

## 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<sup>®</sup>, which has certain features reminiscent of expert systems software, JMP<sup>®</sup>, and MINITAB<sup>®</sup>. (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

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

## 1.2 OBJECTIVES FOR EXPERIMENTAL DESIGNS

3

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  $\mu$  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  $\epsilon$  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,  $Y = f(X_1, X_2, \dots, X_k) + \epsilon$ , with  $X_1, X_2, \dots, X_k$  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

1.3 PLANNED EXPERIMENTATION VERSUS USE OF OBSERVATIONAL DATA 5

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  $\sigma_{\epsilon}^2$ , the variance of the error term for the model that is used. (Multiple readings, however, may lead to underestimation of  $\sigma_{\epsilon}^2$  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  $\sigma_{\epsilon}^2$  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,  $\mu_1$  for the standard method and  $\mu_2$  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  $\Delta$  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  $\sigma_1$ , 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  $\Delta$ , 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  $\varphi$  are constant, with  $\varphi = [\delta^2/(\nu_1 + 1)]^{1/2}$ ,

with  $\delta^2$  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  $\Delta$ . 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

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

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  $\bar{y}_1$  and  $\bar{y}_2$ , respectively, are 79.2 and 75.8, then

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

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

Notice that this computation is based on the assumption that  $\sigma_1$  and  $\sigma_2$  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  $\Delta$ . Doing so produces

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

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  $\Delta$  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

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

with  $s_p$  denoting the square root of the pooled estimate of  $\sigma^2$ , and  $s_p^2$  given by

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

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

factors are involved, must be used judiciously. This is discussed in detail in later chapters, especially Chapter 5.

We should bear in mind, however, that true randomization will often not be possible. That is, whereas it would seem to be easy to randomly assign the girls and the boys to the two classes, physical variables can present insurmountable obstacles. For example, it might be impossible or at least impractical to make frequent temperature changes to ensure randomness, and certain combinations of temperature and other variables might not even be possible. Joiner and Campbell (1976) were among the first to discuss such problems in the statistics literature, and the reader is also referred to papers by Ganju and Lucas (1997, 1999), Bailey (1987), and Youden (1972). Hard-to-change factors and debarred observations are discussed in Sections 4.19 and 13.8, respectively.

Another important concept is *blocking*. It is often said that an experimenter should randomize over factors that can be controlled and block factors that cannot be controlled. For example, suppose that a farmer wants to conduct an experiment to compare two types of feed because he is interested in finding the best feed that will allow him to fatten up his pigs for market as quickly as possible. How quickly the pigs gain weight on a given feed is likely to depend somewhat on the litter that they are from, so it would be logical to use litter as a blocking variable. The use of blocking variables is illustrated in Chapter 3.

## 1.6 STEPS FOR THE DESIGN OF EXPERIMENTS

The steps that one should follow in designing and conducting experiments can be laid out in a very general way, although a cookbook set of steps cannot (or at least should not) be provided. This is because the specifics of procedures to follow will vary somewhat from setting to setting. Nevertheless, Bisgaard (1999) provided a template that is appropriate for factorial experiments (Chapters 4 and 5) that are to be used in a sequential manner. The starting point of course would be a statement of the reason for the experiment(s) and the objective(s). This should be followed by a list of the factors to be studied and the levels of each, a statement of the response variable(s) and how the measurements will be conducted. Bisgaard (1999) went on to list eight subsequent steps that are mostly nonstatistical and include using a flowchart and data collection sheets, planning for problems, assigning responsibilities to team members, and so on. One recommendation that is not generally mentioned in conjunction with designing experiments is the idea of using a Gantt chart that shows the steps to be followed and the associated dates, with a small pilot experiment included. (There is much information on Gantt charts on the Internet. Informative sources are the brief tutorial at <http://www.me.umn.edu/courses/me4054/assignments/gantt.html> and at the web site devoted to Gantt charts: <http://www.ganttchart.com>.)

Coleman and Montgomery (1993) also gave a thorough discussion of considerations that should be made in designing an experiment. They list seven steps that should be performed sequentially: (1) recognition of and statement of the problem, (2) choice of factors and levels, (3) selection of the response variable(s), (4) choice

of experimental design, (5) conduction of the experiment, (6) data analysis, and (7) conclusions and recommendations. See also Van Matre and Diamond (1996–1997) for a discussion of the importance of teamwork in designing and carrying out experiments, as well as Bisgaard (1989), who gave an interesting example of team problem solving.

We will consider the first two steps in some detail and will provide some additional insight.

### 1.6.1 Recognition and Statement of the Problem

Montgomery (1996) points out that the problem statement is often too broad. The problem should be specific enough and the conditions under which the experiment will be performed should be understood so that an appropriate design for the experiment can be selected.

### 1.6.2 Selection of Factors and Levels

This issue has been addressed by Hahn (1977, 1984) and Cox (1958), in addition to Coleman and Montgomery (1993), with a more recent and more extensive discussion given by Czitrom (2003). We will review their recommendations and provide some additional insight.

#### 1.6.2.1 Choice of Factors

The factors that are studied in the initial stages of sequential experimentation are those that are believed to be important. The set can be reduced in later stages, so it is better to start with a large set than with a small set that may not include some important factors. Of course if an experimenter knew which factors were important, then the number of stages normally used in experimentation could be reduced, but such prior knowledge is generally unavailable. Sometimes a factor may not be recognized as important simply because its level isn't changed. Indeed, Myers and Montgomery (1995, p. 636) stated, "Often we have found that a variable was not realized to be important simply because it had never been changed."

See also Barton (1997), who presented a method for identifying and classifying seemingly important factors before an experiment is run. In particular, Barton uses the term *intermediate variable* in referring to variables that may be related to the dependent value, but the values of intermediate variables are determined by the values of independent variables. An example of this was given. This is an important point because the factors used in an experimental design must be true independent variables so that the levels of the factors can be selected without having to worry about what levels will be used for the other factors. There will frequently be restrictions on combinations of factor levels that translate into restriction regions of operability (Section 13.8) and debarred observations (also Section 13.8), but that is quite different from having *every* possible level of a factor dependent upon the levels of other factors, as is the case for an intermediate variable. There is a class of designs, namely, supersaturated designs (Section 13.4.2) for which the sets of factor levels used in the design

will be slightly correlated, but this is a consequence of the design (and specifically because the number of design points is less than the number of factors), not because of a deterministic relationship between any of the factors.

### 1.6.2.2 Choice of Levels

The experimenter must also address the question of how many levels to use and how the levels are to be selected. If only two or three levels are likely to be used in the immediate future, then those levels should be used in the experiment and the inference that is drawn will apply only to those levels. If, however, there is interest in a range of possible levels for the factors of interest, then those levels should be randomly selected from that range. In the first case the factor(s) would be classified as *fixed*; in the second case the factor(s) would be classified as *random*. The distinction between a fixed factor and a random factor is important, as will be seen in succeeding chapters.

Briefly, if a factor were fixed, we would be interested in testing the equality of the population means for those levels. If a factor were random, we would be interested in testing whether or not the variance of the effects of the factor levels is zero for the range of levels of interest. In designs involving more than one factor, the classification of factors as fixed or random determines how certain hypothesis tests are performed. So this distinction is quite important.

In the initial stages of an experimental investigation, the first objective is to determine which factors are important. This necessitates the use of a *screening design* if there are more than a few factors that are being considered. Screening designs usually have two levels, this being necessary because  $p^k$  can be quite large when  $k$  is not small and when  $p > 2$ , with  $k$  denoting the number of factors and  $p$  denoting the number of levels for each factor, with  $p$  usually the same for each factor. Thus,  $p^k$  is the number of observations unless the experiment is replicated.

The number of levels can be increased in the secondary stages of a sequential investigation, after the relevant factors have been identified; this is illustrated in later chapters, including Chapter 10, where it would logically be done since response surface designs generally do not have a large number of factors, but the important factors must be defined first.

Now assume that  $k = 1$ . For qualitative factors, the selection of levels is generally not a major issue, and indeed the word “levels” in the case of alternatives to a standard process that would be compared to the standard process in an experiment simply denotes the number of such alternatives to be examined in an experiment plus one (the standard process).

For quantitative factors, the selection of factor levels can be critical. Assume that we have one factor with two levels, with the factor denoted by  $X$  and  $X_1$  and  $X_2$  denoting the low level and the high level, respectively.

How should the levels be chosen? We obviously want to detect a significant effect of the factor if it exists. We have to exercise some care in defining what we mean by “effect of the factor” as well as variability of the response. In particular, we cannot speak of the latter independent of the levels of a factor.

For the sake of illustration, assume that the response variable has a normal distribution over values of a factor that has not been controlled but has freely varied within

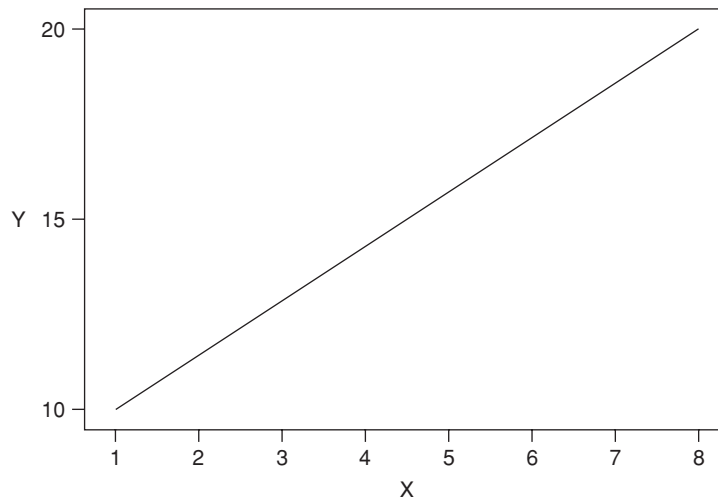


Figure 1.1 Relationship between  $Y$  and  $X$ .

a wide range. Now assume that the response variable ranges from  $\mu - \sigma$  to  $\mu + \sigma$  when the factor of interest is within a specific range, and the endpoints of the range are to be used in an experiment to determine if the factor is significant.

Of course we don't know what the range of the response will be—which is why we conduct experiments—but what should happen if we conduct an experiment with the endpoints of the specific range of the factor used as the two levels in an experiment, with  $n$  observations used at each level? Let's say that the two values of the factor are fixed within the aforementioned range, and assume that the conditional distribution of  $Y|X_1 \sim \text{Normal}(\mu_1, (\sigma^*)^2)$  and  $Y|X_2 \sim \text{Normal}(\mu_2, (\sigma^*)^2)$ . That is, the conditional distributions have a common variance,  $(\sigma^*)^2$ , and of course different means.

The questions to be addressed are as follows: (1) is the difference between  $\mu_1$  and  $\mu_2$  of practical significance, and if so, (2) are  $X_1$  and  $X_2$  far enough apart, relative to  $\sigma^*$ , to allow the difference to be detected with a high probability? If the expected change is only  $\sigma^*$  and the spread from  $X_1$  to  $X_2$  is the largest spread that could be tolerated in, say, a manufacturing setting, then there is not much point in performing experimentation with the factor  $X$ . These are technical questions, the answers to which are beyond the intended level of this chapter.

At this point a simple graphical answer may suffice. Consider Figure 1.1.

Assume that 1–8 on the  $X$ -axis denote possible levels of  $X$ , all of which are feasible, with the line representing the relationship between  $Y$  and  $X$ , which we would expect to eventually become nonlinear if the line were extended far enough in each direction. Assume that an experiment is conducted with levels corresponding to  $X = 2$  and  $X = 3$ . The difference between the corresponding  $Y$ -values in the graph is approximately 1.5, which is thus the difference between the conditional expectations  $E(Y|X_1)$  and  $E(Y|X_2)$ . Let the estimators of these conditional means be denoted by

$\bar{Y}_1$  and  $\bar{Y}_2$ , respectively. Assume that the conditional variances are also equal and are denoted by  $(\sigma^*)^2$ , in accordance with the previous notation, and for the sake of illustration we will assume that this is known. Since  $\text{Var}(\bar{Y}_1 - \bar{Y}_2) = (\sigma^*)^2(2/n)$ , assuming an equal number of observations,  $n$ , at each value of  $X$ , we might ask whether  $E(\bar{Y}_1 - \bar{Y}_2)$  exceeds  $2(\sigma^*)\sqrt{(2/n)}$  (i.e., whether the expected value of the test statistic exceeds 2), as a value of the test statistic much greater than 2 would certainly suggest a difference in the conditional means for the two values of  $X$ .

Obviously the further apart the  $X$ -values are, the more likely this condition is to be met, as is apparent from Figure 1.1. Realize, of course, that spreading out the  $X$ -values will increase the spread of the  $Y$ -values, which will increase the denominator of the test statistic. Furthermore, this apparent desiderata regarding the spread of the  $X$ -values is offset by the fact that the wider the ranges of factors, the more likely interactions between factors are to be judged significant. We will see in Section 4.2 and subsequent sections how significant interactions complicate analyses.

Of course relationships such as the one depicted in Figure 1.1 will in general be unknown. Czitrom (2003, p. 14) states, "In the absence of other engineering considerations, attempt to give each factor, a priori, an "equal opportunity" to influence the result, or set the factors at, say, 10% above and below a value of interest such as the current operating conditions." Of course, feasible factor levels will frequently serve as boundaries for level settings in experimentation.

Although we speak of the variability of the response variable for a given factor, our real interest is in examining the variability of the response between factor levels, as the extent of that variability determines whether or not the factor is significant. Accordingly, the levels must be far enough apart to allow the response variable the opportunity to vary enough between levels to result in the factor being judged significant.

At times we will discuss important nonstatistical considerations that should be made in designing experiments. Detailed discussions of steps (nonstatistical steps) that are followed for an actual experiment are generally not given in the literature; an exception is Vivacqua and de Pinho (2004), which relates to the subject matter in Chapter 9 and is discussed there. We might view the selection of factors and factor levels as being almost nonstatistical considerations, but certainly very important considerations. As Czitrom (2003) stated, "... detailed guidelines for selecting factors and factor levels, which are critical inputs to an experiment, are seldom presented systematically."

The discussion to this point has been restricted to placement of the levels of factors. *If* only a single experiment were to be run, which of course is usually not a good idea, the *number* of such levels would be dictated by our belief regarding the order of a likely model for the response. Any such prior beliefs are of secondary importance when experimentation is sequential, however (as preferred), and an absence of such prior beliefs would be one way to motivate sequential experimentation.

Three levels are of course needed to detect curvature, and three levels might be a substitute for four levels if a 3-level factor had the same extreme levels as would be used if four levels were employed. Five levels are involved in central composite designs, which are response surface designs (Chapter 10), and although Czitrom

(2003, p. 16) mentions an experiment in which 20 levels were used to capture a sinusoidal behavior in the response variable, we would rarely want to use more than five levels. Generally the levels that are used should be equally spaced. Sometimes the levels of factors won't be known and will have to be approximated. In such cases the levels will almost certainly be unequally spaced.

### 1.7 PROCESSES SHOULD IDEALLY BE IN A STATE OF STATISTICAL CONTROL

Although this is a book on design of experiments, and not on process control, the importance of maintaining processes in a state of statistical control when designed experiments are performed cannot be overemphasized. In particular, process drift could cause a new setting for a process variable to appear to produce significantly worse results than the standard setting.

Simply stated, when designed experimentation is performed, controllable factors that could influence the value of the response variable should be kept in a state of statistical control. Ideally, the mean of the response variable should not be affected by controllable extraneous factors during an experiment, but can this idealized state be achieved? Some would say that it cannot be achieved. Box (1990) stated, in responding to the question of whether statistical control can be achieved: "The truth is that we all live in a nonstationary world; a world in which external factors never stay still. Indeed the idea of stationarity—of a stable world in which, without our intervention, things stay put over time—is a purely conceptual one." If we accept this, then we should strive for processes being in a state of near-statistical control. That might not seem particularly difficult to achieve, but Bisgaard (1994) stated that technicians often do not have very good control of their processes during the early stages of experimentation.

Does this mean that control should not be sought during experimentation? Various industrial personnel, who have learned the hard way what the consequences of out-of-control processes can be, would respond with a resounding "No!"

An example of how an out-of-control process can lead to misleading results was given in Ryan (2000, p. 361), using hypothetical data. Specifically, assume that an experiment is being conducted using two levels of a factor (e.g., the standard temperature level versus an experimental level). If a variable that is not under study suddenly goes out of statistical control (generally defined as no longer being within three standard deviations of its mean), and this causes a sharp change in the response variable, this change could erroneously be attributed to a change in the level of the factor under study if that change occurred near the point at which the extraneous variable went out of control. This problem can happen frequently in industrial and other experiments if care is not exercised.

If factors (variables) not part of a study are out of control, how can this be detected? One suggested approach is to make runs at the standard operating conditions at the beginning and end of an experiment and compare the results (see Croarkin and Tobias, 2002, Section 5.2.2 of the *NIST/SEMATECH e-Handbook of Statistical Methods*:

1.7 PROCESSES SHOULD IDEALLY BE IN A STATE OF STATISTICAL CONTROL **19**

<http://www.itl.nist.gov/div898/handbook/pri/section2/pri22.htm>). If the results differ greatly, one or more of the variables not in the study are probably out of control, although of course the variance of the difference of the two numbers is  $2\sigma^2$  when/if the process(es) is in control, so the difference between the numbers may not be small when a state of statistical control does exist if process variability is relatively large but stable.

Unfortunately, this placement of check runs is seldom done, however, as additional design points are generally not strategically placed among the experimental runs. *This is a major weakness in the use of experimental designs.*

Of course, process control techniques should be applied routinely to variables that could easily go out of control, regardless of whether experiments that involve those variables are being performed. Unfortunately, the need to have processes in control when experiments are performed is generally not recognized in the statistical literature. Notable exceptions are the experiment described in Jiang, Turnbull, and Clark (1999) and case study #2 in Inman et al. (1992). Clark (1999) stated the importance of statistical control about as well as it can be stated in discussing the importance of the tooling process being in control for a resistance welding operation experiment that was undertaken: "Control of the tooling proves to be as important as parameter evaluation is in achieving optimum nugget size in resistance welding. Otherwise you are trying to make a prediction about an unpredictable event. Without stability, you have nothing." See also the discussion about the need for processes being in control in Arvidsson and Gremyr (2003).

The use of statistical process control (SPC) methods to ensure that processes are in control when experimentation is performed is very important and so we will at times reiterate the importance of this in subsequent chapters.

Let's assume that a process cannot be brought under a state of statistical control. What should be done? It might seem as though experiments simply should not be run under such conditions, but Box, Bisgaard, and Fung (1990, p. 190) recommend that processes be brought into the best state of control possible, and experimentation be performed in such a way that blocks are formed to represent intervals of time when the process appears to be in control. One potential impediment to the implementation of such an approach, however, is that it may be difficult to determine when changes have occurred, unless the changes are large. There are, however, methods that have been developed for trying to determine change points.

The safest approach would be to strive for tight control of processes so that the results of experiments can be interpreted unambiguously. Since statistical control methods should be used anyway, at least for critical processes, this should not entail any extra work. Furthermore, if these statistical process control methods are in force, check runs as described earlier in this section should be unnecessary, although experimenters might still want to use them if experimentation at the standard operating conditions is inexpensive.

The SPC methods that are employed include control charts in addition to more sophisticated and more efficient methods, including cumulative sum (CUSUM) and exponentially weighted moving average (EWMA) techniques. These and other SPC methods are covered extensively in Ryan (2000) and other books on the subject.

## 1.8 TYPES OF EXPERIMENTAL DESIGNS

There are many types of experimental designs, some of which have been in use for over 70 years. Experimental design has its roots in agriculture and much of the pioneering work was performed at the Rothamsted Experiment Station in England. Although the use of experimental designs has broadened into many areas, including such diverse fields as manufacturing and medicine, some of the terminology of the distant past remains, such as the word “plot” in the names of designs (e.g., a split-plot design) even though there will usually not be a physical “plot,” such as a plot of land, in an experiment. The basic objective remains unchanged, however: to determine the factors that are influencing the variability in the response values and to use whichever experimental design will best provide this insight for a given amount of resources.

Many different experimental designs are presented in the succeeding chapters. These include designs for a single factor, with and without blocking, in addition to designs for multiple factors, including a large number of factors. Various realities of design construction must be faced, including the fact that not all combinations of factor levels are possible, and restrictions on randomization are also encountered very frequently. There is as yet no general solution to the problem of design construction when not all factor-level combinations are possible, and only during the past 10 years have the consequences of restricted randomization been discussed extensively in the literature.

The world of experimental design is thus quite different from what it was several decades ago, and contrary to what some might say, there are still important research contributions in certain areas of design that are emerging and more research that is needed.

New methods of analysis are also needed, and conditional effects analyses are presented herein for the first time as a general method of analysis. Established methods that are typically not used with experimental designs must also be considered. This includes Analysis of Means (ANOM), which is presented in the next section and is used in the following chapters in conjunction with various designs.

## 1.9 ANALYSIS OF MEANS

Almost 50 years ago, Ellis Ott, the founding chairman of the statistics department at Rutgers University, needed a better way to communicate with the engineers for whom he performed consulting work, as they found Analysis of Variance (ANOVA) not to be very intuitive. And indeed it isn't intuitive since squared quantities are involved, thus resulting in the unit of measurement being “squared” in the numerical analysis. Although there is squared inches, the square of most units doesn't make any sense, such as squared temperature and squared yield.

So Ott sought a method that did not cause the unit of measurement to be lost, and consequently invented Analysis of Means (ANOM), developing it in Ott (1958), and introducing it into the literature in Ott (1967); the first edition of Ott's text (1975)

contains many illustrative examples. The most recent edition of Ott's book is Ott, Schilling, and Neubauer (2005), but it should be noted that there are people who consider the first edition of Ott's book to be the best version. Indeed, the graphical display of interactions in factorial and fractional factorial designs (those designs are covered herein in Chapters 4 and 5, respectively) has become distorted relative to the way that they were displayed originally by Ott.

The only book on ANOM is Nelson, Wludyka, and Copeland (2005). Other recent books that have a moderate amount of space devoted to ANOM include Ryan (2000) and Nelson, Coffin, and Copeland (2003), with the display of interactions in Ryan (2000) and in this book being the same as in Ott (1975), except for the minor difference that the latter used slanted lines instead of vertical lines.

Ott's 1967 paper appeared in the Walter Shewhart Memorial issue of *Industrial Quality Control*, and it is fitting that the January 1983 issue of the *Journal of Quality Technology* (which was previously named *Industrial Quality Control*) contained articles only on ANOM, with the purpose of serving as a tribute to Ellis Ott upon his passing.

The state of the art has advanced during the past two decades, primarily because of the research contributions of the late Peter Nelson and his students. This research has allowed ANOM to be applied to data from various types of designs with fixed factors. (One limitation is the restriction to fixed factors; recall the discussion of fixed and random factors in Section 1.6.2.2.)

To many people, the term "analysis of means" undoubtedly conjures up notions of analysis of variance (ANOVA), however, which is also concerned with the analysis of means, and which is much better known and more widely used than ANOM. The latter is apt to have more appeal to engineers and other industrial personnel than does ANOVA, however, since ANOM is inherently a graphical procedure and is somewhat similar to a control chart. The importance of using good analytic graphical procedures cannot be overemphasized, and the similarity to a control chart is also a plus since control charts or related procedures should be used in conjunction with experimental design, anyway, as was emphasized in Section 1.7.

Readers familiar with ANOVA will recall that with that method the experimenter concludes that either all of the means are equal or at least one of the means differs from the others. If the latter is concluded, then a multiple comparison procedure, of which there are many, is generally used to determine which means differ significantly.

With ANOM, however, the user will see whether or not one or more means differ from the average of all the means. Thus, what is being tested is different for the two procedures, so the results will not necessarily agree. In particular, when  $k - 1$  sample averages are bunched tightly together but the  $k$ th sample average (i.e., the other one) differs considerably from the  $k - 1$  averages, the  $F$ -value in ANOVA would likely be relatively small (thus indicating that the population means are equal), whereas the difference would probably be detected using ANOM. Conversely, if the differences between adjacent sample averages are both sizable and similar, the (likely) difference in the population means is more apt to be detected with ANOVA than with ANOM.

One procedure need not be used to the exclusion of the other, however. As Ott (1967) indicates, ANOM can be used either alone or as a supplement to ANOVA.

Since the methods produce similar results, it is not surprising that the assumptions for each method are the same. That is, the variances must be the same within each level of the factor (and within each cell, in general, when there is more than one factor), and the populations for each level (cell) must be normally distributed. Of course, these assumptions are almost impossible to check with a small amount of data per level (cell), so fortunately the normality assumption is not crucial and the equal variances assumption is also not a problem for equal numbers of observations per level (cell) unless the variances differ greatly.

When the one or more factors all have two levels, ANOM is simply a “graphical  $t$ -test” and the results will then agree with the results obtained using either the  $t$ -test or ANOVA. Since two-level factors predominate in practice and good graphical procedures are always useful, ANOM is an important technique for analyzing data from designed experiments. Accordingly, it will be used and discussed in subsequent chapters.

Almost all of the ANOM methods are based on the assumption of a normal distribution for the plotted statistic, or the adequacy of the normal approximation to the binomial or Poisson distribution. An exception is the nonparametric procedure of Bakir (1989).

In general, the use of graphical methods in analyzing data from designed experiments is quite important. See Barton (1999) for many illustrative examples as well as standard and novel graphical methods.

## 1.10 MISSING DATA

Missing data is really an analysis problem, not a design problem, but since inferences are drawn from the data that are collected when a designed experiment is performed, the problem (which will often occur, such as when an experimental run is botched) must be addressed. Accordingly, it is discussed in subsequent chapters, such as Section 2.1.3.2.

## 1.11 EXPERIMENTAL DESIGNS AND SIX SIGMA

Because the Six Sigma phenomenon is still going strong after several years, there should probably be at least brief mention of the role of design of experiments in Six Sigma, a topic that is covered more fully in the context of robust designs in Chapter 8. For those readers unfamiliar with the term “Six Sigma,” it refers to a collection of tools (and even a company way of life), both statistical and nonstatistical that are used to improve the quality of products and processes. Experimental design plays a key role in this because the factors that influence product and process quality must be identified. Readers interested in how design of experiments fits into Six Sigma may want to read Goh (2002).

## 1.12 QUASI-EXPERIMENTAL DESIGN

One topic that is *not* covered in this book is *quasi-experimental design*. The term is used to refer to experimentation in which there is no attempt to adhere to the tenets of experimental designs. For example, there is generally no randomization, and studies are often longitudinal with the time aspect playing a prominent role. In some fields, such as the social and behavioral sciences, it isn't possible to conduct experiments in accordance with the principles of experimental design. Quasi-experimental designs can be very useful, especially in the field of education. Readers interested in quasi-experimentation are referred to Campbell and Stanley (2005), Shadish (2001), Shadish, Cook, and Campbell (2001), Trochim (1986), Cook and Campbell (1979), and Manly (1992).

## 1.13 SUMMARY

Experimental design is used in a very long list of application areas, including areas where we might not expect to see it being used, such as marketing, and in areas where we might want to see greater use, such as engineering. The proper use of experimental designs requires considerable thought, however, and various obstacles discussed in later chapters, such as hard-to-change factors and restricted regions of operability, make optimal or at least judicious use of experimental designs a challenging task.

The minimal size of an effect that an experimenter wants to detect will often be used to determine the number of observations and thus the size of the experiment that will be employed.

It is very important that processes that will impact an experiment be in at least a reasonable state of statistical control when an experiment is performed. The consequences of not doing so can be severe, with the results possibly being very misleading. If experimenters are not used to checking for control, results that run counter to engineering knowledge would likely result in a repeated experiment, with the results possibly being quite different from the first experiment if processes are out of control in a way that is different from what occurred when the first experiment was run.

Appropriate analysis of data from designed experiments is also important and experimenters have certain options, including ANOM as a possible supplement to ANOVA or a replacement for it for fixed factors. The effect that restricted randomization, if present, has on analyses must also be considered. This is considered in Section 4.19 in the context of hard-to-change factors.

## REFERENCES

- Arvidsson, M. and I. Gremyr (2003). Deliberate choices of restrictions in complete randomization. *Quality and Reliability Engineering International*, **19**, 87–99.
- Atkinson, A. C. and R. A. Bailey (2001). One hundred years of the design of experiments on and off the pages of *Biometrika*. *Biometrika*, **88**, 53–97.

- Bailey, R. A. (1987). Restricted randomization: A practical example. *Journal of the American Statistical Association*, **82**, 712–719.
- Bakir, S. T. (1989). Analysis of means using ranks. *Communications in Statistics—Simulation and Computation*, **18**(2), 757–776.
- Barton, R. R. (1997). Pre-experiment planning for designed experiments: Graphical methods. *Journal of Quality Technology*, **29**(3), 307–316.
- Barton, R. R. (1999). *Graphical Methods for the Design of Experiments*. New York: Springer-Verlag.
- Beveridge, W. I. (1960). *Art of Scientific Investigation*. New York: Vantage Press.
- Bisgaard, S. (1989). The quality detective: A case study. *Philosophical Transactions of the Royal Society A327*, 499–511. (This is also available as Report No. 32 of the Center for Productivity and Quality Improvement, University of Wisconsin-Madison and can be downloaded at <http://www.engr.wisc.edu/centers/cppi/reports/pdfs/r032.pdf>.)
- Bisgaard, S. (1992). The early years of designed experiments in industry: Case study references and some historical anecdotes. *Quality Engineering*, **4**(4), 547–562.
- Bisgaard, S. (1994). Blocking generators for small  $2^{k-p}$  designs. *Journal of Quality Technology*, **26**(4), 288–296.
- Bisgaard, S. (1999). Proposals: A mechanism for achieving better experiments. *Quality Engineering*, **11**(4), 645–649.
- Bowman, K. O. and M. A. Kastenbaum (1974). Potential pitfalls of portable power. *Technometrics*, **16**(3), 349–352.
- Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, **71**, 791–799.
- Box, G. E. P. (1990). Must we randomize our experiment? *Quality Engineering*, **2**, 497–502.
- Box, G. E. P. (1993). George's Column. *Quality Engineering*, **5**(2), 321–330.
- Box, G. E. P., S. Bisgaard, and C. Fung (1990). *Designing Industrial Experiments*. Madison, WI: BBBF Books.
- Box, G. E. P. and N. R. Draper (1975). Robust designs. *Biometrika*, **62**, 347–352.
- Box, G. E. P., W. G. Hunter, and J. S. Hunter (1978). *Statistics for Experimenters*. New York: Wiley.
- Brown, M. B. and A. B. Forsythe (1974). Robust tests for equality of variances. *Journal of the American Statistical Association*, **69**, 364–367.
- Buckner, J., B. L. Chin, and J. Henri (1997). Prometrix RS35e gauge study in five two-level factors and one three-level factor. In *Statistical Case Studies for Industrial Process Improvement*, Chapter 2 (V. Czitrom and P. D. Spagon, eds.). Philadelphia: Society for Industrial and Applied Mathematics, and Alexandria, VA: American Statistical Association.
- Campbell, D. T. and J. C. Stanley (2005). *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin Company.
- Clark, J. B. (1999). Response surface modeling for resistance welding. In *Annual Quality Congress Proceedings*, American Society for Quality, Milwaukee, WI.
- Coleman, D. E. and D. C. Montgomery (1993). A systematic approach to planning for a designed industrial experiment. *Technometrics*, **35**(1), 1–12; discussion: 13–27.
- Cook, T. D. and D. T. Campbell (1979). *Quasi-Experimentation: Design and Analysis Issues*. Boston: Houghton Mifflin Company.
- Cox, D. R. (1958). *Planning of Experiments*. New York: Wiley.

REFERENCES

25

- Croarkin, C. and P. Tobias, eds. (2002). *NIST/SEMATECH e-Handbook of Statistical Methods* (<http://www.itl.nist.gov/div898/handbook>), a joint effort of the National Institute of Standards and Technology and International SEMATECH.
- Czitrom, V. (2003). Guidelines for selecting factors and factor levels for an industrial designed experiment. In *Handbook of Statistics, Vol. 22: Statistics in Industry* (R. Khattree and C. R. Rao, eds.). Amsterdam: Elsevier Science B. V.
- Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*. New York: Wiley.
- Dean, A. and D. Voss (1999). *Design and Analysis of Experiments*. New York: Springer-Verlag.
- Ganju, J. and J. M. Lucas (1997). Bias in test statistics when restrictions in randomization are caused by factors. *Communications in Statistics—Theory and Methods*, **26**(1), 47–63.
- Ganju, J. and J. M. Lucas (1999). Detecting randomization restrictions caused by factors. *Journal of Statistical Planning and Inference*, **81**, 129–140.
- Goh, T. N. (2002). The role of statistical design of experiments in Six Sigma: Perspectives of a practitioner. *Quality Engineering*, **14**(4), 659–671.
- Hahn, G. J. (1977). Some things engineers should know about experimental design. *Journal of Quality Technology*, **9**(1), 13–20.
- Hahn, G. J. (1984). Experimental design in the complex world. *Technometrics*, **26**(1), 19–31.
- Inman, J., J. Ledolter, R. V. Lenth, and L. Niemi (1992). Two case studies involving an optical emission spectrometer. *Journal of Quality Technology*, **24**(1), 27–36.
- Jiang, W., B. W. Turnbull, and L. C. Clark (1999). Semiparametric regression models for repeated events with random effects and measurement errors. *Journal of the American Statistical Association*, **94**, 111–124.
- John, P. W. M. (2003). Plenary presentation at the 2003 *Quality and Productivity Research Conference*, IBM T. J. Watson Research Center, Yorktown Heights, NY, May 21–23. The talk is available at [http://www.research.ibm.com/stat/qprc/papers/Peter\\_John.pdf](http://www.research.ibm.com/stat/qprc/papers/Peter_John.pdf).
- Joiner, B. L. and C. Campbell (1976). Designing experiments when run order is important. *Technometrics*, **18**, 249–260.
- Layard, M. W. J. (1973). Robust large-sample tests for homogeneity of variances. *Journal of the American Statistical Association*, **68**(341), 195–198.
- Levene, H. (1960). Robust Tests for the Equality of Variance. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (I. Olkin, et al., eds.). Palo Alto, CA: Stanford University Press, pp. 278–292.
- Lucas, J. M. (1994). How to achieve a robust process using response surface methodology. *Journal of Quality Technology*, **26**(3), 248–260.
- Lynch, R. O. (1993). Minimum detectable effects for  $2^{k-p}$  experimental plans. *Journal of Quality Technology*, **25**(1), 12–17.
- Manly, B. F. (1992). *The Design and Analysis of Research Studies*. Cambridge, UK: Cambridge University Press.
- Mathews, P. (2004). *Designing Experiments with MINITAB*. Milwaukee, WI: Quality Press.
- Montgomery, D. C. (1996). Some practical guidelines for designing an industrial experiment. In *Statistical Applications in Process Control* (J. B. Keats and D. C. Montgomery, eds.). New York: Marcel Dekker.
- Myers, R. H. and D. C. Montgomery (1995). *Response Surface Methodology. Process and Product Optimization using Designed Experiments*, 2nd ed. (2002). New York: Wiley.

- Nelson, P. R., M. Coffin, and K. A. F. Copeland (2003). *Introductory Statistics for Engineering Experimentation*. San Diego, CA: Academic Press.
- Nelson, P. R., P. S. Wludyka, and K. A. F. Copeland (2005). *The Analysis of Means: A Graphical Method for Comparing Means, Rates and Proportions*. Society for Industrial and Applied Mathematics and American Statistical Association: Philadelphia and Alexandria, VA, respectively.
- Nester, M. R. (1996). An applied statistician's creed. *Applied Statistics*, **45**, 401–410.
- Oehlert, G. and P. Whitcomb (2001). Sizing fixed effects for computing power in experimental designs. *Quality and Reliability Engineering International*, **17**(4), 291–306.
- Ott, E. R. (1958). Analysis of means. Technical Report #1. Department of Statistics, Rutgers University.
- Ott, E. R. (1967). Analysis of means—a graphical procedure. *Industrial Quality Control*, **24**(2), 101–109.
- Ott, E. R. (1975). *Process Quality Control*. New York: McGraw-Hill.
- Ott, E. R., E. G. Schilling, and D. V. Neubauer (2005). *Process Quality Control: Troubleshooting and Interpretation of Data*, 4th ed. New York: McGraw-Hill.
- Pearson, E. S. and H. O. Hartley (1972). *Biometrika Tables for Statisticians*, Vol. 2. Cambridge, UK: Cambridge University Press.
- Roach, M. (2003). *Stiff: The Curious Lives of Human Cadavers*. New York: Norton.
- Ryan, T. P. (2000). *Statistical Methods for Quality Improvement*, 2nd ed. New York: Wiley.
- Ryan, T. P. (2004). *Lead Recovery Data. Case Study*. Gaithersburg, MD: Statistical Engineering Division, National Institute of Standards and Technology. (<http://www.itl.nist.gov/div898/casestud/casest3f.pdf>)
- Shadish, W. R. (2001). Quasi-experimental designs. In *International Encyclopedia of the Social and Behavioral Sciences* (N. J. Smelser and P. B. Baltes, eds.). New York: Elsevier, pp. 12655–12659.
- Shadish, W. R., T. D. Cook, and D. T. Campbell (2001). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company.
- Trochim, W. M. K., ed. (1986). *Advances in Quasi-Experimental Design and Analysis*, Vol. 31. Hoboken, NJ: Jossey-Bass.
- Vivacqua, C. A. and A. L. S. de Pinho (2004). On the path to Six Sigma through DOE. In *Annual Quality Transactions*, American Society for Quality, Milwaukee, WI. (available to ASQ members at [http://www.asq.org/members/news/aqc/58\\_2004/20116.pdf](http://www.asq.org/members/news/aqc/58_2004/20116.pdf))
- Van Matre, J. G. and N. Diamond (1996–1997). Team work and design of experiments. *Quality Engineering*, **9**(2), 343–348. (This article is also available as Report No. 144, Center for Quality and Productivity Improvement, University of Wisconsin-Madison and can be downloaded at <http://www.engr.wisc.edu/centers/cqpi/reports/pdfs/r144.pdf>.)
- Wheeler, R. E. (1974). Portable power. *Technometrics*, **16**, 193–201.
- Youden, W. J. (1972). Randomization and experimentation. *Technometrics*, **14**, 13–22.

## EXERCISES

- 1.1** To see the importance of maintaining processes in a state of statistical control when experiments are performed, as stressed in Section 1.7, consider the

following scenario. An industrial experiment is conducted with temperature set at two levels, 300 and 400° F. Assume that no attempt at process control is made during the experiment and consequently there is a 3-sigma increase in the average conductivity from the original value of 18.6, due to some cause other than the change in temperature, which occurs right when the temperature is changed. Assume that  $\sigma$  is known to be 4.5 and that 20 observations were made at each of the two temperature levels.

- (a) Using the assumed value of  $\sigma$ , what is the expected value of the appropriate test statistic for this scenario and the corresponding expected  $p$ -value?
- (b) Answer these same two questions if the process went out of control, by the same amount, when the first 10 observations had been made at the second temperature. What argument do these two sets of numbers support?

- 1.2 Assume that Analysis of Variance calculations have been performed for a problem where there is a single factor with two levels. This would produce results equivalent to an independent sample  $t$ -test provided that the alternative hypothesis for the  $t$ -test is (choose one): (a) greater than, (b) not equal to, (c) less than, (d) none of the above.
- 1.3 It has been stated that “the best time to design an experiment is after it has just been run.” Explain what this means. Is this a problem if we view experimentation as being sequential?
- 1.4 An experiment with a single factor and two levels, 1 and 2, was used and the results were as follows:

1	6.1	8.2	7.3	8.4	8.0	7.6	8.7	9.3	6.8	7.5
2	6.3	8.0	7.7	8.1	8.6	7.2	8.4	9.7	6.8	7.2

Would you use the methodology that was used in Section 1.4.4 in analyzing these data? Why, or why not? If yes, perform the analysis and state your conclusion.

- 1.5 Give an example of an experimental situation in your field in which repeated readings are made instead of replications. Then indicate how the replications would be performed and state whether or not this would have any nonstatistical disadvantages, such as a considerable increase in the cost of running the experiment.
- 1.6 Explain the difference between replication and repeated readings.
- 1.7 Explain why the usual practice of randomly assigning treatments to experimental units is objectionable.

- 1.8 Assume that an experimenter wants to use six levels for an experiment involving only one factor, but since the runs will be somewhat expensive, he can't afford more than 24 observations for the experiment. Explain what problems, if any, this poses.
- 1.9 Identify, if possible, a scenario in your field of application for which the failure to maintain one or more processes in a state of statistical control could seriously undermine experimentation of a specific type.
- 1.10 Critique the following statement regarding the data in Exercise 1.4: "Since the data vary considerably within each level but there is much less variation between corresponding values for the two levels, if the data are time-ordered, then there must be at least one statistical process that is out of control."
- 1.11 Assume a single factor with two levels. Use either appropriate software or the Java applet in Section 1.4.4 to determine the power of detecting a difference of the two means of at least 2.5 when  $\sigma_1 = \sigma_2 = \sigma = 4$ ,  $\sigma = .05$ , and there are 10 observations made at each level of the factor. Will the power be greater or less if there are more observations and  $\sigma$  is smaller, or can that be determined? Explain.
- 1.12 Explain why an experimenter would consider using factors with more than two levels.
- 1.13 Assume that the time order of the observations from an experiment is unknown, and it is suspected that an important process may have gone out of control during experimentation with a single factor but the run sequence was lost. Could the data still be analyzed? What does this imply about the need to keep processes in a state of statistical control and to note the run sequence?
- 1.14 Critique the following statement: "I believe that the expected change in my response variable is approximately 2 units when a particular factor is varied from one level to another level. Thus, I know what should happen so there is no point in performing any experimentation."
- 1.15 Assume that an experiment with four levels of a single factor was run, with randomization, and the value of the response variable was almost strictly increasing during the runs in the experiment. What would you suspect and what would be your recommendation?
- 1.16 If you have a solid statistical background, read the Jiang, Turnbull, and Clark (1999) paper that was mentioned in Section 1.7. Also read case study #2 in Inman et al. (1992), which you will likely find easier to read than the other paper. Write up the use of process control methods in each paper. If you are

EXERCISES

conversant in the latter, do you believe that the methods that they used to try to ensure control were adequate? Explain.

- 1.17 Use available and appropriate software to determine the probability of detecting a difference in the means for two levels of a factor of  $1.5\sigma$  when  $n_1 = n_2 = 20$ . Would the probability increase or decrease if the two sample sizes were increased? Explain.
- 1.18 A discussion of experimental design considerations for tree improvement is given in *Experimental Design and Tree Improvement*, 2nd ed., by E. R. Williams, A. C. Matheson, and C. E. Harwood. The second chapter of the book is available online at <http://www.publish.csiro.au/samples/ExperimentalDesignSample.pdf>. Read the chapter and comment on the nonstatistical design considerations that are involved in seeking tree improvement through experimental design.
- 1.19 Is sequential experimentation feasible in your field of study? If not, explain why not. If it is possible, can a sequence of experiments be performed in a short or at least reasonable period of time? Explain.
- 1.20 If possible, give an example of an experiment from your field where complete randomization and the design of a one-shot experiment could be risky.
- 1.21 Buckner, Chin, and Henri (1997, references) performed a gauge study using a design with three factors, but we will analyze only one of the three: the operator effect. The data for the sheet resistance for the 7 kÅ tungsten film under test is shown below, along with the corresponding operator, which is labeled here simply as 1, 2, and 3.

Sheet resistance (mΩ)	84.86	84.92	84.81	84.80	84.86	84.93	84.80	84.94
Operator	1	2	2	3	3	2	2	1
Sheet resistance (mΩ)	84.91	84.86	84.78	84.86	84.96	84.89	84.90	84.90
Operator	1	2	3	2	2	2	3	1

Notice that the design relative to this factor is unbalanced. Does that create a problem with the analysis of the data? Why, or why not? If the data can be analyzed despite this imbalance, do so and determine whether or not there was a significant operator effect.

- 1.22 The following data appear on the Internet, preceded by the statement “In a comparison of the finger-tapping speed of males and females the following data was [*sic*] collected.”

Males	43	56	32	45	36	48			
Females	41	63	72	53	68	49	51	59	60

This is all the information that was given relative to the experiment, if it can be called such, and the data.

- (a) Based on the information given, could this be called an experiment. If so, what is the factor? If not, explain why it isn't an experiment.
- (b) Can the data be analyzed, based on what is given? Why, or why not? If the data can be analyzed, do so and draw a conclusion.

**1.23** The example given in Ryan (2000), which was mentioned in Section 1.7, showed what can happen to the conclusions of an experiment when a process goes out of control and affects the mean of the experimental level of a factor. Would the experiment similarly be undermined if the variance of either or both of the levels was increased as a result of an out-of-control condition? Explain.