1 INTRODUCTION

1.1 THE THEME OF THE BOOK

This book is about the power of statistical experiments. In the increasingly competitive global economy, firms are constantly under pressure to reduce costs, increase productivity, and improve quality. Testing or experimentation in the business world is commonplace, and the usual approach is to change one factor at a time while holding other factors constant. To some, this approach seems logical, simple, and therefore appealing. But as we will show, it is highly inefficient, and it may fail to identify important factors and lead to wrong conclusions. The better method is to test all factors simultaneously. Doing so not only reduces the costs of experimenting but, as we will demonstrate, also provides the experimenter with more and better information.

Elementary courses in statistics that cover topics such as probability, hypothesis testing, confidence intervals, and regression analysis often appear abstract; and although they are illustrated with numerous examples, they typically seem far removed from practical issues. In this book we use and build on basic statistical concepts to explore approaches for solving real-world problems. Although our focus is on practice, it is important to keep in mind that statistics is a science, and science is based on theory. While computer software has made the implementation of statistical methods much easier, there is a danger in relying on a cookbook approach in which the user fails to understand the underlying concepts. In contrast, this book's presentation combines theory and practice, and focuses on strengthening the reader's understanding of fundamental statistical ideas.

Our goal in writing this book is to share our passion for the subject and to provide students, practitioners, and managers with a set of highly relevant, interesting, and valuable tools. In the past, in the area of experimental design, nearly all the attention was focused on manufacturing rather than services. In contrast, most of the applications and examples in this book will involve marketing and service operations. In the next section, we give a brief introduction to some of the cases that are included.

1.2 A PREVIEW OF CASES

Throughout the book we illustrate concepts with practical examples. In addition, we include a group of real cases based on the actual implementation of experimental design methods. In this section, we discuss the highlights of a number of these cases.

In the marketing area, consumer testing is an important and widely used tool. But most marketing professionals hold firmly to the approach of changing only one variable at a time, which is often called "split-run testing" (also referred to as A/B splits, test-control, or champion-challenger testing). Only recently have marketing managers begun to embrace multifactor techniques that simultaneously test marketing variables. These experimental design methods are particularly well suited to product testing in supermarkets.

In one significant marketing application, and one of the cases in the book, a major magazine publisher sought to increase sales of its popular magazine in a chain of supermarkets. The firm identified 10 factors to test, including a discount on multiple copies (no or yes), an additional display rack in the snack food area (no or yes), and an on-shelf advertisement (no or yes). After considering a number of alternatives, the publisher implemented a 24-run Plackett-Burman experimental design (see Chapter 6). Each run consisted of a particular combination of settings of each of the 10 factors.

A key part of the experiment was to decide how many stores to include and how long to test, in order to achieve statistically significant results. A total of 48 stores were included, and the experiment ran for two weeks. As a result, the firm identified several changes that increased sales by 20%, and equally important, it gained insights into which changes would have a negative effect or no effect.

Direct mail is a common marketing channel, and firms use it for a wide range of products including credit cards, clothing, and magazines. Typically, response rates are very low, and a small increase in response can mean large financial benefits. *Mother Jones* magazine had extensive experience in direct mail testing aimed at increasing their subscription rates. Their protocol was to test only one change, such as the color of the envelope, in each mailing to potential subscribers. Using a fractional factorial design (Chapter 5) the firm was able to test seven factors simultaneously in a single mailing, gaining valuable and immediate insights that led to large increases in response. Moreover, the results were attained with a sample size (the number of people receiving the mailing) that was much smaller than would have been needed if the seven factors had been tested one factor at a time.

A leading office supplies retailer designed and implemented an e-mail test targeted at small business customers, a group the retailer wanted to attract to their stores and Web site. The retailer identified 13 factors that it wanted to include in this experiment, with each factor having two possible values. The factors included the background color of the e-mail (white or blue), a discount offer (normal price or 15% discount), a free gift (no gift or a pen-and-pencil set), and products pictured (few or many). Testing all possible combinations of the 13 factors would have required $2^{13} = 8{,}192$ different e-mail designs! But using a fractional factorial design, a methodology that we discuss in Chapter 5, the firm was able to successfully test all 13 factors with just 32 different designs.

Peak Electronics, a manufacturer of printed circuit boards, was faced with a recurring problem. In the circuit board production process, most of the holes on each board are plated with a thin layer of copper so that current can flow from one side to the other. Some holes, however, are not meant to be plated and instead are *tented*, meaning that they are protected by a thin layer of photographic film. During the manufacturing process, a significant number of these tents were breaking, and their holes were being plated. The result was the number-one cause of rework at the firm, because the copper in these holes had to be scraped out.

At the time, Peak was using film supplied by Dupont. The sales representative of Hercules, a competing filmmaker, suggested that Peak perform an experiment using the Hercules film to test the effect on broken tents of a number of key manufacturing variables. The sales representative designed the test and helped Peak analyze the results.

With the explosive growth of the Internet, Web site design has become an important issue, as firms attempt to attract a greater number of people to visit their sites and order their products or services. PhoneHog is a subscription-based service through which consumers get free long-distance phone calls. Participants sign up for the program and earn phone minutes by visiting Internet sites, entering sweepstakes, or trying new products and services. The PhoneHog case in Chapter 8 describes how experimental design can be used to improve a Web site to obtain more customers. In this case there were 10 factors to be tested with the number of variations, or levels, for each factor ranging between 2 and 10. For example, the top image on the Web page had four possible designs: (1) photos of five people talking on the phone with the PhoneHog logo on the right, (2) a cartoon image of a pig peeking through the O in the PhoneHog logo on a blue background, (3) the same image of the pig on a white background, and (4) the photos of the five people talking on the phone with a different PhoneHog logo on the right. If every possible combination of factor levels were included in the experiment, a total of 1,658,880 test Web pages would have been required. In fact, the experiment consisted of just 45 different Web pages (each page a combination of factor levels), with each person arriving to the site randomly assigned to one of them. The number of visitors to the site and the number of visitors who click on an icon to request additional information were recorded. As a result of this experiment, the click-through rate, which is the number of clicks divided by the number of visitors, increased by 35%.

1.3 A BRIEF HISTORY OF EXPERIMENTAL DESIGN

The field of experimental design began with the pioneering work of Sir Ronald Fisher, whose classic book, *The Design of Experiments*, was published in 1935. Fisher was responsible for statistical analysis at an agricultural experiments station in England, and his early work on experimental design was applied to improving crop yields and solving other agricultural problems. Over the years, applications of experimental design to industrial problems have been widespread, with particular attention given to problems in the chemical industry, such as maximizing chemical yields and assays. In 1978, George E. P. Box, William G. Hunter, and J. Stuart Hunter published *Statistics for Experimenters* (second edition, 2005), a book that became, and still is, a standard text in the field.

Beginning about 1980, U.S. manufacturing firms, faced with competitive challenges, especially from Japanese companies, took a renewed interest in quality management and design of experiments. This period spurred renewed interest among U.S. manufacturers in experimental design, and in the 1980s the American Society for Quality (ASQ) and many other organizations started to offer numerous seminars on experimental design. However, little or no attention was given to the application of experimental design to service organizations.

More recently that has slowly begun to change, and several articles have appeared showing that multivariable experimental design techniques provide powerful approaches to service problems. "The New Mantra: MVT" (*Forbes*, March 11, 1996) discussed the experimental design applications to services by a quality consulting firm, while "Tests Lead Lowe's to Revamp Strategy" (*Wall Street Journal*, March 11, 1999) explained how that firm helped Lowe's improve its advertising policy. The article, "Boost Your Marketing ROI with Experimental Design" (Almquist and Wyner, *Harvard Business Review*, October 1, 2001), told how another consulting firm used experimental design to improve marketing decisions. In short, business leaders are beginning to realize that experimental design has widespread applications to management decision making, particularly in service organizations.

1.4 OUTLINE OF THE BOOK

Chapter 2 presents important basic concepts of probability and statistics. It is meant to be a concise review and provides the common language and notation that we use throughout the book. While writing it, we assumed that readers have previously had an exposure to most of the material covered in the chapter. We discuss a number of important distributions such as the binomial, normal, t-, and F-distributions. In subsequent chapters, they are used extensively. We also discuss useful tools for displaying data such as dot plots, histograms, and scatter diagrams. Chapter 2 also shows how confidence intervals and tests of hypotheses are constructed based on sample information and are used to make inferences about a population mean or the difference in two population means. These important statistical tools are applied in later chapters to identify statistically significant factors.

In Chapter 3, we extend the discussion in Chapter 2 and focus on comparing more than two population means. For example, we might want to compare the effectiveness of three different advertising strategies by testing them in a number of stores. We present two statistical models: the completely randomized design and the randomized block design. In Chapter 3, we emphasize two important ideas, randomization and blocking, that are used throughout the book.

The heart of the book begins with Chapter 4, where we focus on so-called 2-level factorial designs. In these designs, there are k factors to be tested, and each factor is studied at two different values (levels). For example, in a Web site test, one factor might be the banner headline (version 1 vs. version 2), while another might be the image under the headline (product photo vs. happy user). In the full factorial design, the experimenter tests all combinations of factors and levels, with each combination called a run. With k factors, there are 2^k runs. Thus, testing two factors requires $2^2 = 4$ runs, testing three factors requires $2^3 = 8$ runs, and so forth. The main effect of a factor is the difference in response at one level of the factor versus the other. For example, for the image under the banner headline, the main effect of that factor is the difference in response if the happy user image is employed rather than the product photo. In some instances, there may be an interaction between factors. For example, the difference in response between version 1 and version 2 of the banner headline may depend on which image under the banner is used. In Chapter 4, we show how main and interaction effects are estimated and discuss the various approaches for determining which effects are statistically significant.

The focus of Chapter 5 is on 2-level fractional factorial designs. Full factorial designs are useful for experimenting with relatively few factors. As the number of factors increases, the number of runs required in a full factorial design increases dramatically. In fact, the inclusion of each additional factor in a full factorial design doubles the number of runs required, with 4 factors requiring $2^4 = 16$ runs, 5 factors requiring $2^5 = 32$ runs, and so forth. If full factorials were the only option, the experimental design approach would have limited value. In a fractional factorial design, the experiment requires only a fraction of the number of runs needed for a full factorial design. For example, a full factorial design with seven factors requires $2^7 = 128$ runs, or separate experiments. But as we shall see, it is possible to construct a fractional design requiring only 16 runs that provides nearly as much information as in a full factorial design. In some instances, a fractional experiment may produce results that are difficult to interpret. We show how a follow-up experiment can be designed and executed to resolve these ambiguities.

In Chapter 6, we discuss Plackett-Burman designs. The number of runs required in a fractional factorial design is a power of 2. Thus, the number of runs would be 8, 16, 32, 64, and so forth. In a Plackett-Burman design, the number of runs required is a multiple of 4, so the number of runs would be 4, 8, 12, 16, and so forth. For example, in a particular situation, if the experimenter were limited to fractional factorial designs, she might have to choose between a design of 16 runs and a design of 32 runs. There is a rather large gap between allowable run sizes. The Plackett-Burman designs give the experimenter additional options that may be advantageous. We discuss the characteristics of Plackett-Burman designs and illustrate their use with several case examples.

The designs in Chapters 4 through 6 are all 2-level designs, with each factor being set at one value or another. In Chapter 7, we extend the analysis to include designs in which factors may be at more than two levels. We show how regression analysis can be used to estimate effects, and we discuss the construction and analysis of simple fractional designs that include factors at more than two levels.

The last chapter of the book, Chapter 8, is devoted mainly to the most advanced topic. The designs discussed in earlier chapters have an important property called *orthogonality*. In an orthogonal design, effects are estimated independently of one another. That means that the particular estimate of one effect is not influenced by the estimated value of another. In Chapter 8, we consider nonorthogonal designs involving many factors and several levels. We show how regression analysis can be used to analyze these designs, and we illustrate the approach with the PhoneHog case, which was described earlier in this chapter. Chapter 8 ends with a discussion of experimental design software focusing on two software products, Minitab and JMP.

1.5 NOBODY ASKED US, BUT . . .

R. A. Fisher, The Life of a Scientist, is an interesting biography of Sir Ronald Fisher written by his daughter, Joan Fisher Box (1978). Fisher is one of the statisticians included in *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*, by David Salsburg (2001). The title of the book comes from a paper that Fisher wrote, which is included in Fisher's *The Design of Experiments*. As the story goes, a lady claimed that by tasting it she could tell whether milk or tea was put into the cup first. Fisher designed an experiment to test her claim. Salsburg's book has stories of other great statisticians including William Gossett, famous for the *t*-distribution, which we discuss in Chapter 2. The online (and free) encyclopedia Wikipedia (www.wikipedia.org) has interesting biographical information on Fisher, Gossett, and many other important figures in the world of statistics.

The NBC television white paper "If Japan Can, Why Can't We?" which was broadcast in 1980, was a milestone that marked the beginning of a quality revolution in manufacturing in the United States. W. Edwards Deming was featured on the program, and he castigated American firms for shoddy quality. Deming, a statistician with a Ph.D. in physics, gave a series of lectures in Japan in 1950 that greatly influenced that country's quality efforts. The Deming prize, the highest award for quality in Japan, is named in his honor. Deming's (1982) book, *Out of the Crisis*, is a good source for learning about his quality management ideas.

Not long after that NBC program, the work of Genichi Taguchi, a Japanese consultant and former professor, began receiving widespread attention from manufacturers in the United States, particularly in the automobile industry. Taguchi methods became a familiar buzzword for his approaches to experimental design. Statisticians have often criticized Taguchi's statistical methods, but there is general agreement that his engineering ideas are very useful. He is probably best known for two concepts: robust design and the Taguchi loss function. *Robust design* means designing a product or process that is insensitive to environmental factors. For example, a robust cake recipe would produce a good cake even with considerable variation in baking time and oven temperature. The *Taguchi loss function* is an appealing alternative to the traditional approach to determine whether a product or process meets customer specifications. For example, to illustrate the traditional approach, suppose the plating thickness in millimeters of a printed circuit board is acceptable if it falls within certain upper and lower specification limits. So, a board having thickness just below the upper specification would be judged acceptable, whereas a board whose thickness was just above that limit would be classified as defective. In reality, there is a target that is ideal, and the closer each board comes to that target, the better. In contrast, under Taguchi's loss function, the loss associated with an individual board would be equal to a constant times the squared deviation between the board's thickness and the target. With this function, doubling the distance from the target would quadruple the loss. In the traditional approach, where each board is either in or out of specifications, two processes might have the same fraction of boards meeting specifications but, in reality, very different quality levels. One process might have most of its acceptable boards with thicknesses close to the target, whereas the other might have a more uniform distribution with board thicknesses evenly spread within the window defined by the specification limits. This process would have much lower quality than the other, but under the traditional approach, the quality of products produced under the two systems would be judged as equal.

In recent years, the Six Sigma approach to quality has been embraced by numerous organizations. Six Sigma was originally developed at Motorola in the mid-1980s and refined first by Allied Signal and more recently by General Electric. Six Sigma has many similarities to total quality management (TQM) and other programs in the past, but it also has some distinctive characteristics. One is its focus on defining and responding to customer needs. In doing so, it takes a broader view of quality management compared to some more narrowly focused programs of the past, better integrating quality activities into all areas of the organization and aligning these activities with the strategic goals of the firm. In addition, Six Sigma programs have been more widely applied to service processes, including many implementations in hospitals and other health care organizations. The design of efficient experiments is an important component of the Six Sigma approach. One of the many books on Six Sigma is *The Six Sigma Way: How GE, Motorola, and Other Top Companies are Honing Their Performance*, by Peter S. Pande, Robert P. Neuman, and Roland R. Cavanagh (2000).

EXERCISES

Exercise 1 Search the Web for the work of Sir Ronald Fisher on experimental design, including his earliest efforts performing agricultural experiments at the Rothamsted Experimental Station in the United Kingdom.

Exercise 2 Read *Mother Jones* (Case 3 in the case study appendix) and *Peak Electronics: The Broken Tent Problem* (Case 4). Both of these cases describe a company's first exposure to experimental design methods.

(a) *Mother Jones*: Suppose the organization wanted to test each of the seven factors in a separate mailing. What specific shortcomings would this approach have compared to the approach in the case?

(b) *Peak Electronics*: Suppose the company did not use experimental design to examine and solve the broken tent problem. Imagine how they would have approached the problem instead. What difficulties would they have likely encountered? Would it have been possible to identify interactions between factors? If so, how?

Exercise 3 Pick a Web site on the Internet. Suppose you were designing an experiment for increasing visitors' response to a product or service offered on the site. What seven factors do you think would be most important to test? In each case, if possible, specify two levels (values) for each of the factors.

2 A REVIEW OF BASIC STATISTICAL CONCEPTS

2.1 INTRODUCTION

This chapter reviews basic concepts that we use in the remainder of this book. Section 2.2 reviews discrete and continuous probability distributions including two important special cases, the binomial and normal distributions. In Section 2.3, we focus on the graphical display and numerical summary of information. Topics covered include bar and pie charts for categorical data; dot diagrams; histograms and scatter plots for continuous data; and summary measures including the mean, median, standard deviation, and correlation coefficient. In Section 2.4, we discuss sampling and random sampling, and in Section 2.5, we review the basics of statistical inference. We discuss confidence intervals and hypothesis tests for a single mean and a single proportion, and determine the sample size that is required for estimates to achieve a given level of precision. We also address the comparison of two populations, using data from the completely randomized as well as the randomized block experiment. A case study on the effectiveness of two advertising strategies completes the chapter.

2.2 PROBABILITY DISTRIBUTIONS

The world is uncertain, and measurements on products and processes vary. Probability distributions describe the variability among the measurements.

Random variables are variables whose outcomes are uncertain. For example, the purchasing response of a customer who receives a catalog or an e-mail offer can be "yes" or "no"—or in coded form, 1 for "yes," and 0 for "no." Similarly, the soldering quality of a circuit board, expressed in terms of the number of flaws, is a random variable. The board may have zero flaws, exactly one error, two errors, and so on.

Random variables with a discrete number of possible outcomes (in the first example, 0 and 1; in the second example, 0, 1, 2, . . .) are called *discrete random variables*. We use discrete probability distributions to describe the uncertainty. Later in this chapter, we discuss the binomial distribution, the most important discrete distribution.

Variables such as the length or the width of a product, the amount spent on purchases, the commuting time to work, the gas mileage of a car, or the yield of a process are continuous