Modern Experimental Design

THOMAS P. RYAN
Acworth, GA

WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

Modern Experimental Design

THE WILEY BICENTENNIAL–KNOWLEDGE FOR GENERATIONS

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

William J. Pesce

WILLIAM J. PESCE
PRESIDENT AND CHIEF EXECUTIVE OFFICER

Peter Booth Wiley

PETER BOOTH WILEY
CHAIRMAN OF THE BOARD

Modern Experimental Design

THOMAS P. RYAN
Acworth, GA

WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

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

Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Ryan, Thomas P.
Modern experimental design / by Thomas P. Ryan
p. cm.
Includes bibliographical references and index.
ISBN 978-0-471-21077-1

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

Contents

Preface

Although there is a moderate amount of data analysis, especially in certain chapters, the emphasis in this book is on the statistical design of experiments. Such emphasis is justified by the widely held view that data from a well-designed experiment are easy to analyze. Certain types of designs are not simple, however, such as those covered in Chapters 7, 8, and 11, and the problem is compounded by the fact that some popular statistical software packages have quite limited capability for those designs.

The book would be suitable for an undergraduate one-semester course in design of experiments. For a course taught to nonstatistics majors, an instructor may wish to cover Chapters 1–4, part of Chapter 5, and then pick and choose from the other chapters in accordance with the needs of the students. The selection might include either or both of Chapters 10 and 12 and then cover sections of interest in Chapter 13.

For statistics majors, the book would be suitable for use in an advanced undergraduate course, perhaps covering Chapters 1–5, 7, 8, and much of Chapter 13. There is also enough advanced material for the book to be useful as a reference book in a graduate course taught to statistics majors, and might also be used in a graduate course for nonstatistics majors, depending on the needs and backgrounds of the students.

There is also enough material for a two-semester course, with the first course perhaps covering Chapters 1–6 and the second course covering Chapters 7–12 and 14, and parts of Chapter 13.

There is a considerable amount of material that is not covered to any extent, if at all, in other books on the subject, and some or all of this material might be used in special topics courses. These topics include conditional effects, uniform designs, and designs for restricted operating regions. (I have covered this material in an Internet course.)

A two-semester course in statistical methods should provide more than enough background for the book since the emphasis is on designs rather than statistical concepts. Matrix algebra is used in various places in the book, although it is not used extensively. Nevertheless, proficiency in the basics of matrix algebra is necessary for following some of the material.

One of the special features of the book is the emphasis on conditional effects in Chapters 4, 5, 6, and 10. This is an important topic that is not covered to any extent in most books and is addressed in very few journal articles. Another somewhat unique feature is moderate use of URLs, especially links to published articles that are available to the general public as well as article preprints and technical reports. There are other links for articles that are available to certain groups, such as members of the American Society for Quality. Some of those URLs might of course become outdated but I decided to list them since many of them, such as links to journal articles, will probably not become outdated in the near future. They make available to the reader a considerate amount of important resource material.

It is worth noting that this book does not contain catalogs of designs, as are given in some other books on the subject. Rather, the emphasis is on understanding design concepts and properties, the software that is available for generating specific designs and when to use those designs, and as stated, a moderate amount of analysis of data from experiments in which the designs are used, with extensive analysis provided in some case studies. Although there is some hand computation, the emphasis is on using appropriate software to generate output and interpret the output.

It is also worth noting that whereas there are case studies and a moderate amount of data analyses, there is not a "full" analysis of any dataset as that would include checking for outliers and influential observations, testing assumptions, and so on, which are covered in books on statistical methods. This is important but comes under the heading of data analysis rather than design and analysis of experiments. Although this book has more analysis than most books on design of experiments, it is not intended to be a handbook on data analysis.

I wish to gratefully acknowledge my editor, Steve Quigley, who motivated me to write this book, in addition to the contributions of associate editor Susanne Steitz, production editor Rosalyn Farkas, and colleagues who have made helpful comments, including Dennis Lin and Ivelisse Aviles, plus the helpful comments of three anonymous reviewers.

THOMAS P. RYAN

CHAPTER 1

Introduction

The statistical design of experiments plays a prominent role in experimentation. As George Box has stated, to see how a system functions when you have interfered with it, you have to interfere with it. That "interference" must be done in a systematic way so that the data from the experiment produce meaningful information.

The design of an experiment should be influenced by (1) the objectives of the experiment, (2) the extent to which sequential experimentation will be performed, if at all, (3) the number of factors under investigation, (4) the possible presence of identifiable and nonidentifiable extraneous factors, (5) the amount of money available for the experimentation, and (6) the purported model for modeling the response variable. Inman, Ledolter, Lenth, and Niemi (1992) stated, "Finally, it is impossible to overemphasize the importance of using a statistical model that matches the experimental design that was actually used." If we turn that statement around, we should use a design that matches a tentative model, recognizing that we won't know the model exactly.

In general, the design that is used for an experiment should be guided by these objectives. In many cases, the conditions and objectives will lead to an easy choice of a design, but this will not always be the case. Software is almost indispensable in designing experiments, although commonly used software will sometimes be inadequate, such as when there is a very large number of factors. Special-purpose software, not all of which is commercially available, will be needed in some circumstances. Various software programs are discussed throughout the book, with strong emphasis on Design-Expert®, which has certain features reminiscent of expert systems software, JMP®, and MINITAB®. (Readers intending to use the latter for designing experiments and analyzing the resultant data may be interested in Mathews (2004), although the latter is largely an introductory statistics book. Parts of the book are available online to members of the American Society for Quality (ASQ) at http://qualitypress.asq.org/chapters/H1233.pdf.) Although it is freeware, GOSSET is far more powerful than typical freeware. It is especially good for optimal designs (see Section 13.7) and runs on Unix, Linux, and Mac operating systems. Since GOSSET

Modern Experimental Design By Thomas P. Ryan

is not as well known by experimenters, its Web site has been given here, which is http://www.research.att.com/~njas/gosset/index.html.

Design-Expert is a registered trademark of Stat-Ease, Inc. JMP is a registered trademark of SAS Institute, Inc. MINITAB is a registered trademark of MINITAB, Inc.

1.1 EXPERIMENTS ALL AROUND US

People perform experiments all of the time: workers who are new to a city want to find the shortest and/or fastest route to work, chefs experiment with new recipes, computer makers try to make better and faster computers, and so on. Improvement in processes is often the objective, as is optimality, such as finding the shortest route to work.

A pharmaceutical company that invents a new drug it believes is effective in combating a particular disease has to support its belief with the results of clinical trials, a form of experimentation. A scientist who believes he or she has made an important discovery needs to have the result supported by the results of experimentation. Although books on design of experiments did not begin to appear until well into the twentieth century, experimentation is certainly about as old as mankind.

Undoubtedly, all kinds of experiments were performed centuries ago that did not become a part of recorded history. About 100 years ago some rather extreme and bizarre experiments performed by Duncan MacDougall, MD, did become part of recorded history, however. He postulated that the human soul has measurable mass that falls within a specific range of weights. To prove this, he performed experiments on humans and dogs. In experimentation described at http://www.snopes.com/religion/soulweight.asp, Dr. MacDougall supposedly used six terminal patients and weighed them before, during, and after the process of death. The first patient lost three-fourths of an ounce and Dr. MacDougall, who apparently sought to conduct his experiments in a manner approximating the scientific method (see, e.g., Beveridge, 1960), ruled out all possible physiological explanations for the loss of weight. Since 3/4 ounce equals 21.26 grams, the result of this experimentation is believed to form the basis for the title of the movie *21 Grams* that was released in 2003 and starred Sean Penn and Naomi Watts.

To help confirm his conclusion, Dr. MacDougall decided to perform the same experiment on 15 dogs and found that the weight of the dogs did not change. As he stated, "the ideal test on dogs would be obtained in those dying from some disease that rendered them much exhausted and incapable of struggle." Unfortunately, he found that "it was not my good fortune to find dogs dying from such sickness." This prompted author Mary Roach (2003) to write "barring a local outbreak of distemper, one is forced to conclude that the good doctor calmly poisoned fifteen healthy canines for his little exercise in biological theology."

Accounts of Dr. MacDougall's experiments were published in the journal *American Medicine* and in *The New York Times*: "Soul has weight, physician thinks," March 11, 1907, p. 5, and "He weighed human soul," October 16, 1920, p. 13, with the latter published at the time of his death. MacDougall admitted that his experiments would have to be repeated many times with similar results before any conclusions could be drawn. Today his work is viewed as suffering from too small a sample size and an

imprecise measuring instrument, and is viewed as nothing more than a curiosity (see, for example, http://www.theage.com.au/articles/2004/02/20/1077072838871.html.)

Although such experimentation is quite different from most types of experimentation that involve statistically designed experiments, small sample sizes and imprecise measuring instruments can undermine any experiment. Accordingly, attention is devoted in Section 1.4.3 and in other chapters on necessary minimum sample sizes for detecting significant effects in designed experiments.

More traditional experiments, many of which were performed more than 50 years ago, are in the 113 case studies of statistically designed experiments given by Bisgaard (1992).

When we consider all types of experiments that are performed, we find that certainly most experiments are not guided by statistical principles. Rather, most experimentation is undoubtedly trial-and-error experimentation. Much experimentation falls in the one-factor-at-a-time (OFAT) category, with each of two or more factors varied one at a time while the other factors are held fixed. Misleading information can easily result from such experiments, although OFAT designs can occasionally be used beneficially. These designs are discussed in Section 13.1.

1.2 OBJECTIVES FOR EXPERIMENTAL DESIGNS

The objectives for each experiment should be clearly delineated, as these objectives will dictate the construction of the designs, with sequential experimentation generally preferred. The latter is usually possible, depending upon the field of application. Bisgaard (1989) described a sequence of experiments and how, after considerable frustration, a satisfactory end result was finally achieved.

As explained by John (2003), however, sequential experimentation isn't very practical in the field of agronomy, as the agronomist must plan his or her experiment in the spring and harvest all of the data in the fall. Such obstacles to sequential experimentation do not exist in engineering applications, nor do they exist in most other fields of application. (John (2003) is recommended reading for its autobiographical content on one of the world's leading researchers in experimental design over a period of several decades.)

Following are a few desirable criteria for an experimental design:

(1) The design points should exert equal influence on the determination of the regression coefficients and effect estimates, as is the case with almost all the designs discussed in this book.
(2) The design should be able to detect the need for nonlinear terms.
(3) The design should be robust to model misspecification since all models are wrong.
(4) Designs in the early stage of the use of a sequential set of designs should be constructed with an eye toward providing appropriate information for follow-up experiments.

Box and Draper (1975) gave a list of 14 properties that a response surface design (see Chapter 10) should possess, and most of the properties are sufficiently general as to be

applicable to virtually all types of designs. That list was published over 30 years ago and many advancements have occurred since then, although some properties, such as "provide data that will allow visual analysis," will stand the test of time.

Assume that a marathon runner would like to identify the training and nutritional regimens that will allow him or her to perform at an optimal level in a forthcoming race. Let Y denote the runner's race time and let μ denote what his or her theoretical average time would be over all training and nutritional regimens that he or she would consider and over all possible weather conditions. If no controllable or uncontrollable factors could be identified that would affect the runner's time, then the model for the race time would be

$$Y = \mu + \epsilon$$

with ϵ denoting a random error term that represents that the race time should vary in a random manner from the overall mean.

If this were the true model, then all attempts at discovering the factors that affect this person's race time would be unsuccessful. But we know this cannot be the correct model because, at the very least, weather conditions will have an affect. Weather conditions are, of course, uncontrollable, and so being able to identify weather conditions as an important factor would not be of great value to our runner. However, he or she would still be interested in knowing the effect of weather conditions on performance, just as a company would like to know how its products perform when customers use the products in some way other than the intended manner.

The runner would naturally prefer not to be greatly affected by weather conditions nor by the difficulty of the course, just as a toy manufacturer would not want its toys to fall apart if children are somewhat rough on them.

In experimental design applications we want to be able to identify both controllable and uncontrollable factors that affect our response variable (Y). We must face the fact, however, that we cannot expect to identify all of the relevant factors and the true model that is a function of them. As G. E. P. Box stated (e.g., Box, 1976), "All models are wrong, but some are useful." Our objective, then, is to identify a useful model, $Y = f(X_1, X_2, \ldots, X_k) + \epsilon$, with $X_1, X_2, \ldots, X_k$ having been identified as significant factors. Each factor is either *quantitative* or *qualitative*, and a useful model might contain a mixture of the two. For example, the type of breakfast that a runner eats would be a qualitative factor.

Since we will never have the correct model, we cannot expect to run a single experiment and learn all that we need to learn from that experiment. Indeed, Box (1993) quoted R. A. Fisher: "The best time to design an experiment is after you have done it." Thus, experimentation should (ideally) be sequential, with subsequent experiments designed using knowledge gained from prior experiments, and budgets should be constructed with this in mind. Opinions do vary on how much of the budget should be spent on the first experiment. Daniel (1976) recommends using 50–67 percent of the resources on the first experiment, whereas Box, Hunter, and Hunter (1978) more stringently recommend that at most 25 percent of the resources be used for the first experiment. Since sequential experimentation could easily involve

more than two experiments, depending upon the overall objective(s), the latter seems preferable.

1.3 PLANNED EXPERIMENTATION VERSUS USE OF OBSERVATIONAL DATA

Many universities model college grade point average (GPA) as a function of variables such as high school GPA and aptitude test scores. As a simple example, assume that the model contains high school GPA and SAT total as the two variables. Clearly these two variables should be positively correlated. That is, if one is high the other will probably also be high. When we have two factors (i.e., variables) in an experimental design, we want to isolate the effect of each factor and also to determine if the interaction of the two factors is important (interaction is discussed and illustrated in detail in Section 4.2).

A factor can be either *quantitative or qualitative*. For a quantitative factor, inferences can be drawn regarding the expected change in the response variable per unit change in the factor, within the range of the experimentation, whereas, say, the "midpoint" between two levels of a qualitative factor, such as two cities, generally won't have any meaning. Quantitative and qualitative factors are discussed further in Section 1.6.2.2.

For the scenario just depicted, we do not have an experimental design, however. Rather, we have observational data, as we would "observe" the data that we would obtain in our sample of records from the Registrar's office. We can model observational data, but we cannot easily determine the separate effects of the factors since they will almost certainly be correlated, at least to some degree.

However, assume that we went to the Registrar's office and listed 25 combinations of the two variables that we wanted, and the student's college GPA was recorded for each combination. Since the values of the two variables are thus fixed, could we call this planned experimentation? No, it is still observational data. Furthermore, it would be nonrandom data if we wanted our "design" to have good properties, as we would, for example, be trying to make the two variables appear to be uncorrelated (i.e., an orthogonal design), which are actually highly correlated. So the results that were produced would probably be extremely misleading.

Returning to the runner example, let's say that our runner uses two nutritional supplement approaches (heavy and moderate), and two training regimes (intense and less intense). He wants to isolate the effects of these two factors, and he will use a prescribed course and record his running time. Assume that he is to make four runs and for two of these runs he uses a heavy supplement approach and an intense training regime, and for the other two he uses a moderate supplement approach and a less intense training regime.

Would the data obtained from this experiment be useful? No, this would be a classic example of how *not* to design an experiment. If the running time decreased when the intensity of the training regimen increased, was the decrease in running time due to the training regimen change or was it due to the increase in supplementation? In

statistical parlance, these two effects are completely *confounded* and cannot be separated. (The terms *confounding* and *partial confounding* are discussed and illustrated in Section 5.1.)

Obviously the correct way to design the experiment if four runs are to be used is to use all four combinations of the two factors. Then we could identify the effects of each factor separately, as will be seen in Section 4.1 when we return to this example.

1.4 BASIC DESIGN CONCEPTS

Assume that a math teacher in an elementary school has too many students in her class one particular semester, so her class will be split and she will teach each of the two classes. She has been considering a new approach to teaching certain math concepts, and this unexpected turn of events gives her an opportunity to test the new approach against the standard approach. She will split the 40 students (20 boys and 20 girls) into two classes, and she wonders how she should perform the split so that the results of her experiment will be valid.

One obvious possibility would be to have the boys in one class and the girls in the other class. In addition to being rather unorthodox, this could create a *lurking variable* (i.e., an extraneous factor) that could undermine the results since it has been conjectured for decades that boys may take to math better than do girls. What if the split were performed alphabetically? Some people believe that there is a correlation between intelligence and the closeness to the beginning of the alphabet of the first letter in the person's last name. Although this is probably more folklore than fact, why take a chance? The safest approach would obviously be to use some random number device to assign the students to the two classes. That is, *randomization* is used. (Although this would likely create different numbers of boys and girls in each class if the 40 students were randomly divided between the two classes, the imbalance would probably be slight and not of any real concern.)

1.4.1 Randomization

IMPORTANT POINT

Randomization should be used *whenever possible and practical* so as to eliminate or at least reduce the possibility of confounding effects that could render an experiment practically useless.

Randomization is, loosely speaking, the random assignment of factor levels to experimental units. Ideally, the randomization method described by Atkinson and Bailey (2001) should be used whenever possible, although it is doubtful that hardly any experimenters actually use it. Specifically, they state, "In a completely randomised design the treatments, with their given replications, are first assigned to the experimental units systematically, and then a permutation is chosen at random from the $n!$ permutations

of the experimental units (p. 57)." This is preferable to assigning the treatments (i.e., factor levels) at random to the experimental units, because a random assignment if performed sequentially will result, for example, in the last factor level being assigned to the last available experimental unit, which is clearly not a random assignment. The randomization method espoused by Atkinson and Bailey (2001) avoids these types of problems. Of course we could accomplish the same thing by, assuming t treatments, randomly selecting one of the $t!$ orderings, and then randomly selecting one of the $n!$ permutations of the experimental units and elementwise combining the juxtaposed lists.

Randomization is an important part of design of experiments because it reduces the chances of extraneous factors undermining the results, as illustrated in the preceding section. Czitrom (2003, p. 25) stated, "The results of many semiconductor experiments have been compromised by lack of randomization in the assignment of the wafers in a lot (experimental units) to experimental conditions."

Notice the words "whenever possible and practical" in italics in the Important Point, however, as randomization should not automatically be used.

In particular, randomization is not always possible, and this is especially true in regard to a randomized run order, as it will often not be possible to change factor levels at will and use certain combinations of factor levels. If randomization is not performed, however, and the results are unexpected, it may be almost impossible to quantitatively assess the effect of any distortion caused by the failure to randomize. This is an important consideration.

There are various detailed discussions of randomization in the literature, perhaps the best of which is Box (1990). The position taken by the author, which is entirely reasonable, is that randomization should be used if it only slightly complicates the experiment; it should not be used if it more than slightly complicates the experiment, but there is a strong belief that process stability has been achieved and is likely to continue during the experiment; and the experiment should not be run at all if the process is so unstable that the results would be unreliable without randomization but randomization is not practical.

The issue of process stability and its importance is discussed further in Section 1.7.

Undoubtedly there are instances, although probably rare, when the use of randomization in the form of randomly ordering the runs can cause problems. John (2003) gave an example of the random ordering of runs for an experiment with a 2^4 design (covered in Chapter 4) that created a problem. Specifically, the machinery broke down after the first week so that only 8 of the 16 runs could be made. Quoting John (2003), "It would have been so much better if we had not randomized the order. If only we had made the first eight points be one of the two resolution IV half replicates. We could have also chosen the next four points to make a twelve-point fraction of resolution V, and, then, if all was going well, complete the full factorial." (These designs are covered in Chapters 4 and 5.)

1.4.2 Replication versus Repeated Measurements

Another important concept is *replication*, and the importance of this (and the importance of doing it properly) can be illustrated as follows.

IMPORTANT POINT

Replication should be used whenever possible so as to provide an estimate of the standard deviation of the experimental error. It is important to distinguish between replicates and multiple readings. To replicate an experiment is to start from scratch and repeat an entire experiment, not to simply take more readings at each factor-level condition without resetting factor levels and doing the other things necessary to have a true replicated experiment.

The distinction between replication and multiple readings is an important one, as values of the response variable Y that result from replication can be used to estimate σ_ϵ^2, the variance of the error term for the model that is used. (Multiple readings, however, may lead to underestimation of σ_ϵ^2 because the multiple readings might be misleadingly similar.) Values of Y that result from experiments that do not meet all the requirements of a replicated experiment may have variation due to extraneous factors, which would cause σ_ϵ^2 to be overestimated, with the consequence that significant factors may be erroneously declared not significant. For the moment we will simply note that many experiments are performed that are really not true replicated experiments, and indeed the fraction of such experiments that are presumed to be replicated experiments is undoubtedly quite high. One example of such an experiment is the lead extraction from paint experiment described in Ryan (2004), which although being "close" to a replicated experiment (and assumed to be such) wasn't quite that because the specimens could not be ground down to an exact particle size, with the size of the specimen expected to influence the difficulty in grinding to the exact desired particle size. Thus, the experimental material was not quite identical between replicates, or even within replicates. Undoubtedly, occurrences of this type are very common in experimentation.

One decision that must be made when an experiment is replicated is whether or not "replications" should be isolated as a factor. If replications are to extend over a period of time and the replicated observations can be expected to differ over time, then replications should be treated as a factor.

1.4.3 Example

Let's think back a century or more when there were many rural areas, and schools in such areas might have some very small classes. Consider the extreme case where the teacher has only two students; so one student receives one method of instruction and the other student receives the other method of instruction. Then there will be two test scores, one for each method.

We could see which score is larger, but could we draw any meaningful conclusion from this? Obviously we cannot do so. We would have no estimate of the variability of the test scores for each method, and without a measure of variability we cannot make a meaningful comparison.

Now consider the other extreme and assume that we start with 600 students so that 300 will be in each class (of course many college classes are indeed of this size, and larger).

What do we gain by having such a large *sample size*? Quite frankly, we may gain something that we don't want. We are in essence testing the hypothesis that the average score will be the same for the two methods, if the process of selecting a set of students and splitting the group into two equal groups were continued a very large number of times. The larger the sample sizes, the more likely we are to conclude that there is a difference in the true means (say, μ_1 for the standard method and μ_2 for the new method), although the actual difference might be quite small and not of any practical significance. (The determination of an appropriate sample size has been described by some as a way of equating statistical significance with practical significance.)

From a practical standpoint we *know* that the true means are almost certainly different. If we record the means to the nearest hundredth of a point (e.g., 87.45), is there much chance the means could be the same? Of course not. If we rounded the means to the nearest integer, there would be a reasonable chance of equality, but then we would not be using the actual means.

The point to be made is that in some ways *hypothesis testing* is somewhat of a mindless exercise that has been criticized by many, although certain types of hypothesis tests, such as testing for a normal distribution and hoping that we don't see a great departure from normality, do make sense and are necessary. See, for example, Nester (1996) and the references cited therein regarding hypothesis testing.

A decision must be reached in some manner, however, so the teacher would have to decide the smallest value of $\mu_2 - \mu_1$ that he or she would consider to be of practical significance. Let Δ denote this difference, so that the alternative hypothesis is H_a : $\mu_2 - \mu_1 > \Delta$. If the standard method has been used for many semesters, a reasonably good estimate of σ_1, the standard deviation of scores for that method, is presumably available. If we assume $\sigma_1 \doteq \sigma_2$ (probably not an unrealistic assumption for this scenario), then following Wheeler (1974), using a significance level of $\alpha = .05$ and a probability of .90 of detecting a difference of at least Δ, we might determine the total sample size, n, as

$$n = \left(\frac{4r\sigma}{\Delta}\right)^2 \tag{1.1}$$

with r denoting the number of levels, 2 in this case, of the factor "teaching method." Thus, for example, if $\sigma_1 = \sigma_2 = \sigma = 15/8 = 1.875$ and the teacher selects $\Delta = 3$, then $n = 25$ students so use 26 in order to have 13 in each of the two classes.

Equation (1.1), although appealing because of its simplicity and for that reason has probably been used considerably and has been mentioned in various literature articles (e.g., Lucas, 1994), is an omnibus formula that does not reduce to the exact expression when $r = 2$. Furthermore, Bowman and Kastenbaum (1974) pointed out that Eq. (1.1) resulted from incorrectly applying the charts of Pearson and Hartley (1972). More specifically, Bowman and Kastenbaum (1974) stated that Eq. (1.1) is based on the false assumption that values of φ are constant, with $\varphi = [\delta^2/(\nu_1 + 1)]^{1/2}$,

with δ^2 denoting the noncentrality parameter and $\nu_1 + 1$ denoting the number of levels of a factor.

An important general point made by Wheeler (1974) is that when an effect is not significant, the experimenter should state that if the factor has an effect, it is less than approximately Δ. Clearly this is preferable to stating that the factor has no effect, which is the same as saying that $\Delta = 0$, a statement that would not be warranted.

The appropriate expression for the number of observations to be used in *each* of $r = 2$ groups is given in many introductory statistics books and is

$$\frac{n}{2} = \frac{(z_\alpha + z_\beta)^2\left(\sigma_1^2 + \sigma_2^2\right)}{\Delta^2} \tag{1.2}$$

Using this formula produces

$$\begin{aligned}\frac{n}{2} &= \frac{(1.645 + 1.28)^2(1.875^2 + 1.875^2)}{3^2} \\ &= 6.68\end{aligned}$$

with $1.645 = z_{.05}$ and $1.28 = z_{.10}$ being the standard normal variates corresponding to $\alpha = .05$ and the power of the test of .90, respectively. Thus, 7 students would be used in each class rather than 13, which is the result from the use of Eq. (1.1).

There are various other methods available for determining sample sizes in designed experiments, such as the more complicated iterative procedure given by Dean and Voss (1999, p. 50). The utility of Eq. (1.1) of course lies in its simplicity, although its approximate nature should be kept in mind and variations of it will be needed for certain types of designs, with some variations given by Wheeler (1974). If the test averages for the two classes, denoted by $\overline{y}_1$ and $\overline{y}_2$, respectively, are 79.2 and 75.8, then

$$\begin{aligned}z &= \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{2\sigma^2/n}} \\ &= \frac{75.8 - 79.2}{\sqrt{2(15/8)^2/13}} \\ &= -4.62\end{aligned}$$

Since, assuming (approximate) normality for the statistic z, $P(z < -4.62 \mid \mu_1 = \mu_2) = 1.9 \times 10^{-6}$, we would conclude that there is a significant difference between the two teaching methods.

Notice that this computation is based on the assumption that σ_1 and σ_2 were known and that $\sigma_1 = \sigma_2$. Generally we want to test assumptions, so it would be advisable to use the data to test the assumption that the two variances are equal. (Of course the standard deviations will be equal if the variances are equal but the proposed tests are for testing the equality of the variances.) Preferably, we should use a test that is not sensitive to the assumption of normality, and tests such as those given by Layard

(1973) and Levene (1960) are therefore recommended, in addition to the Brown and Forsythe (1974) modification of Levene's test. (The latter is used in Section 2.1.2.1.1.)

We should also test the assumption of normality of each of the two populations. This can be done graphically by using normal probability plots (see, e.g., Section 4.9) and/or by using numerical tests. Preferably, the two types of tests should be used together.

1.4.4 Size of an Effect That Can Be Detected

It is useful to turn Eq. (1.2) and similar formulas around and solve for Δ. Doing so produces

$$\Delta = \frac{(z_\alpha + z_\beta)2\sigma}{\sqrt{n}} \tag{1.3}$$

assuming $\sigma_1 = \sigma_2 = \sigma$. For the example in Section 1.4.3, we thus have

$$\begin{aligned} \Delta &= \frac{2.925(2\sigma)}{\sqrt{n}} \\ &= \frac{5.850\sigma}{\sqrt{n}} \end{aligned}$$

With $n = 14$, the smallest difference that can be detected with a probability of .90 and a significance level of $\alpha = .05$ is $1.56\sigma = 1.56(1.875) = 2.925$, which is slightly less than 3 because the sample size was rounded up to the next integer. (We should keep in mind that Eq. (1.3) is for a one-sided test.)

We will return to Eq. (1.3) and related formulas in subsequent chapters when we consider the magnitude of effects that can be detected with factorial designs (covered in Chapter 4) and other designs, especially small factorial designs, because it is important to know the magnitude of effect sizes that can be detected. This is something that is often overlooked. Indeed, Wheeler (1974, p. 200) stated, "The omission of such statements (crude though the numbers in them may be) is a major shortcoming of many statistical analyses."

There are various Java applets that can determine n, or Δ for a given value of n; perhaps the best known of these is the one that is due to Russ Lenth, which is found at http://www.stat.uiowa.edu/~rlenth/Power/index.html. Entering $n = 9, \sigma = 1.87$, and $\Delta = 3$, results in a power of .8896. There will not be exact agreement between the results obtained using this applet and the results using the previously stated equations, however, because the latter are based on the use of z, whereas that is not one of the options when the applet is used. Instead, these numbers result when the use of a t-test is assumed.

Software can of course also be used to compute power, and Design-Expert can be used for this purpose for any specific design.

In addition to these applets and software, Lynch (1993) gave tables for use in determining the minimum detectable effects in two-level fractional factorial designs

(i.e., 2^{k-p} designs), which are covered in Chapter 5. These tables were computed using the noncentrality parameter of the t-distribution, which was given as

$$\lambda = \left(\frac{\Delta}{\sigma}\right)\frac{\sqrt{n}}{2}$$

with n denoting the total number of runs in the experiment and the test statistic given by

$$t = \frac{\text{Effect estimate}}{2(s_{\text{p}}/\sqrt{n})}$$

with s_{p} denoting the square root of the pooled estimate of σ^2, and s_{p}^2 given by

$$s_{\text{p}}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

with s_1^2 and s_2^2 denoting the sample variances for the first and second levels of the factor, respectively, and n_1 and n_2 denoting the corresponding sample sizes, with $n_1 + n_2 = n$.

The results essentially show that 2^{k-p} designs with $2^{k-p} < 16$ (i.e., 8-point designs) have poor detection properties. This is discussed in more detail in Section 5.1.

A general method for computing power for a variety of designs, including many that are given in this book, was given by Oehlert and Whitcomb (2001).

1.5 TERMINOLOGY

The terms *randomization* and *replication* were used in Sections 1.4.1 and 1.4.2, respectively. There are other terms that will be used frequently in subsequent chapters. In the example involving the math teacher, which was given in Section 1.4.3, the students are the *experimental units* to whom the two *treatments* (i.e., the two methods of teaching the class) are applied.

In that experiment the possibility of having all the girls in one class and all the boys in the other class was mentioned—and quickly dismissed. If the experiment had been conducted in this manner, this would be an example of an experiment in which factors are confounded. That is, we would estimate the gender effect—if we were interested in doing so—by taking the difference of the average of the girls' scores on the first test and the average of the boys' scores on that test. But this is exactly the same way that we would estimate the teaching method effect. Thus, one number would estimate two effects; so we would say that the effects are *confounded*. Obviously we would want to avoid confounding the two effects if we believe that they both may be statistically significant. Therefore, confounding, which is essentially unavoidable in most experiments, due to cost considerations when more than a few