The design for this experiment, shown in Table 4.8, is a 5 × 5 Latin square. After coding by subtracting 25 from each observation, we have the data in Table 4.11. The sums of squares for the total, batches (rows), and operators (columns) are computed as follows:

$$SS_T = \sum_i \sum_j \sum_k y_{ijk}^2 - \frac{y_{...}^2}{N}$$

$$= 680 - \frac{(10)^2}{25} = 676.00$$

$$SS_{\text{Batches}} = \frac{1}{p}\sum_{i=1}^{p} y_{i..}^2 - \frac{y_{...}^2}{N}$$

$$= \frac{1}{5}\left[(-14)^2 + 9^2 + 5^2 + 3^2 + 7^2\right]$$

$$- \frac{(10)^2}{25} = 68.00$$

$$SS_{\text{Operators}} = \frac{1}{p}\sum_{k=1}^{p} y_{..k}^2 - \frac{y_{...}^2}{N}$$

$$= \frac{1}{5}\left[(-18)^2 + 18^2 + (-4)^2 + 5^2 + 9^2\right]$$

$$- \frac{(10)^2}{25} = 150.00$$

The totals for the treatments (Latin letters) are

Latin Letter	Treatment Total
A	$y_{.1.} = 18$
B	$y_{.2.} = -24$
C	$y_{.3.} = -13$
D	$y_{.4.} = 24$
E	$y_{.5.} = 5$

The sum of squares resulting from the formulations is computed from these totals as

$$SS_{\text{Formulations}} = \frac{1}{p}\sum_{j=1}^{p} y_{.j.}^2 - \frac{y_{...}^2}{N}$$

$$= \frac{18^2 + (-24)^2 + (-13)^2 + 24^2 + 5^2}{5}$$

$$- \frac{(10)^2}{25} = 330.00$$

The error sum of squares is found by subtraction

$$SS_E = SS_T - SS_{\text{Batches}} - SS_{\text{Operators}} - SS_{\text{Formulations}}$$

$$= 676.00 - 68.00 - 150.00 - 330.00 = 128.00$$

The analysis of variance is summarized in Table 4.12. We conclude that there is a significant difference in the mean burning rate generated by the different rocket propellant formulations. There is also an indication that differences between operators exist, so blocking on this factor was a good precaution. There is no strong evidence of a difference between batches of raw material, so it seems that in this particular experiment we were unnecessarily concerned about this source of variability. However, blocking on batches of raw material is usually a good idea.

■ **TABLE 4.11**
Coded Data for the Rocket Propellant Problem

Batches of	Operators					
Raw Material	1	2	3	4	5	$y_{i..}$
1	$A = -1$	$B = -5$	$C = -6$	$D = -1$	$E = -1$	−14
2	$B = -8$	$C = -1$	$D = 5$	$E = 2$	$A = 11$	9
3	$C = -7$	$D = 13$	$E = 1$	$A = 2$	$B = -4$	5
4	$D = 1$	$E = 6$	$A = 1$	$B = -2$	$C = -3$	3
5	$E = -3$	$A = 5$	$B = -5$	$C = 4$	$D = 6$	7
$y_{..k}$	−18	18	−4	5	9	$10 = y_{...}$

■ **TABLE 4.12**
Analysis of Variance for the Rocket Propellant Experiment

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	*P*-Value
Formulations	330.00	4	82.50	7.73	0.0025
Batches of raw material	68.00	4	17.00		
Operators	150.00	4	37.50		
Error	128.00	12	10.67		
Total	676.00	24			

As in any design problem, the experimenter should investigate the adequacy of the model by inspecting and plotting the residuals. For a Latin square, the residuals are given by

$$\begin{aligned} e_{ijk} &= y_{ijk} - \hat{y}_{ijk} \\ &= y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...} \end{aligned}$$

The reader should find the residuals for Example 4.2 and construct appropriate plots.

A Latin square in which the first row and column consists of the letters written in alphabetical order is called a **standard Latin square**, which is the design shown in Example 4.3. A standard Latin square can always be obtained by writing the first row in alphabetical order and then writing each successive row as the row of letters just above shifted one place to the left. Table 4.13 summarizes several important facts about Latin squares and standard Latin squares.

As with any experimental design, the observations in the Latin square should be taken in random order. The proper randomization procedure is to select the particular square employed at random. As we see in Table 4.13, there are a large number of Latin squares of a particular size, so it is impossible to enumerate all the squares and

■ **TABLE 4.13**
Standard Latin Squares and Number of Latin Squares of Various Sizes[a]

Size	3 × 3	4 × 4	5 × 5	6 × 6	7 × 7	*p* × *p*
Examples of standard squares	*A B C* *B C A* *C A B*	*A B C D* *B C D A* *C D A B* *D A B C*	*A B C D E* *B A E C D* *C D A E B* *D E B A C* *E C D B A*	*A B C D E F* *B C F A D E* *C F B E A D* *D E A B F C* *E A D F C B* *F D E C B A*	*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*	*ABC . . . P* *BCD . . . A* *CDE . . . B* ⋮ *PAB . . .* (*P* − 1)
Number of standard squares	1	4	56	9408	16,942,080	—
Total number of Latin squares	12	576	161,280	818,851,200	61,479,419,904,000	*p*!(*p* − 1)!× (number of standard squares)

[a] Some of the information in this table is found in Fisher and Yates (1953). Little is known about the properties of Latin squares larger than 7 × 7.

select one randomly. The usual procedure is to select an arbitrary Latin square from a table of such designs, as in Fisher and Yates (1953), or start with a standard square, and then arrange the order of the rows, columns, and letters at random. This is discussed more completely in Fisher and Yates (1953).

Occasionally, one observation in a Latin square is missing. For a $p \times p$ Latin square, the missing value may be estimated by

$$y_{ijk} = \frac{p(y'_{i..} + y'_{.j.} + y'_{..k}) - 2y'_{...}}{(p-2)(p-1)} \tag{4.28}$$

where the primes indicate totals for the row, column, and treatment with the missing value, and $y'_{...}$ is the grand total with the missing value.

Latin squares can be useful in situations where the rows and columns represent factors the experimenter actually wishes to study and where there are no randomization restrictions. Thus, three factors (rows, columns, and letters), each at p levels, can be investigated in only p^2 runs. This design assumes that there is no interaction between the factors. More will be said later on the subject of interaction.

Replication of Latin Squares. A disadvantage of small Latin squares is that they provide a relatively small number of error degrees of freedom. For example, a 3 × 3 Latin square has only two error degrees of freedom, a 4 × 4 Latin square has only six error degrees of freedom, and so forth. When small Latin squares are used, it is frequently desirable to replicate them to increase the error degrees of freedom.

A Latin square may be replicated in several ways. To illustrate, suppose that the 5 × 5 Latin square used in Example 4.3 is replicated n times. This could have been done as follows:

1. Use the same batches and operators in each replicate.
2. Use the same batches but different operators in each replicate (or, equivalently, use the same operators but different batches).
3. Use different batches and different operators.

The analysis of variance depends on the method of replication.

Consider case 1, where the same levels of the row and column blocking factors are used in each replicate. Let y_{ijkl} be the observation in row i, treatment j, column k, and replicate l. There are $N = np^2$ total observations. The ANOVA is summarized in Table 4.14.

Now consider case 2 and assume that new batches of raw material but the same operators are used in each replicate. Thus, there are now five new rows (in general, p new rows) within each replicate. The ANOVA is summarized in Table 4.15. Note that the source of variation for the rows really measures the variation between rows within the n replicates.

■ **TABLE 4.14**

Analysis of Variance for a Replicated Latin Square, Case 1

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$\frac{1}{np}\sum_{j=1}^{p} y_{.j..}^2 - \frac{y_{....}^2}{N}$	$p - 1$	$\frac{SS_{\text{Treatments}}}{p-1}$	$\frac{MS_{\text{Treatments}}}{MS_E}$
Rows	$\frac{1}{np}\sum_{i=1}^{p} y_{i...}^2 - \frac{y_{....}^2}{N}$	$p - 1$	$\frac{SS_{\text{Rows}}}{p-1}$	
Columns	$\frac{1}{np}\sum_{k=1}^{p} y_{..k.}^2 - \frac{y_{....}^2}{N}$	$p - 1$	$\frac{SS_{\text{Columns}}}{p-1}$	
Replicates	$\frac{1}{p^2}\sum_{l=1}^{n} y_{...l}^2 - \frac{y_{....}^2}{N}$	$n - 1$	$\frac{SS_{\text{Replicates}}}{n-1}$	
Error	Subtraction	$(p - 1)[n(p + 1) - 3]$	$\frac{SS_E}{(p-1)[n(p+1)-3]}$	
Total	$\sum\sum\sum\sum y_{ijkl}^2 - \frac{y_{....}^2}{N}$	$np^2 - 1$		

■ **TABLE 4.15**

Analysis of Variance for a Replicated Latin Square, Case 2

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$\frac{1}{np}\sum_{j=1}^{p} y_{.j..}^2 - \frac{y_{....}^2}{N}$	$p - 1$	$\frac{SS_{\text{Treatments}}}{p-1}$	$\frac{MS_{\text{Treatments}}}{MS_E}$
Rows	$\frac{1}{p}\sum_{l=1}^{n}\sum_{i=1}^{p} y_{i..l}^2 - \sum_{l=1}^{n}\frac{y_{...l}^2}{p^2}$	$n(p - 1)$	$\frac{SS_{\text{Rows}}}{n(p-1)}$	
Columns	$\frac{1}{np}\sum_{k=1}^{p} y_{..k.}^2 - \frac{y_{....}^2}{N}$	$p - 1$	$\frac{SS_{\text{Columns}}}{p-1}$	
Replicates	$\frac{1}{p^2}\sum_{l=1}^{n} y_{...l}^2 - \frac{y_{....}^2}{N}$	$n - 1$	$\frac{SS_{\text{Replicates}}}{n-1}$	
Error	Subtraction	$(p - 1)(np - 1)$	$\frac{SS_E}{(p-1)(np-1)}$	
Total	$\sum_i\sum_j\sum_k\sum_l y_{ijkl}^2 - \frac{y_{....}^2}{N}$	$np^2 - 1$		

■ **TABLE 4.16**
Analysis of Variance for a Replicated Latin Square, Case 3

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$\frac{1}{np}\sum_{j=1}^{p} y_{.j..}^2 - \frac{y_{....}^2}{N}$	$p - 1$	$\frac{SS_{\text{Treatments}}}{p-1}$	$\frac{MS_{\text{Treatments}}}{MS_E}$
Rows	$\frac{1}{p}\sum_{l=1}^{n}\sum_{i=1}^{p} y_{i..l}^2 - \sum_{l=1}^{n}\frac{y_{...l}^2}{p^2}$	$n(p-1)$	$\frac{SS_{\text{Rows}}}{n(p-1)}$	
Columns	$\frac{1}{p}\sum_{l=1}^{n}\sum_{k=1}^{p} y_{..kl}^2 - \sum_{l=1}^{n}\frac{y_{...l}^2}{p^2}$	$n(p-1)$	$\frac{SS_{\text{Columns}}}{n(p-1)}$	
Replicates	$\frac{1}{p^2}\sum_{l=1}^{n} y_{...l}^2 - \frac{y_{....}^2}{N}$	$n - 1$	$\frac{SS_{\text{Replicates}}}{n-1}$	
Error	Subtraction	$(p-1)[n(p-1)-1]$	$\frac{SS_E}{(p-1)[n(p-1)-1]}$	
Total	$\sum_i\sum_j\sum_k\sum_l y_{ijkl}^2 - \frac{y_{....}^2}{N}$	$np^2 - 1$		

Finally, consider case 3, where new batches of raw material and new operators are used in each replicate. Now the variation that results from both the rows and columns measures the variation resulting from these factors within the replicates. The ANOVA is summarized in Table 4.16.

There are other approaches to analyzing replicated Latin squares that allow some interactions between treatments and squares (refer to Problem 4.35).

Crossover Designs and Designs Balanced for Residual Effects. Occasionally, one encounters a problem in which time periods are a factor in the experiment. In general, there are p treatments to be tested in p time periods using np experimental units. For example, a human performance analyst is studying the effect of two replacement fluids on dehydration in 20 subjects. In the first period, half of the subjects (chosen at random) are given fluid A and the other half fluid B. At the end of the period, the response is measured and a period of time is allowed to pass in which any physiological effect of the fluids is eliminated. Then the experimenter has the subjects who took fluid A take fluid B and those who took fluid B take fluid A. This design is called a **crossover design**. It is analyzed as a set of 10 Latin squares with two rows (time periods) and two treatments (fluid types). The two columns in each of the 10 squares correspond to subjects.

The layout of this design is shown in Figure 4.7. Notice that the rows in the Latin square represent the time periods and the columns represent the subjects. The 10 subjects who received fluid A first (1, 4, 6, 7, 9, 12, 13, 15, 17, and 19) are randomly determined.

An abbreviated analysis of variance is summarized in Table 4.17. The subject sum of squares is computed as the corrected sum of squares among the 20 subject totals, the period sum of squares is the corrected sum of squares

	Latin Squares																			
	I		II		III		IV		V		VI		VII		VIII		IX		X	
Subject	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Period 1	A	B	B	A	B	A	A	B	A	B	B	A	A	B	A	B	A	B	A	B
Period 2	B	A	A	B	A	B	B	A	B	A	A	B	B	A	B	A	B	A	B	A

■ **FIGURE 4.7** **A crossover design**

■ **TABLE 4.17**
Analysis of Variance for the Crossover Design in Figure 4.7

Source of Variation	Degrees of Freedom
Subjects (columns)	19
Periods (rows)	1
Fluids (letters)	1
Error	18
Total	39

among the rows, and the fluid sum of squares is computed as the corrected sum of squares among the letter totals. For further details of the statistical analysis of these designs, see Cochran and Cox (1957), John (1971), and Anderson and McLean (1974).

It is also possible to employ Latin square type designs for experiments in which the treatments have a **residual effect**—that is, for example, if the data for fluid *B* in period 2 still reflected some effect of fluid *A* taken in period 1. Designs balanced for residual effects are discussed in detail by Cochran and Cox (1957) and John (1971).

4.3 The Graeco-Latin Square Design

Consider a $p \times p$ Latin square, and superimpose on it a second $p \times p$ Latin square in which the treatments are denoted by Greek letters. If the two squares when superimposed have the property that each Greek letter appears once and only once with each Latin letter, the two Latin squares are said to be **orthogonal**, and the design obtained is called a **Graeco-Latin square**. An example of a 4 × 4 Graeco-Latin square is shown in Table 4.18.

The Graeco-Latin square design can be used to control systematically three sources of extraneous variability, that is, to block in *three* directions. The design allows investigation of four factors (rows, columns, Latin letters, and Greek letters), each at p levels in only p^2 runs. Graeco-Latin squares exist for all $p \geq 3$ except $p = 6$.

The statistical model for the Graeco-Latin square design is

$$y_{ijkl} = \mu + \theta_i + \tau_j + \omega_k + \Psi_l + \epsilon_{ijkl} \begin{cases} i = 1, 2, \ldots, p \\ j = 1, 2, \ldots, p \\ k = 1, 2, \ldots, p \\ l = 1, 2, \ldots, p \end{cases} \tag{4.29}$$

where y_{ijkl} is the observation in row i and column l for Latin letter j and Greek letter k, θ_i is the effect of the ith row, τ_j is the effect of Latin letter treatment j, ω_k is the effect of Greek letter treatment k, Ψ_l is the effect of column l, and ϵ_{ijkl} is an NID$(0, \sigma^2)$ random error component. Only two of the four subscripts are necessary to completely identify an observation.

■ **TABLE 4.18**
4 × 4 Graeco-Latin Square Design

	Column			
Row	**1**	**2**	**3**	**4**
1	$A\alpha$	$B\beta$	$C\gamma$	$D\delta$
2	$B\delta$	$A\gamma$	$D\beta$	$C\alpha$
3	$C\beta$	$D\alpha$	$A\delta$	$B\gamma$
4	$D\gamma$	$C\delta$	$B\alpha$	$A\beta$

■ **TABLE 4.19**
Analysis of Variance for a Graeco-Latin Square Design

Source of Variation	Sum of Squares	Degrees of Freedom
Latin letter treatments	$SS_L = \frac{1}{p}\sum_{j=1}^{p} y_{.j..}^2 - \frac{y_{....}^2}{N}$	$p - 1$
Greek letter treatments	$SS_G = \frac{1}{p}\sum_{k=1}^{p} y_{..k.}^2 - \frac{y_{....}^2}{N}$	$p - 1$
Rows	$SS_{\text{Rows}} = \frac{1}{p}\sum_{i=1}^{p} y_{i...}^2 - \frac{y_{....}^2}{N}$	$p - 1$
Columns	$SS_{\text{Columns}} = \frac{1}{p}\sum_{l=1}^{p} y_{...l}^2 - \frac{y_{....}^2}{N}$	$p - 1$
Error	SS_E (by subtraction)	$(p - 3)(p - 1)$
Total	$SS_T = \sum_i \sum_j \sum_k \sum_l y_{ijkl}^2 - \frac{y_{....}^2}{N}$	$p^2 - 1$

The analysis of variance is very similar to that of a Latin square. Because the Greek letters appear exactly once in each row and column and exactly once with each Latin letter, the factor represented by the Greek letters is orthogonal to rows, columns, and Latin letter treatments. Therefore, a sum of squares due to the Greek letter factor may be computed from the Greek letter totals, and the experimental error is further reduced by this amount. The computational details are illustrated in Table 4.19. The null hypotheses of equal row, column, Latin letter, and Greek letter treatments would be tested by dividing the corresponding mean square by mean square error. The rejection region is the upper tail point of the $F_{p-1,(p-3)(p-1)}$ distribution.

EXAMPLE 4.3

Suppose that in the rocket propellant experiment of Example 4.2 an additional factor, test assemblies, could be of importance. Let there be five test assemblies denoted by the Greek letters $\alpha, \beta, \gamma, \delta,$ and ϵ. The resulting 5×5 Graeco-Latin square design is shown in Table 4.20.

Notice that because the totals for batches of raw material (rows), operators (columns), and formulations (Latin letters) are identical to those in Example 4.2, we have

$$SS_{\text{Batches}} = 68.00, \quad SS_{\text{Operators}} = 150.00, \quad \text{and} \quad SS_{\text{Formulations}} = 330.00$$

The totals for the test assemblies (Greek letters) are

Greek Letter	Test Assembly Total
α	$y_{..1.} = 10$
β	$y_{..2.} = -6$
γ	$y_{..3.} = -3$
δ	$y_{..4.} = -4$
ϵ	$y_{..5.} = 13$

Thus, the sum of squares due to the test assemblies is

$$\begin{aligned} SS_{\text{Assemblies}} &= \frac{1}{p}\sum_{k=1}^{p} y_{..k.}^2 - \frac{y_{....}^2}{N} \\ &= \frac{1}{5}[10^2 + (-6)^2 + (-3)^2 \\ &\quad + (-4)^2 + 13^2] - \frac{(10)^2}{25} = 62.00 \end{aligned}$$

The complete ANOVA is summarized in Table 4.21. Formulations are significantly different at 1 percent. In comparing Tables 4.21 and 4.12, we observe that removing the variability due to test assemblies has decreased the experimental error. However, in decreasing the experimental error, we have also reduced the error degrees of freedom from 12 (in the Latin square design of Example 4.2) to 8. Thus, our estimate of error has fewer degrees of freedom, and the test may be less sensitive.

■ **TABLE 4.20**
Graeco-Latin Square Design for the Rocket Propellant Problem

Batches of Raw Material	Operators 1	2	3	4	5	$y_{i...}$
1	$A\alpha = -1$	$B\gamma = -5$	$C\epsilon = -6$	$D\beta = -1$	$E\delta = -1$	−14
2	$B\beta = -8$	$C\delta = -1$	$D\alpha = 5$	$E\gamma = 2$	$A\epsilon = 11$	9
3	$C\gamma = -7$	$D\epsilon = 13$	$E\beta = 1$	$A\delta = 2$	$B\alpha = -4$	5
4	$D\delta = 1$	$E\alpha = 6$	$A\gamma = 1$	$B\epsilon = -2$	$C\beta = -3$	3
5	$E\epsilon = -3$	$A\beta = 5$	$B\delta = -5$	$C\alpha = 4$	$D\gamma = 6$	7
$y_{...l}$	−18	18	−4	5	9	$10 = y_{....}$

■ **TABLE 4.21**
Analysis of Variance for the Rocket Propellant Problem

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	*P*-Value
Formulations	330.00	4	82.50	10.00	0.0033
Batches of raw material	68.00	4	17.00		
Operators	150.00	4	37.50		
Test assemblies	62.00	4	15.50		
Error	66.00	8	8.25		
Total	676.00	24			

The concept of orthogonal pairs of Latin squares forming a Graeco-Latin square can be extended somewhat. A $p \times p$ **hypersquare** is a design in which three or more orthogonal $p \times p$ Latin squares are superimposed. In general, up to $p + 1$ factors could be studied if a complete set of $p - 1$ orthogonal Latin squares is available. Such a design would utilize all $(p + 1)\ (p - 1) = p^2 - 1$ degrees of freedom, so an independent estimate of the error variance is necessary. Of course, there must be no interactions between the factors when using hypersquares.

4.4 Balanced Incomplete Block Designs

In certain experiments using randomized block designs, we may not be able to run all the treatment combinations in each block. Situations like this usually occur because of shortages of experimental apparatus or facilities or the physical size of the block. For example, in the vascular graft experiment (Example 4.1), suppose that each batch of material is only large enough to accommodate testing three extrusion pressures. Therefore, each pressure cannot be tested in each batch. For this type of problem it is possible to use randomized block designs in which every treatment is not present in every block. These designs are known as **randomized incomplete block designs**.

When all treatment comparisons are equally important, the treatment combinations used in each block should be selected in a balanced manner, so that any pair of treatments occur together the same number of times as any other pair. Thus, a **balanced incomplete block design (BIBD)** is an incomplete block design in which any two treatments appear together an equal number of times. Suppose that there are a treatments and that each block can hold exactly k ($k < a$) treatments. A balanced incomplete block design may be constructed by taking $\binom{a}{k}$ blocks and assigning a different combination of treatments to each block. Frequently, however, balance can be obtained with fewer than $\binom{a}{k}$ blocks. Tables of BIBDs are given in Fisher and Yates (1953), Davies (1956), and Cochran and Cox (1957).

As an example, suppose that a chemical engineer thinks that the time of reaction for a chemical process is a function of the type of catalyst employed. Four catalysts are currently being investigated. The experimental procedure consists of selecting a batch of raw material, loading the pilot plant, applying each catalyst in a separate run of the pilot plant, and observing the reaction time. Because variations in the batches of raw material may affect the performance of the catalysts, the engineer decides to use batches of raw material as blocks. However, each batch is only large enough to permit three catalysts to be run. Therefore, a randomized incomplete block design must be used. The balanced incomplete block design for this experiment, along with the observations recorded, is shown in Table 4.22. The order in which the catalysts are run in each block is randomized.

4.4.1 Statistical Analysis of the BIBD

As usual, we assume that there are a treatments and b blocks. In addition, we assume that each block contains k treatments, that each treatment occurs r times in the design (or is replicated r times), and that there are $N = ar = bk$ total observations. Furthermore, the number of times each pair of treatments appears in the same block is

$$\lambda = \frac{r(k-1)}{a-1}$$

If $a = b$, the design is said to be **symmetric**.

The parameter λ must be an integer. To derive the relationship for λ, consider any treatment, say treatment 1. Because treatment 1 appears in r blocks and there are $k-1$ other treatments in each of those blocks, there are $r(k-1)$ observations in a block containing treatment 1. These $r(k-1)$ observations also have to represent the remaining $a-1$ treatments λ times. Therefore, $\lambda(a-1) = r(k-1)$.

The **statistical model** for the BIBD is

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \tag{4.30}$$

where y_{ij} is the ith observation in the jth block, μ is the overall mean, τ_i is the effect of the ith treatment, β_j is the effect of the jth block, and ϵ_{ij} is the NID$(0, \sigma^2)$ random error component. The total variability in the data is expressed by the total corrected sum of squares:

$$SS_T = \sum_i \sum_j y_{ij}^2 - \frac{y_{..}^2}{N} \tag{4.31}$$

■ **TABLE 4.22**
Balanced Incomplete Block Design for Catalyst Experiment

Treatment	Block (Batch of Raw Material)				
(Catalyst)	1	2	3	4	$y_{i.}$
1	73	74	—	71	218
2	—	75	67	72	214
3	73	75	68	—	216
4	75	—	72	75	222
$y_{.j}$	221	224	207	218	$870 = y_{i.}$

Total variability may be partitioned into

$$SS_T = SS_{\text{Treatments(adjusted)}} + SS_{\text{Blocks}} + SS_E$$

where the sum of squares for treatments is **adjusted** to separate the treatment and the block effects. This adjustment is necessary because each treatment is represented in a different set of r blocks. Thus, differences between unadjusted treatment totals $y_{1.}, y_{2.}, \ldots, y_{a.}$ are also affected by differences between blocks.

The block sum of squares is

$$SS_{\text{Blocks}} = \frac{1}{k}\sum_{j=1}^{b} y_{.j}^2 - \frac{y_{..}^2}{N} \tag{4.32}$$

where $y_{.j}$ is the total in the jth block. SS_{Blocks} has $b - 1$ degrees of freedom. The adjusted treatment sum of squares is

$$SS_{\text{Treatments(adjusted)}} = \frac{k\sum_{i=1}^{a} Q_i^2}{\lambda a} \tag{4.33}$$

where Q_i is the adjusted total for the ith treatment, which is computed as

$$Q_i = y_{i.} - \frac{1}{k}\sum_{j=1}^{b} n_{ij}y_{.j} \quad i = 1, 2, \ldots, a \tag{4.34}$$

with $n_{ij} = 1$ if treatment i appears in block j and $n_{ij} = 0$ otherwise. The adjusted treatment totals will always sum to zero. $SS_{\text{Treatments(adjusted)}}$ has $a - 1$ degrees of freedom. The error sum of squares is computed by subtraction as

$$SS_E = SS_T - SS_{\text{Treatments(adjusted)}} - SS_{\text{Blocks}} \tag{4.35}$$

and has $N - a - b + 1$ degrees of freedom.

The appropriate statistic for testing the equality of the treatment effects is

$$F_0 = \frac{MS_{\text{Treatments(adjusted)}}}{MS_E}$$

The ANOVA is summarized in Table 4.23.

■ **TABLE 4.23**
Analysis of Variance for the Balanced Incomplete Block Design

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments (adjusted)	$\frac{k\sum Q_i^2}{\lambda a}$	$a - 1$	$\frac{SS_{\text{Treatments(adjusted)}}}{a-1}$	$F_0 = \frac{MS_{\text{Treatments(adjusted)}}}{MS_E}$
Blocks	$\frac{1}{k}\sum y_{.j}^2 - \frac{y_{..}^2}{N}$	$b - 1$	$\frac{SS_{\text{Blocks}}}{b-1}$	
Error	SS_E (by subtraction)	$N - a - b + 1$	$\frac{SS_E}{N-a-b+1}$	
Total	$\sum\sum y_{ij}^2 - \frac{y_{..}^2}{N}$	$N - 1$		

EXAMPLE 4.4

Consider the data in Table 4.22 for the catalyst experiment. This is a BIBD with $a = 4, b = 4, k = 3, r = 3, \lambda = 2$, and $N = 12$. The analysis of this data is as follows. The total sum of squares is

$$SS_T = \sum_i \sum_j y_{ij}^2 - \frac{y_{..}^2}{12}$$

$$= 63{,}156 - \frac{(870)^2}{12} = 81.00$$

The block sum of squares is found from Equation 4.32 as

$$SS_{\text{Blocks}} = \frac{1}{3}\sum_{j=1}^{4} y_{.j}^2 - \frac{y_{..}^2}{12}$$

$$= \frac{1}{3}[(221)^2 + (207)^2 + (224)^2 + (218)^2] - \frac{(870)^2}{12}$$

$$= 55.00$$

To compute the treatment sum of squares adjusted for blocks, we first determine the adjusted treatment totals using Equation 4.34 as

$$Q_1 = (218) - \tfrac{1}{3}(221 + 224 + 218) = -9/3$$

$$Q_2 = (214) - \tfrac{1}{3}(207 + 224 + 218) = -7/3$$

$$Q_3 = (216) - \tfrac{1}{3}(221 + 207 + 224) = -4/3$$

$$Q_4 = (222) - \tfrac{1}{3}(221 + 207 + 218) = 20/3$$

The adjusted sum of squares for treatments is computed from Equation 4.33 as

$$SS_{\text{Treatments(adjusted)}} = \frac{k\sum_{i=1}^{4} Q_i^2}{\lambda a}$$

$$= \frac{3[(-9/3)^2 + (-7/3)^2 + (-4/3)^2 + (20/3)^2]}{(2)(4)}$$

$$= 22.75$$

The error sum of squares is obtained by subtraction as

$$SS_E = SS_T - SS_{\text{Treatments(adjusted)}} - SS_{\text{Blocks}}$$

$$= 81.00 - 22.75 - 55.00 = 3.25$$

The analysis of variance is shown in Table 4.24. Because the *P*-value is small, we conclude that the catalyst employed has a significant effect on the time of reaction.

■ **TABLE 4.24**

Analysis of Variance for Example 4.4

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	*P*-Value
Treatments (adjusted for blocks)	22.75	3	7.58	11.66	0.0107
Blocks	55.00	3	—		
Error	3.25	5	0.65		
Total	81.00	11			

If the factor under study is fixed, tests on individual treatment means may be of interest. If orthogonal contrasts are employed, the contrasts must be made on the **adjusted treatment totals,** the $\{Q_i\}$ rather than the $\{y_{i.}\}$. The contrast sum of squares is

$$SS_c = \frac{k\left(\sum_{i=1}^{a} c_i Q_i\right)^2}{\lambda a \sum_{i=1}^{a} c_i^2}$$

where $\{c_i\}$ are the contrast coefficients. Other multiple comparison methods may be used to compare all the pairs of adjusted treatment effects, which we will find in Section 4.4.2 are estimated by $\hat{\tau}_i = kQ_i/(\lambda a)$. The standard error of an adjusted treatment effect is

$$s = \sqrt{\frac{kMS_E}{\lambda a}} \tag{4.36}$$

In the analysis that we have described, the total sum of squares has been partitioned into an adjusted sum of squares for treatments, an unadjusted sum of squares for blocks, and an error sum of squares. Sometimes we would like to assess the block effects. To do this, we require an alternate partitioning of SS_T, that is,

$$SS_T = SS_{\text{Treatments}} + SS_{\text{Blocks(adjusted)}} + SS_E$$

Here $SS_{\text{Treatments}}$ is unadjusted. If the design is symmetric, that is, if $a = b$, a simple formula may be obtained for $SS_{\text{Blocks(adjusted)}}$. The adjusted block totals are

$$Q'_j = y_{.j} - \frac{1}{4}\sum_{i=1}^{a} n_{ij}y_{i.} \quad j = 1, 2, \ldots, b \tag{4.37}$$

and

$$SS_{\text{Blocks(adjusted)}} = \frac{r\sum_{j=1}^{b}(Q'_j)^2}{\lambda b} \tag{4.38}$$

The BIBD in Example 4.4 is symmetric because $a = b = 4$. Therefore,

$$Q'_1 = (221) - \tfrac{1}{3}(218 + 216 + 222) = 7/3$$

$$Q'_2 = (224) - \tfrac{1}{3}(218 + 214 + 216) = 24/3$$

$$Q'_3 = (207) - \tfrac{1}{3}(214 + 216 + 222) = -31/3$$

$$Q'_4 = (218) - \tfrac{1}{3}(218 + 214 + 222) = 0$$

and

$$SS_{\text{Blocks(adjusted)}} = \frac{3[(7/3)^2 + (24/3)^2 + (-31/3)^2 + (0)^2]}{(2)(4)} = 66.08$$

Also,

$$SS_{\text{Treatments}} = \frac{(218)^2 + (214)^2 + (216)^2 + (222)^2}{3} - \frac{(870)^2}{12} = 11.67$$

A summary of the analysis of variance for the symmetric BIBD is given in Table 4.25. Notice that the sums of squares associated with the mean squares in Table 4.25 do not add to the total sum of squares, that is,

$$SS_T \neq SS_{\text{Treatments(adjusted)}} + SS_{\text{Blocks(adjusted)}} + SS_E$$

This is a consequence of the nonorthogonality of treatments and blocks.

Computer Output. There are several computer packages that will perform the analysis for a balanced incomplete block design. The SAS General Linear Models procedure is one of these and Minitab and JMP are others. The upper portion of Table 4.26 is the Minitab General Linear Model output for Example 4.4. Comparing Tables 4.26 and 4.25, we see that Minitab has computed the adjusted treatment sum of squares and the adjusted block sum of squares (they are called "Adj *SS*" in the Minitab output).

The lower portion of Table 4.26 is a multiple comparison analysis, using the Tukey method. Confidence intervals on the differences in all pairs of means and the Tukey test are displayed. Notice that the Tukey method would lead us to conclude that catalyst 4 is different from the other three.

■ **TABLE 4.25**
Analysis of Variance for Example 4.4, Including Both Treatments and Blocks

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	*P*-Value
Treatments (adjusted)	22.75	3	7.58	11.66	0.0107
Treatments (unadjusted)	11.67	3			
Blocks (unadjusted)	55.00	3			
Blocks (adjusted)	66.08	3	22.03	33.90	0.0010
Error	3.25	5	0.65		
Total	81.00	11			

4.4.2 Least Squares Estimation of the Parameters

Consider estimating the treatment effects for the BIBD model. The least squares normal equations are

$$\begin{aligned}
&\mu : N\hat{\mu} + r\sum_{i=1}^{a}\hat{\tau}_i + k\sum_{j=1}^{b}\hat{\beta}_j = y_{..} \\
&\tau_i : r\hat{\mu} + r\hat{\tau}_i + \sum_{j=1}^{b} n_{ij}\hat{\beta}_j = y_{i.} \qquad i = 1, 2, \ldots, a \\
&\beta_j : k\hat{\mu} + \sum_{i=1}^{a} n_{ij}\hat{\tau}_i + k\hat{\beta}_j = y_{.j} \qquad j = 1, 2, \ldots, b
\end{aligned} \tag{4.39}$$

Imposing $\sum \hat{\tau}_i = \sum \hat{\beta}_j = 0$, we find that $\hat{\mu} = \overline{y}_{..}$. Furthermore, using the equations for $\{\beta_j\}$ to eliminate the block effects from the equations for $\{\tau_i\}$, we obtain

$$rk\hat{\tau}_i - r\hat{\tau}_i - \sum_{j=1}^{b}\sum_{\substack{p=1\\p\neq 1}}^{a} n_{ij}n_{pj}\hat{\tau}_p = ky_{i.} - \sum_{j=1}^{b} n_{ij}y_{.j} \tag{4.40}$$

Note that the right-hand side of Equation 4.41 is kQ_i, where Q_i is the ith adjusted treatment total (see Equation 4.34). Now, because $\sum_{J=1}^{b} n_{ij}n_{pj} = \lambda$ if $p \neq i$ and $n_{pj}^2 = n_{pj}$ (because $n_{pj} = 0$ or 1), we may rewrite Equation 4.40 as

$$r(k-1)\hat{\tau}_i - \lambda\sum_{\substack{p=1\\p\neq 1}}^{a}\hat{\tau}_p = kQ_i \quad i = 1, 2, \ldots, a \tag{4.41}$$

Finally, note that the constraint $\sum_{i=1}^{a}\hat{\tau}_i = 0$ implies that $\sum_{\substack{p=1\\p\neq 1}}^{a}\hat{\tau}_p = -\hat{\tau}_i$ and recall that $r(k-1) = \lambda(a-1)$ to obtain

$$\lambda a\hat{\tau}_i = kQ_i \quad i = 1, 2, \ldots, a \tag{4.42}$$

Therefore, the least squares estimators of the treatment effects in the balanced incomplete block model are

$$\hat{\tau}_i = \frac{kQ_i}{\lambda a} i = 1, 2, \ldots, a \tag{4.43}$$

■ **TABLE 4.26**
Minitab (General Linear Model) Analysis for Example 4.4

General Linear Model

```
Factor          Type      Levels       Values
Catalyst       fixed           4      1 2 3 4
Block          fixed           4      1 2 3 4

Analysis of Variance for Time, using Adjusted SS for Tests
Source        DF      Seq SS      Adj SS      Adj MS         F           P
Catalyst       3      11.667      22.750       7.583     11.67       0.011
Block          3      66.083      66.083      22.028     33.89       0.001
Error          5       3.250       3.250       0.650
Total         11      81.000

Tukey 95.0% Simultaneous Confidence Intervals
Response Variable Time
All Pairwise Comparisons among Levels of Catalyst

Catalyst = 1 subtracted from:

Catalyst       Lower      Center      Upper   ---------+--------+--------+------
2             -2.327      0.2500      2.827    (--------*--------)
3             -1.952      0.6250      3.202     (--------*--------)
4              1.048      3.6250      6.202               (--------*--------)
                                              ----------+--------+--------+-----_
                                                        0.0      2.5       5.0

Catalyst = 2 subtracted from:
Catalyst       Lower      Center      Upper   ---------+--------+--------+------
3             -2.202      0.3750      2.952    (--------*--------)
4              0.798      3.3750      5.952              (--------*--------)
                                              ----------+--------+--------+-----_
                                                        0.0      2.5       5.0

Catalyst = 3 subtracted from:
Catalyst       Lower      Center      Upper   ---------+--------+--------+------
4             0.4228       3.000      5.577             (--------*--------)
                                              ----------+--------+--------+-----_
                                                        0.0      2.5       5.0

Tukey Simultaneous Tests
Response Variable Time
All Pairwise Comparisons among Levels of Catalyst

Catalyst = 1 subtracted from:

Level           Difference            SE of                       Adjusted
Catalyst          of Means       Difference        T-Value         P-Value
2                   0.2500           0.6982         0.3581          0.9825
3                   0.6250           0.6982         0.8951          0.8085
4                   3.6250           0.6982         5.1918          0.0130

Catalyst = 2 subtracted from:

Level           Difference            SE of                       Adjusted
Catalyst          of Means       Difference        T-Value         P-Value
3                   0.3750           0.6982         0.5371          0.9462
4                   3.3750           0.6982         4.8338          0.0175

Catalyst = 3 subtracted from:

Level           Difference            SE of                       Adjusted
Catalyst          of Means       Difference        T-Value         P-Value
4                    3.000           0.6982          4.297          0.0281
```

As an illustration, consider the BIBD in Example 4.4. Because $Q_1 = -9/3, Q_2 = -7/3, Q_3 = -4/3$, and $Q_4 = 20/3$, we obtain

$$\hat{\tau}_1 = \frac{3(-9/3)}{(2)(4)} = -9/8 \quad \hat{\tau}_2 = \frac{3(-7/3)}{(2)(4)} = -7/8$$
$$\hat{\tau}_3 = \frac{3(-4/3)}{(2)(4)} = -4/8 \quad \hat{\tau}_4 = \frac{3(20/3)}{(2)(4)} = 20/8$$

as we found in Section 4.4.1.

4.4.3 Recovery of Interblock Information in the BIBD

The analysis of the BIBD given in Section 4.4.1 is usually called the **intrablock analysis** because block differences are eliminated and all contrasts in the treatment effects can be expressed as comparisons between observations in the same block. This analysis is appropriate regardless of whether the blocks are fixed or random. Yates (1940) noted that, if the block effects are uncorrelated random variables with zero means and variance σ_β^2, one may obtain additional information about the treatment effects τ_i. Yates called the method of obtaining this additional information the **interblock analysis**.

Consider the block totals $y_{.j}$ as a collection of b observations. The model for these observations [following John (1971)] is

$$y_{.j} = k\mu + \sum_{i=1}^{a} n_{ij}\tau_i + \left(k\beta_j + \sum_{i=1}^{a} \epsilon_{ij} \right) \tag{4.44}$$

where the term in parentheses may be regarded as error. The interblock estimators of μ and τ_i are found by minimizing the least squares function

$$L = \sum_{j=1}^{b} \left(y_{.j} - k\mu - \sum_{i=1}^{a} n_{ij}\tau_i \right)^2$$

This yields the following least squares normal equations:

$$\begin{aligned} \mu &: N\tilde{\mu} + r\sum_{i=1}^{a} \tilde{\tau}_i = y_{..} \\ \tau_i &: kr\tilde{\mu} + r\tilde{\tau}_i + \lambda \sum_{\substack{p=1 \\ p\neq 1}}^{a} \tilde{\tau}_p = \sum_{j=1}^{b} n_{ij}y_{.j} \qquad i = 1, 2, \ldots, a \end{aligned} \tag{4.45}$$

where $\tilde{\mu}$ and $\tilde{\tau}_i$ denote the **interblock estimators**. Imposing the constraint $\sum_{i=1}^{a} \hat{\tau}_i = 0$, we obtain the solutions to Equations 4.45 as

$$\tilde{\mu} = \bar{y}_{..} \tag{4.46}$$

$$\tilde{\tau}_i = \frac{\sum_{j=1}^{b} n_{ij}y_{.j} - kr\bar{y}_{..}}{r - \lambda} \quad i = 1, 2, \ldots, a \tag{4.47}$$

It is possible to show that the interblock estimators $\{\tilde{\tau}_i\}$ and the intrablock estimators $\{\hat{\tau}_i\}$ are uncorrelated.

The interblock estimators $\{\tilde{\tau}_i\}$ can differ from the intrablock estimators $\{\hat{\tau}_i\}$. For example, the interblock estimators for the BIBD in Example 4.4 are computed as follows:

$$\tilde{\tau}_1 = \frac{663 - (3)(3)(72.50)}{3 - 2} = 10.50$$
$$\tilde{\tau}_2 = \frac{649 - (3)(3)(72.50)}{3 - 2} = -3.50$$
$$\tilde{\tau}_3 = \frac{652 - (3)(3)(72.50)}{3 - 2} = -0.50$$
$$\tilde{\tau}_4 = \frac{646 - (3)(3)(72.50)}{3 - 2} = -6.50$$

Note that the values of $\sum_{j=1}^{b} n_{ij}y_{.j}$ were used previously on page 164 in computing the adjusted treatment totals in the intrablock analysis.

Now suppose we wish to combine the interblock and intrablock estimators to obtain a single, unbiased, minimum variance estimate of each τ_i. It is possible to show that both $\hat{\tau}_i$ and $\tilde{\tau}_i$ are unbiased and also that

$$V(\hat{\tau}_i) = \frac{k(a-1)}{\lambda a^2}\sigma^2 \qquad \text{(intrablock)}$$

and

$$V(\tilde{\tau}_i) = \frac{k(a-1)}{a(r-\lambda)}(\sigma^2 + k\sigma_\beta^2) \qquad \text{(intrablock)}$$

We use a linear combination of the two estimators, say

$$\tau_i^* = \alpha_1\hat{\tau}_i + \alpha_2\tilde{\tau}_i \tag{4.48}$$

to estimate τ_i. For this estimation method, the minimum variance unbiased combined estimator τ_i^* should have weights $\alpha_1 = u_1/(u_1 + u_2)$ and $\alpha_2 = u_2/(u_1 + u_2)$, where $u_1 = 1/V(\hat{\tau}_i)$ and $u_2 = 1/V(\tilde{\tau}_i)$. Thus, the optimal weights are inversely proportional to the variances of $\hat{\tau}_i$ and $\tilde{\tau}_i$. This implies that the best combined estimator is

$$\tau_i^* = \frac{\hat{\tau}_i \dfrac{k(a-1)}{a(r-\lambda)}(\sigma^2 + k\sigma_\beta^2) + \tilde{\tau}_i \dfrac{k(a-1)}{\lambda a^2}\sigma^2}{\dfrac{k(a-1)}{\lambda a^2}\sigma^2 + \dfrac{k(a-1)}{a(r-\lambda)}(\sigma^2 + k\sigma_\beta^2)} \qquad i = 1, 2, \ldots, a$$

which can be simplified to

$$\tau_i^* = \frac{kQ_i(\sigma^2 + k\sigma_\beta^2) + \left(\sum_{j=1}^{b} n_{ij}y_{.j} - kr\bar{y}_{..}\right)\sigma^2}{(r-\lambda)\sigma^2 + \lambda a(\sigma^2 + k\sigma_\beta^2)} \qquad i = 1, 2, \ldots, a \tag{4.49}$$

Unfortunately, Equation 4.49 cannot be used to estimate the τ_i because the variances σ^2 and σ_β^2 are unknown. The usual approach is to estimate σ^2 and σ_β^2 from the data and replace these parameters in Equation 4.49 by the estimates. The estimate usually taken for σ^2 is the error mean square from the intrablock analysis of variance, or the **intrablock error**. Thus,

$$\hat{\sigma}^2 = MS_E$$

The estimate of σ_β^2 is found from the mean square for blocks adjusted for treatments. In general, for a balanced incomplete block design, this mean square is

$$MS_{\text{Blocks(adjusted)}} = \frac{\left(\dfrac{k\sum_{i=1}^{a} Q_i^2}{\lambda a} + \sum_{j=1}^{b}\dfrac{y_{.j}^2}{k} - \sum_{i=1}^{a}\dfrac{y_{i.}^2}{r}\right)}{(b-1)} \tag{4.50}$$

and its expected value [which is derived in Graybill (1961)] is

$$E[MS_{\text{Blocks(adjusted)}}] = \sigma^2 + \frac{a(r-1)}{(b-1)}\sigma_\beta^2$$

Thus, if $MS_{\text{Blocks(adjusted)}} > MS_E$, the estimate of $\hat{\sigma}^2_\beta$ is

$$\hat{\sigma}^2_\beta = \frac{[MS_{\text{Blocks(adjusted)}} - MS_E](b-1)}{a(r-1)} \tag{4.51}$$

and if $MS_{\text{Blocks(adjusted)}} \leq MS_E$, we set $\hat{\sigma}^2_\beta = 0$. This results in the combined estimator

$$\tau_i^* = \begin{cases} \dfrac{kQ_i(\hat{\sigma}^2 + k\hat{\sigma}^2_\beta) + \left(\displaystyle\sum_{j=1}^{b} n_{ij} y_{\cdot j} - kr\bar{y}_{\cdot\cdot}\right)\hat{\sigma}^2}{(r-\lambda)\hat{\sigma}^2 + \lambda a(\hat{\sigma}^2 + k\hat{\sigma}^2_\beta)}, & \hat{\sigma}^2_\beta > 0 \quad \textbf{(4.52a)} \\[2ex] \dfrac{y_{i.} - (1/a)y_{..}}{r}, & \hat{\sigma}^2_\beta = 0 \quad \textbf{(4.52b)} \end{cases}$$

We now compute the combined estimates for the data in Example 4.4. From Table 4.25 we obtain $\hat{\sigma}^2 = MS_E = 0.65$ and $MS_{\text{Blocks(adjusted)}} = 22.03$. (Note that in computing $MS_{\text{Blocks(adjusted)}}$ we make use of the fact that this is a symmetric design.) In general, we must use Equation 4.50. Because $MS_{\text{Blocks(adjusted)}} > MS_E$, we use Equation 4.51 to estimate σ^2_β as

$$\hat{\sigma}^2_\beta = \frac{(22.03 - 0.65)(3)}{4(3-1)} = 8.02$$

Therefore, we may substitute $\hat{\sigma}^2 = 0.65$ and $\hat{\sigma}^2_\beta = 8.02$ into Equation 4.52a to obtain the combined estimates listed below. For convenience, the intrablock and interblock estimates are also given. In this example, the combined estimates are close to the intrablock estimates because the variance of the interblock estimates is relatively large.

Parameter	Intrablock Estimate	Interblock Estimate	Combined Estimate
τ_1	−1.12	10.50	−1.09
τ_2	−0.88	−3.50	−0.88
τ_3	−0.50	−0.50	−0.50
τ_4	2.50	−6.50	2.47

4.5 Problems

4.1 Suppose that a single-factor experiment with four levels of the factor has been conducted. There are six replicates and the experiment has been conducted in blocks. The error sum of squares is 500 and the block sum of squares is 250. If the experiment had been conducted as a completely randomized design the estimate of the error variance σ^2 would be.

(a) 25.0 **(b)** 25.5 **(c)** 35.0

(d) 37.5 **(e)** None of the above

4.2 Suppose that a single-factor experiment with five levels of the factor has been conducted. There are three replicates and the experiment has been conducted as a complete randomized design. If the experiment had been conducted in blocks, the pure error degrees of freedom would be reduced by

(a) 3 **(b)** 5 **(c)** 2

(d) 4 **(e)** None of the above

4.3 Blocking is a technique that can be used to control the variability transmitted by uncontrolled nuisance factors in an experiment.

(a) True

(b) False

4.4 The number of blocks in the RCBD must always equal the number of treatments or factor levels.

(a) True

(b) False

4.5 The key concept of the phrase "Block if you can, randomize if you can't." is that:

(a) It is usually better to not randomize within blocks.

(b) Blocking violates the assumption of constant variance.

(c) Create blocks by using each level of the nuisance factor as a block and randomize within blocks.

(d) Randomizing the runs is preferable to randomizing blocks.

4.6 The ANOVA from a randomized complete block experiment output is shown below.

```
Source        DF        SS         MS         F     P
Treatment      4   1010.56          ?     29.84     ?
Block          ?         ?     64.765         ?     ?
Error         20    169.33          ?
Total         29   1503.71
```

(a) Fill in the blanks. You may give bounds on the *P*-value.

(b) How many blocks were used in this experiment?

(c) What conclusions can you draw?

4.7 Consider the single-factor completely randomized experiment shown in Problem 3.8. Suppose that this experiment had been conducted in a randomized complete block design and that the sum of squares for blocks was 80.00. Modify the ANOVA for this experiment to show the correct analysis for the randomized complete block experiment.

4.8 A chemist wishes to test the effect of four chemical agents on the strength of a particular type of cloth. Because there might be variability from one bolt to another, the chemist decides to use a randomized block design, with the bolts of cloth considered as blocks. She selects five bolts and applies all four chemicals in random order to each bolt. The resulting tensile strengths follow. Analyze the data from this experiment (use $\alpha = 0.05$) and draw appropriate conclusions.

	Bolt				
Chemical	**1**	**2**	**3**	**4**	**5**
1	73	68	74	71	67
2	73	67	75	72	70
3	75	68	78	73	68
4	73	71	75	75	69

4.9 Three different washing solutions are being compared to study their effectiveness in retarding bacteria growth in 5-gallon milk containers. The analysis is done in a laboratory, and only three trials can be run on any day. Because days could represent a potential source of variability, the experimenter decides to use a randomized block design. Observations are taken for four days, and the data are shown here. Analyze the data from this experiment (use $\alpha = 0.05$) and draw conclusions.

	Days			
Solution	**1**	**2**	**3**	**4**
1	13	22	18	39
2	16	24	17	44
3	5	4	1	22

4.10 Plot the mean tensile strengths observed for each chemical type in Problem 4.8 and compare them to an appropriately scaled *t* distribution. What conclusions would you draw from this display?

4.11 Plot the average bacteria counts for each solution in Problem 4.9 and compare them to a scaled *t* distribution. What conclusions can you draw?

4.12 Consider the hardness testing experiment described in Section 4.1. Suppose that the experiment was conducted as described and that the following Rockwell C-scale data (coded by subtracting 40 units) obtained:

	Coupon			
Tip	**1**	**2**	**3**	**4**
1	9.3	9.4	9.6	10.0
2	9.4	9.3	9.8	9.9
3	9.2	9.4	9.5	9.7
4	9.7	9.6	10.0	10.2

(a) Analyze the data from this experiment.

(b) Use the Fisher LSD method to make comparisons among the four tips to determine specifically which tips differ in mean hardness readings.

(c) Analyze the residuals from this experiment.

4.13 A consumer products company relies on direct mail marketing pieces as a major component of its advertising campaigns. The company has three different designs for a new brochure and wants to evaluate their effectiveness, as there are substantial differences in costs between the three designs. The company decides to test the three designs by mailing 5000 samples of each to potential customers in four different regions of the country. Since there are known regional differences in the customer base, regions are considered as blocks. The number of responses to each mailing is as follows.

	Region			
Design	**NE**	**NW**	**SE**	**SW**
1	250	350	219	375
2	400	525	390	580
3	275	340	200	310

(a) Analyze the data from this experiment.

(b) Use the Fisher LSD method to make comparisons among the three designs to determine specifically which designs differ in the mean response rate.

(c) Analyze the residuals from this experiment.

4.14 The effect of three different lubricating oils on fuel economy in diesel truck engines is being studied. Fuel economy is measured using brake-specific fuel consumption after the engine has been running for 15 minutes. Five different truck engines are available for the study, and the experimenters conduct the following RCBD.

	Truck				
Oil	**1**	**2**	**3**	**4**	**5**
1	0.500	0.634	0.487	0.329	0.512
2	0.535	0.675	0.520	0.435	0.540
3	0.513	0.595	0.488	0.400	0.510

(a) Analyze the data from this experiment.

(b) Use the Fisher LSD method to make comparisons among the three lubricating oils to determine specifically which oils differ in brake-specific fuel consumption.

(c) Analyze the residuals from this experiment.

4.15 An article in the *Fire Safety Journal* ("The Effect of Nozzle Design on the Stability and Performance of Turbulent Water Jets," Vol. 4, August 1981) describes an experiment in which a shape factor was determined for several different nozzle designs at six levels of jet efflux velocity. Interest focused on potential differences between nozzle designs, with velocity considered as a nuisance variable. The data are shown below:

Nozzle	Jet Efflux Velocity (m/s)					
Design	**11.73**	**14.37**	**16.59**	**20.43**	**23.46**	**28.74**
1	0.78	0.80	0.81	0.75	0.77	0.78
2	0.85	0.85	0.92	0.86	0.81	0.83
3	0.93	0.92	0.95	0.89	0.89	0.83
4	1.14	0.97	0.98	0.88	0.86	0.83
5	0.97	0.86	0.78	0.76	0.76	0.75

(a) Does nozzle design affect the shape factor? Compare the nozzles with a scatter plot and with an analysis of variance, using $\alpha = 0.05$.

(b) Analyze the residuals from this experiment.

(c) Which nozzle designs are different with respect to shape factor? Draw a graph of the average shape factor for each nozzle type and compare this to a scaled t distribution. Compare the conclusions that you draw from this plot to those from Duncan's multiple range test.

4.16 An article in *Communications of the ACM* (Vol. 30, No. 5, 1987) studied different algorithms for estimating software development costs. Six algorithms were applied to several different software development projects and the percent error in estimating the development cost was observed. Some of the data from this experiment is shown in the table below.

(a) Do the algorithms differ in their mean cost estimation accuracy?

(b) Analyze the residuals from this experiment.

(c) Which algorithm would you recommend for use in practice?

	Project					
Algorithm	**1**	**2**	**3**	**4**	**5**	**6**
1(SLIM)	1244	21	82	2221	905	839
2(COCOMO-A)	281	129	396	1306	336	910
3(COCOMO-R)	220	84	458	543	300	794
4(COCONO-C)	225	83	425	552	291	826
5(FUNCTION POINTS)	19	11	−34	121	15	103
6(ESTIMALS)	−20	35	−53	170	104	199

4.17 An article in *Nature Genetics* (2003, Vol. 34, pp. 85–90) "Treatment-Specific Changes in Gene Expression Discriminate in vivo Drug Response in Human Leukemia Cells" studied gene expression as a function of different treatments for leukemia. Three treatment groups are as follows: mercaptopurine (MP) only; low-dose methotrexate (LDMTX) and MP; and high-dose methotrexate (HDMTX) and MP. Each group contained ten subjects. The responses from a specific gene are shown in the table below.

(a) Is there evidence to support the claim that the treatment means differ?

(b) Check the normality assumption. Can we assume these samples are from normal populations?

(c) Take the logarithm of the raw data. Is there evidence to support the claim that the treatment means differ for the transformed data?

(d) Analyze the residuals from the transformed data and comment on model adequacy.

Treatments	**Observations**									
MP ONLY	334.5	31.6	701	41.2	61.2	69.6	67.5	66.6	120.7	881.9
MP + HDMTX	919.4	404.2	1024.8	54.1	62.8	671.6	882.1	354.2	321.9	91.1
MP + LDMTX	108.4	26.1	240.8	191.1	69.7	242.8	62.7	396.9	23.6	290.4

4.18 Consider the ratio control algorithm experiment described in Section 3.8. The experiment was actually conducted as a randomized block design, where six time periods were selected as the blocks, and all four ratio control algorithms were tested in each time period. The average cell voltage and the standard deviation of voltage (shown in parentheses) for each cell are as follows:

Ratio Control	**Time Period**		
Algorithm	**1**	**2**	**3**
1	4.93 (0.05)	4.86 (0.04)	4.75 (0.05)
2	4.85 (0.04)	4.91 (0.02)	4.79 (0.03)
3	4.83 (0.09)	4.88 (0.13)	4.90 (0.11)
4	4.89 (0.03)	4.77 (0.04)	4.94 (0.05)

Ratio Control	**Time Period**		
Algorithm	**4**	**5**	**6**
1	4.95 (0.06)	4.79 (0.03)	4.88 (0.05)
2	4.85 (0.05)	4.75 (0.03)	4.85 (0.02)
3	4.75 (0.15)	4.82 (0.08)	4.90 (0.12)
4	4.86 (0.05)	4.79 (0.03)	4.76 (0.02)

(a) Analyze the average cell voltage data. (Use $\alpha = 0.05$.) Does the choice of ratio control algorithm affect the average cell voltage?

(b) Perform an appropriate analysis on the standard deviation of voltage. (Recall that this is called "pot noise.") Does the choice of ratio control algorithm affect the pot noise?

(c) Conduct any residual analyses that seem appropriate.

(d) Which ratio control algorithm would you select if your objective is to reduce both the average cell voltage *and* the pot noise?

4.19 An aluminum master alloy manufacturer produces grain refiners in ingot form. The company produces the product in four furnaces. Each furnace is known to have its own unique operating characteristics, so any experiment run in the foundry that involves more than one furnace will consider furnaces as a nuisance variable. The process engineers suspect that stirring rate affects the grain size of the product. Each furnace can be run at four different stirring rates. A randomized block design is run for a particular refiner, and the resulting grain size data is as follows.

	Furnace			
Stirring Rate (rpm)	**1**	**2**	**3**	**4**
5	8	4	5	6
10	14	5	6	9
15	14	6	9	2
20	17	9	3	6

(a) Is there any evidence that stirring rate affects grain size?

(b) Graph the residuals from this experiment on a normal probability plot. Interpret this plot.

(c) Plot the residuals versus furnace and stirring rate. Does this plot convey any useful information?

(d) What should the process engineers recommend concerning the choice of stirring rate and furnace for this particular grain refiner if small grain size is desirable?

4.20 Analyze the data in Problem 4.9 using the general regression significance test.

4.21 Assuming that chemical types and bolts are fixed, estimate the model parameters τ_i and β_j in Problem 4.8.

4.22 Draw an operating characteristic curve for the design in Problem 4.9. Does the test seem to be sensitive to small differences in the treatment effects?

4.23 Suppose that the observation for chemical type 2 and bolt 3 is missing in Problem 4.8. Analyze the problem by estimating the missing value. Perform the exact analysis and compare the results.

4.24 Consider the hardness testing experiment in Problem 4.12. Suppose that the observation for tip 2 in coupon 3 is missing. Analyze the problem by estimating the missing value.

4.25 ***Two missing values in a randomized block.*** Suppose that in Problem 4.8 the observations for chemical type 2 and bolt 3 and chemical type 4 and bolt 4 are missing.

(a) Analyze the design by iteratively estimating the missing values, as described in Section 4.1.3.

(b) Differentiate SS_E with respect to the two missing values, equate the results to zero, and solve for estimates of the missing values. Analyze the design using these two estimates of the missing values.

(c) Derive general formulas for estimating two missing values when the observations are in *different* blocks.

(d) Derive general formulas for estimating two missing values when the observations are in the *same* block.

4.26 An industrial engineer is conducting an experiment on eye focus time. He is interested in the effect of the distance of the object from the eye on the focus time. Four different distances are of interest. He has five subjects available for the experiment. Because there may be differences among individuals, he decides to conduct the experiment in a randomized block design. The data obtained follow. Analyze the data from this experiment (use $\alpha = 0.05$) and draw appropriate conclusions.

	Subject				
Distance (ft)	**1**	**2**	**3**	**4**	**5**
4	10	6	6	6	6
6	7	6	6	1	6
8	5	3	3	2	5
10	6	4	4	2	3

4.27 The effect of five different ingredients (*A, B, C, D, E*) on the reaction time of a chemical process is being studied. Each batch of new material is only large enough to permit five runs to be made. Furthermore, each run requires approximately $1\frac{1}{2}$ hours, so only five runs can be made in one day. The experimenter decides to run the experiment as a Latin square so that day and batch effects may be systematically controlled. She obtains the data that follow. Analyze the data from this experiment (use $\alpha = 0.05$) and draw conclusions.

	Day				
Batch	**1**	**2**	**3**	**4**	**5**
1	$A = 8$	$B = 7$	$D = 1$	$C = 7$	$E = 3$
2	$C = 11$	$E = 2$	$A = 7$	$D = 3$	$B = 8$
3	$B = 4$	$A = 9$	$C = 10$	$E = 1$	$D = 5$
4	$D = 6$	$C = 8$	$E = 6$	$B = 6$	$A = 10$
5	$E = 4$	$D = 2$	$B = 3$	$A = 8$	$C = 8$

4.28 An industrial engineer is investigating the effect of four assembly methods (*A, B, C, D*) on the assembly time for a color television component. Four operators are selected for the study. Furthermore, the engineer knows that each assembly method produces such fatigue that the time required for the last assembly may be greater than the time required for the first, regardless of the method. That is, a trend develops in the required assembly time. To account for this source of variability, the engineer uses the Latin square design that follows. Analyze the data from this experiment ($\alpha = 0.05$) and draw appropriate conclusions.

Order of	Operator			
Assembly	**1**	**2**	**3**	**4**
1	$C = 10$	$D = 14$	$A = 7$	$B = 8$
2	$B = 7$	$C = 18$	$D = 11$	$A = 8$
3	$A = 5$	$B = 10$	$C = 11$	$D = 9$
4	$D = 10$	$A = 10$	$B = 12$	$C = 14$

4.29 Consider the randomized complete block design in Problem 4.9. Assume that the days are random. Estimate the block variance component.

4.30 Consider the randomized complete block design in Problem 4.12. Assume that the coupons are random. Estimate the block variance component.

4.31 Consider the randomized complete block design in Problem 4.14. Assume that the trucks are random. Estimate the block variance component.

4.32 Consider the randomized complete block design in Problem 4.16. Assume that the software projects that were used as blocks are random. Estimate the block variance component.

4.33 Consider the gene expression experiment in Problem 4.17. Assume that the subjects used in this experiment are random. Estimate the block variance component.

4.34 Suppose that in Problem 4.27 the observation from batch 3 on day 4 is missing. Estimate the missing value and perform the analysis using the value.

4.35 Consider a $p \times p$ Latin square with rows (α_i), columns (β_k), and treatments (τ_j) fixed. Obtain least squares estimates of the model parameters α_i, β_k, and τ_j.

4.36 Derive the missing value formula (Equation 4.28) for the Latin square design.

4.37 ***Designs involving several Latin squares.*** [See Cochran and Cox (1957), John (1971).] The $p \times p$ Latin square contains only p observations for each treatment. To obtain more replications, the experimenter may use several squares, say n. It is immaterial whether the squares used are the same or different. The appropriate model is

$$y_{ijkh} = \begin{array}{r} \mu + \rho_h + \alpha_{i(h)} \\ + \tau_j + \beta_{k(h)} \\ + (\tau\rho)_{jh} + \epsilon_{ijkh} \end{array} \begin{cases} i = 1, 2, \ldots, p \\ j = 1, 2, \ldots, p \\ k = 1, 2, \ldots, p \\ h = 1, 2, \ldots, n \end{cases}$$

where y_{ijkh} is the observation on treatment j in row i and column k of the hth square. Note that $\alpha_{i(h)}$ and $\beta_{k(h)}$ are the row and column effects in the hth square, ρ_h is the effect of the hth square, and $(\tau\rho)_{jh}$ is the interaction between treatments and squares.

(a) Set up the normal equations for this model, and solve for estimates of the model parameters. Assume that appropriate side conditions on the parameters are $\sum_h \hat{\rho}_h = 0$, $\sum_i \hat{\alpha}_{i(h)} = 0$, and $\sum_k \hat{\beta}_{k(h)} = 0$ for each h, $\sum_j \hat{\tau}_j = 0$, $\sum_j (\hat{\tau\rho})_{jh} = 0$ for each h, and $\sum_h (\hat{\tau\rho})_{jh} = 0$ for each j.

(b) Write down the analysis of variance table for this design.

4.38 Discuss how you would determine the sample size for use with the Latin square design.

4.39 Suppose that in Problem 4.27 the data taken on day 5 were incorrectly analyzed and had to be discarded. Develop an appropriate analysis for the remaining data.

4.40 The yield of a chemical process was measured using five batches of raw material, five acid concentrations, five standing times (*A, B, C, D, E*), and five catalyst concentrations ($\alpha, \beta, \gamma, \delta, \epsilon$). The Graeco-Latin square that follows was used. Analyze the data from this experiment (use $\alpha = 0.05$) and draw conclusions.

	Acid Concentration				
Batch	**1**	**2**	**3**	**4**	**5**
1	$A\alpha = 26$	$B\beta = 16$	$C\gamma = 19$	$D\delta = 16$	$E\epsilon = 13$
2	$B\gamma = 18$	$C\delta = 21$	$D\epsilon = 18$	$E\alpha = 11$	$A\beta = 21$
3	$C\epsilon = 20$	$D\alpha = 12$	$E\beta = 16$	$A\gamma = 25$	$B\delta = 13$
4	$D\beta = 15$	$E\gamma = 15$	$A\delta = 22$	$B\epsilon = 14$	$C\alpha = 17$
5	$E\delta = 10$	$A\epsilon = 24$	$B\alpha = 17$	$C\beta = 17$	$D\gamma = 14$

4.41 Suppose that in Problem 4.28 the engineer suspects that the workplaces used by the four operators may represent an additional source of variation. A fourth factor, workplace ($\alpha, \beta, \gamma, \delta$) may be introduced and another experiment conducted, yielding the Graeco-Latin square that follows. Analyze the data from this experiment (use $\alpha = 0.05$) and draw conclusions.

	Operator			
Order of Assembly	**1**	**2**	**3**	**4**
1	$C\beta = 11$	$B\gamma = 10$	$D\delta = 14$	$A\alpha = 8$
2	$B\alpha = 8$	$C\delta = 12$	$A\gamma = 10$	$D\beta = 12$
3	$A\delta = 9$	$D\alpha = 11$	$B\beta = 7$	$C\gamma = 15$
4	$D\gamma = 9$	$A\beta = 8$	$C\alpha = 18$	$B\delta = 6$

4.42 Construct a 5×5 hypersquare for studying the effects of five factors. Exhibit the analysis of variance table for this design.

4.43 Consider the data in Problems 4.28 and 4.41. Suppressing the Greek letters in problem 4.41, analyze the data using the method developed in Problem 4.37.

4.44 Consider the randomized block design with one missing value in Problem 4.24. Analyze this data by using the exact analysis of the missing value problem discussed in Section 4.1.4. Compare your results to the approximate analysis of these data given from Problem 4.24.

4.45 An engineer is studying the mileage performance characteristics of five types of gasoline additives. In the road test he wishes to use cars as blocks; however, because of a time constraint, he must use an incomplete block design. He runs the balanced design with the five blocks that follow. Analyze the data from this experiment (use $\alpha = 0.05$) and draw conclusions.

	Car				
Additive	**1**	**2**	**3**	**4**	**5**
1		17	14	13	12
2	14	14		13	10
3	12		13	12	9
4	13	11	11	12	
5	11	12	10		8

4.46 Construct a set of orthogonal contrasts for the data in Problem 4.45. Compute the sum of squares for each contrast.

4.47 Seven different hardwood concentrations are being studied to determine their effect on the strength of the paper produced. However, the pilot plant can only produce three runs each day. As days may differ, the analyst uses the BIBD that follows. Analyze the data from this experiment (use $\alpha = 0.05$) and draw conclusions.

Hardwood	**Days**			
Concentration (%)	**1**	**2**	**3**	**4**
2	114			
4	126	120		
6		137	117	
8	141		129	149
10		145		150
12			120	
14				136

Hardwood	**Days**		
Concentration (%)	**5**	**6**	**7**
2	120		117
4		119	
6			134
8			
10	143		
12	118	123	
14		130	127

4.48 Analyze the data in Example 4.4 using the general regression significance test.

4.49 Prove that $k\sum_{i=1}^{a} Q_i^2/(\lambda a)$ is the adjusted sum of squares for treatments in a BIBD.

4.50 An experimenter wishes to compare four treatments in blocks of two runs. Find a BIBD for this experiment with six blocks.

4.51 An experimenter wishes to compare eight treatments in blocks of four runs. Find a BIBD with 14 blocks and $\lambda = 3$.

4.52 Perform the interblock analysis for the design in Problem 4.45.

4.53 Perform the interblock analysis for the design in Problem 4.47.

4.54 Verify that a BIBD with the parameters $a = 8$, $r = 8, k = 4$, and $b = 16$ does not exist.

4.55 Show that the variance of the intrablock estimators $\{\hat{\tau}_i\}$ is $k(a-1)\sigma^2/(\lambda a^2)$.

4.56 ***Extended incomplete block designs.*** Occasionally, the block size obeys the relationship $a < k < 2a$. An extended incomplete block design consists of a single replicate of each treatment in each block along with an incomplete block design with $k^* = k - a$. In the balanced case, the incomplete block design will have parameters $k^* = k - a$, $r^* = r - b$, and λ^*. Write out the statistical analysis. (*Hint:* In the extended incomplete block design, we have $\lambda = 2r - b + \lambda^*$.)

4.57 Suppose that a single-factor experiment with five levels of the factor has been conducted. There are three replicates and the experiment has been conducted as a complete randomized design. If the experiment had been conducted in blocks, the pure error degrees of freedom would be reduced by (choose the correct answer):

(a) 3 **(b)** 5 **(c)** 2

(d) 4 **(e)** none of the above

4.58 Physics graduate student Laura Van Ertia has conducted a complete randomized design with a single factor, hoping to solve the mystery of the unified theory and complete her dissertation. The results of this experiment are summarized in the following ANOVA display:

Source	**DF**	**SS**	**MS**	***F***
Factor	?	?	14.18	?
Error	?	37.75	?	
Total	23	108.63		

Answer the following questions about this experiment.

(a) The sum of squares for the factor is ______.

(b) The number of degrees of freedom for the single factor in the experiment is ______.

(c) The number of degrees of freedom for error is ______.

(d) The mean square for error is ______.

(e) The value of the test statistic is ______.

(f) If the significance level is 0.05, your conclusions are not to reject the null hypothesis.

Yes

No

(g) An upper bound on the *P*-value for the test statistic is ______.

(h) A lower bound on the *P*-value for the test statistic is ______.

(i) Laura used ______ levels of the factor in this experiment.

(j) Laura replicated this experiment ______ times.

(k) Suppose that Laura had actually conducted this experiment as a randomized complete block design and the sum of squares for blocks was 12. Reconstruct the ANOVA display above to reflect this new situation. How much has blocking reduced the estimate of experimental error?

4.59 Consider the direct mail marketing experiment in Problem 4.13. Suppose that this experiment had been run as a completely randomized design, ignoring potential regional differences, but that exactly the same data was obtained. Reanalyze the experiment under this new assumption. What difference would ignoring blocking have on the results and conclusions?