CHAPTER 4

# Full Factorial Designs with Two Levels

In this chapter we extend the discussion of designed experiments to experiments involving more than one factor, but restrict consideration to designs with two levels. The most frequently used design with multiple factors is a *factorial design.*

## 4.1 THE NATURE OF FACTORIAL DESIGNS

A factorial design has no direct connections with factorials per se (such as 4!). All possible combinations of factor levels are used (ideally) in a factorial design, so one could argue that "combination designs" would be a better term. The term "crossed design" is also used because each level of each factor is "crossed" with the levels of all the other factors.

To illustrate, the simplest example of a factorial design would be a $2^2$—a design with two factors, each at two levels. This design would hardly ever be used in practice, but it has considerable illustrative value due to its simplicity. For the rest of this chapter, we will assume, unless stated otherwise, that each factor is *fixed*, that is, the levels used in the experiment are the only ones of interest. (Recall the discussion in Section 2.1.1 regarding fixed and random factors.)

In general, factorial designs are represented by the form $s^k$, with $s$ denoting the number of levels and $k$ denoting the number of factors. When $s = 2$, the two levels are usually denoted as "high" and "low" if the factor is quantitative, such as temperature. Of course if the factor is qualitative, with the two "levels" being two operators involved in a production process, such a designation would have no meaning.

For a $2^2$ design there are obviously four combinations of the high and low levels of the two factors and these combinations are often designated by using the presence and absence of lowercase letters to indicate the high level and low level, respectively, of the corresponding factor, which is designated by the appropriate capital letter. For example, $a$ would mean that factor $A$ is at the high level and factor $B$ is at the low level. This takes care of combinations $a$, $b$, and $ab$, but if the combination with each

*Modern Experimental Design* By Thomas P. Ryan

**TABLE 4.1 Design Layout for the $2^2$ Design**

| Treatment Combination | $A$ | $B$ | $AB$ |
|---|---|---|---|
| **(1)** | $-1$ | $-1$ | 1 |
| $a$ | 1 | $-1$ | $-1$ |
| $b$ | $-1$ | 1 | $-1$ |
| $ab$ | 1 | 1 | 1 |

factor at its low level was similarly designated, we would just have a blank space. Since that obviously wouldn't work, we use (1) to indicate each factor at its low level and this is also used when there are more than two factors.

Using a $2^2$ design and estimating all the effects that can be estimated ($A$, $B$, and $AB$, the interaction term to be explained later) implies a tentative Analysis of Variance (ANOVA) model of the form

$$Y_{ij} = \mu + A_i + B_j + (AB)_{ij} \quad i = 1, 2 \;\; j = 1, 2 \tag{4.1}$$

with $Y_{ij}$ denoting the response when the $i$th level of $A$ is used in combination with the $j$th level of $B$, and "1" and "2" denoting the low level and the high level, respectively, of each factor. Note that there is no error term in this model, as there was, for example, in Eq. (2.1). This is because there is an exact fit when this design is used and the design is not replicated. This is illustrated later in the section. There is an error term when the design is replicated, so then the model is

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijk} \quad i = 1, 2 \;\; j = 1, 2 \;\; k = 1, 2, \ldots, c$$

for $c$ replicates.

For the moment, we can view $(AB)_{ij}$ as a product term, analogous to a product term in a regression model, because this is how the "levels" of the term are formed. Specifically, the design layout, including the values of the product $AB$, which of course are forced rather than specifically selected, is as given in Table 4.1.

The high level of each factor is denoted by $+1$ and the low level by $-1$. The levels could have been denoted by 0 and 1, or by 1 and 2, and these designations are used in certain literature articles, but the $(+1, -1)$ designation has some advantages that will become evident shortly, and it is also a natural designation. Regarding the latter, recall that the numerator of a two-sample $t$-test, as discussed in Section 1.4.3 is $\overline{y}_1 - \overline{y}_2$, with the coefficients of the two sample averages thus being $+1$ and $-1$. In general, whenever we compare data from two populations or compare the effects of two levels of a factor, the obvious choice for the coefficients is $+1$ and $-1$.

Notice that the dot product of any pair of these columns is zero. (For example, the dot product of $A$ and $B$ is $(-1)(-1) + (1)(-1) + (-1)(1) + (1)(1) = 0$.) This implies that the design is *orthogonal*, which means that the effects of each factor and the product of the two factors can be estimated independently of each other. The $A$ and $B$ effects are referred to as *main effects*, and the product term will, in the next section, be explained as an *interaction effect*. Note that two levels result when the multiplication is carried out to produce the product term.

The treatment combinations in Table 4.1 are listed in *Yates order*. If there had been three factors, for example, the rest of the sequence would have been *c, ac, bc, abc*. That is, whenever a new letter is listed, it is followed by all possible combinations of it with the previously listed letters. So if there had been a fourth factor, the sequence would have continued with *d, ad, bd, abd, cd, acd, bcd, abcd*, with the entire list having $2^4 = 16$ treatment combinations.

The numbers in the columns in Table 4.1 indicate how the effect of each factor would be estimated. Even if we didn't have these numbers to look at, a logical way to estimate, say, factor $A$ would be to take the average of the observations that result when $A$ is at one of the two levels and subtract from this number the average of the two responses when $A$ is at the other level.

**Example 4.1**

To illustrate, let's assume that the observations of some physical characteristic (tensile strength, perhaps) that result when the experiment is run (with $A$ and $B$ perhaps denoting temperature and pressure, respectively) are 70, 62, 59, and 71, corresponding to the order of the treatment combinations listed in the table.

Assume that the temperature levels are 350 and 370°F. Then we would code these levels as $(370 - 360)/10 = 1$ and $(350 - 360)/10 = -1$, and similarly for the other factor. It is important that analyses be performed using coded values when interaction effects are being estimated, because misleading results can be obtained when analyses are performed using raw values, and these results will not be in agreement with the results obtained using the coded values. This can be illustrated as follows.

As long as we have a single factor, the analysis, in terms of the conclusions that are reached, is invariant to whatever coding, if any, is used. This is also true when there is more than one factor, provided that the model does not contain any interaction terms. To see this, assume that we have two factors, $X_1$ and $X_2$, and an unreplicated $2^2$ design. In coded form, $X_1$, $X_2$, and $X_1X_2$ are of course pairwise orthogonal. That is, the dot product of the following columns

| *A* | *B* | *AB* |
|---|---|---|
| -1 | -1 | 1 |
| 1 | -1 | -1 |
| -1 | 1 | -1 |
| 1 | 1 | 1 |

is zero, as stated previously. Assume that the two levels of factor $A$ are 10 and 15 and the two levels of factor $B$ are 20 and 30. The raw-form representation of the two factors and their interaction is then

| *A* | *B* | *AB* |
|---|---|---|
| 10 | 20 | 200 |
| 20 | 20 | 400 |
| 10 | 30 | 300 |
| 20 | 30 | 600 |

Notice that the last column of numbers cannot be transformed to the corresponding set of coded-form numbers with any method of coding (i.e., any linear transformation). Furthermore, if the analysis were performed in raw form there would be a major multicollinearity problem as the correlation between $A$ and $AB$ is .845 and the correlation between $B$ and $AB$ is .507. If we let the four response values be 6.0, 5.8, 9.7, and 6.3, respectively, it can be shown that the ordinary least squares (OLS) coefficient of $A$ is positive when the raw form is used and negative when the coded form is used.

We can see that the $A$ effect is negative because $6.0 + 9.7$ exceeds $5.8 + 6.3$, using the signs in the $A$ column, which is used to obtain the $A$ effect estimate, and similarly for the other effect estimates. Specifically, the effect estimate is $(5.8 + 6.3)/2 - (6.0 + 9.7)/2 = -1.8$. Therefore, since the effect estimate is negative, the regression coefficient should also be negative. The sign is (correctly) negative when the coded form is used but not when the raw form is used. Thus we have a wrong sign problem, and unlike the "wrong sign" discussed in the regression literature (see, e.g., Ryan, 1997, p. 131), there really *is* a wrong sign problem!

Returning to the temperature data, since $A$ is a quantitative factor, the estimate of the $A$ effect would logically be computed as the average response value at the higher of the two temperatures minus the average response at the lower temperature, and notice that this is what happens when we multiply the column of numbers for $A$ by the corresponding response values.

Of course "higher" is denoted by "1" and "lower" by "$-1$" in terms of the coded levels. The $B$ effect would be computed in the same manner as the $A$ effect. Note that two levels result when $A$ and $B$ are multiplied together to produce the interaction $AB$. Accordingly, it seems reasonable to compute the interaction effect as the average response at the "1 level" minus the average response at the "$-1$ level," and this is how the interaction effect is computed if we multiply the $AB$ column by the corresponding response values.

Thus, the estimate of the $A$ effect is $(62 + 71)/2 - (70 + 59)/2 = 2$. Similarly, the estimate of the $B$ effect is $(59 + 71)/2 - (70 + 62)/2 = -1$, and the estimate of the $AB$ effect is $(70 + 71)/2 - (62 + 59)/2 = 10$.

If we write the fitted model in the form of a regression model, with $b_0$, $b_1$, $b_2$, and $b_{12}$ denoting the OLS regression coefficients, we then have

$$\widehat{Y} = b_0 + b_1 A + b_2 B + b_{12} AB \tag{4.2}$$

with $b_0 = \widehat{\mu} = \overline{Y} = 65.5$ and $b_1$, $b_2$, and $b_{12}$ are, by definition, half the effect estimates for the terms for which they serve as coefficients. (See chapter Appendix B for details.) That is, the $A$ effect estimate was calculated to be 2, so $b_1 = 1$. The effect estimates of $B$ and $AB$ are $-1$ and 10, respectively, so $b_2 = -0.5$ and $b_{12} = 5$.

Notice that the fitted values are equal to the observed values. For example, the first fitted value is $\hat{Y} = 65.5 + (1)(-1) + (-0.5)(-1) + (5)(1) = 70$, and it can be similarly shown that the other fitted values are the same as the observed values.

Of course this does not generally happen when a regression model is fit to a set of data, but in such applications of regression the number of observations is

invariably greater than the number of parameters that are estimated. Here the number of coefficients is equal to the number of observations, so there is an exact fit.

Although it might seem like utopia to be able to exactly reproduce the observed values with a fitted model, this is actually a problem because it stands to reason that there would not be an exact fit if there were more observations. For example, if there were two observations per treatment combination, there would not be an exact fit unless the two observations at each treatment combination were the same, which would be highly unlikely and would suggest that there was a problem with the experiment.

An unreplicated factorial design, be it a $2^2$ design or a design with more factors, is a *saturated design*. That is, the number of effects to be estimated is equal to the number of observations minus one. That this will be the case for all two-level full factorial designs regardless of the number of factors can be seen as follows. The number of observations for a $2^k$ design will of course be $2^k$ since this defines the number of observations. Now think of the "2" as representing the two possibilities, presence or absence, of each factor in an effect to be estimated. If all factors are "absent," then there is nothing to estimate, so the total number of effects is $2^k - 1 = n - 1$.

Hypothesis tests, as were performed in Section 2.1.2, for example, by using information from an ANOVA table, are not available for unreplicated designs because the variance of the error term cannot be estimated because there isn't enough data. Instead, a normal probability plot approach is typically used, which is illustrated and discussed in Sections 4.10 and 4.11 and is also used in subsequent chapters.

**Example 4.2**

Recall the example with the runner, which was given in Section 1.3. Assume that the runner conducts an experiment over a period of several months with the two levels of the training factor being "heavy" and "moderate" and the two levels of nutritional supplementation being "moderate" and "increased." The runner's time in minutes is recorded over a prescribed course for each of the four combinations. Those times are 57, 52, 53, and 58 for the combinations (1), *a, b,* and *ab,* respectively, with factor $A$ being the training factor and factor $B$ the nutritional factor. It is easy to see that the estimate of the $A$ effect is 0, and the estimate of the $B$ effect is 1.0. The estimate of the $AB$ interaction effect is 5.0, which complicates the analysis since it is much bigger than the $A$ effect and $B$ effect. Clearly there is an $A$ effect at low $B(52 - 57)$ and an $A$ effect at high $B(58 - 53)$—the effects simply add to zero. This necessitates the use of conditional effects, which are discussed in Section 4.2.1.

Although the $2^2$ design is the ideal design to use for illustrating basic concepts in factorial designs, we can use Eq. (1.3), appropriately modified for a two-tailed test, to show that it really has no practical value. For an unreplicated $2^2$ design, we obtain $\Delta \doteq \frac{((1.96+1.28)2\sigma)}{\sqrt{n}} = \frac{6.48\sigma}{\sqrt{n}} = 3.24\sigma$ as the smallest effect of $A$, $B$, or $AB$ that could be detected, if $\sigma$ were known, with a probability of .90 and a significance level of $\alpha = .05$. This is far too large a multiple of $\sigma$. The multiplier would have to be much smaller than this for a design to have practical value. We apply this formula to larger factorial designs in later sections.

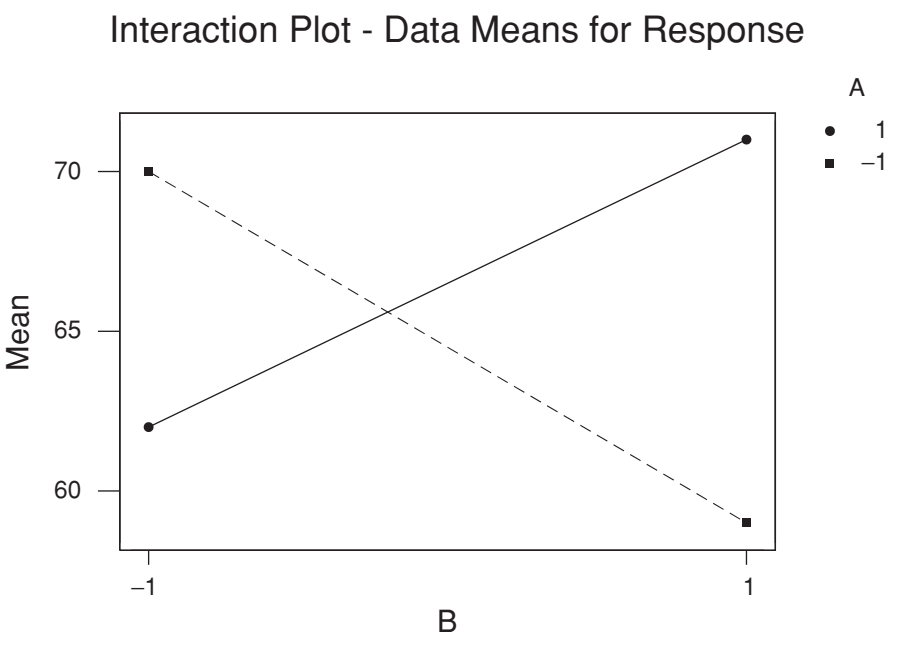



**Figure 4.1** Interaction plot for $2^2$ example.

So while recognizing that the design has no practical value, we will continue to use it for another section or two to illustrate basic concepts.

## 4.2 THE DELETERIOUS EFFECTS OF INTERACTIONS

In the previous section, *AB* was viewed as a product term, and it could be seen from Table 4.1 that it is indeed a product term. In this section we will view it, equivalently, as an *interaction term*, with the term "interaction" unrelated to the interaction of physical variables as in a chemical interaction. Interaction simply means that the effect of a factor depends upon the level(s) of the other factor(s). This will be illustrated later in the section.

The presence of interaction, particularly extreme interaction, can easily result in completely erroneous conclusions being drawn if an experimenter is not careful. This can be seen in the example in the preceding section, as the interaction effect estimate was much larger than either of the main effect estimates. Daniel (1976, p. 21) stresses that data from a designed experiment should not be reported in terms of main effects and interactions if an interaction is more than one-third of a main effect.

In the example in Section 4.1, the main effect estimates are not even 1/3 of the interaction effect estimate! This is a serious interaction problem, which can also be seen graphically. The four points are plotted in Figure 4.1.

As was indicated by the numerical calculations, this plot illustrates severe interaction, with the interaction effect estimate being nonzero if the lines are not parallel. Here the plot deviates sharply from parallelism, almost forming an "X." If an "X" had been formed, the main effect estimates would have been zero and an analysis using main effects would have been totally misleading. In particular, note that there is a very large $A$ effect at the high value of $B$, as represented by the vertical distance between the two lines at $B = 1$. Similarly, there is a very large $B$ effect at the low level of $A$, as reflected by the slope of the dotted line.

### 4.2.1 Conditional Effects

These effects have been called *simple* effects in the literature, as in Glasnapp and Sauls (1976), Bohrer, Chow, Faith, Joshi, and Wu (1981), Woodward and Bonett (1991), Winer, Brown, and Michels (1991), Bonett and Woodward (1993), Kirk (1995), Schabenberger, Gregoire, and Kong (2000), Kuehl (2000), and Hinklemann and Kempthorne (2005, p. 260), and Schabenberger et al. (2000) referred to collections of simple effects as "slices." A more appropriate term, however, probably is *conditional effects* because effect estimates are computed by conditioning on a particular level of one of the factors, which for a two-level factor means that each conditional effect is computed using half of the data, with the computation formed as for a regular (i.e., unconditional) effect estimate, except that only half of the data are being used. The term "conditional effect" has also been used by Kao, Notz, and Dean (1997) and Wu and Hamada (2000). See also Taguchi (1987, p. 279), in which it is termed the "trans-factor" technique.

It is worth noting that the data are also split when CART (Classification and Regression Trees) methodology is used to analyze data from designed experiments, but there the objectives are different, as one objective might be to determine the combination of factor levels that maximizes the response. Another objective is to uncover interactions that might not be easily detectable when small designs are used. The use of CART in analyzing data from factorial designs is illustrated by Wisnowski, Runger, and Montgomery (1999–2000).

Consider the following example, which illustrates the need for examining conditional effects.

**Example 4.3**

As a very simple example, consider an unreplicated $2^2$ design and the following data.

| $A$ | $B$ | Response |
|---|---|---|
| -1 | -1 | 14 |
| 1 | -1 | 14 |
| -1 | 1 | 13 |
| 1 | 1 | 22 |

The interaction effect estimate is $1/2(14 + 22 - 14 - 13) = 4.5$.

Splitting on $B$, the conditional effects of $A$ are 0 at low $B$ and 9.0 at high $B$.
Splitting on $A$, the conditional effects of $B$ are $-1.0$ at low $A$ and 8.0 at high $A$.

The conditional effect estimates are just linear combinations of two numbers for this very small example, but the point to be made here is that none of the conditional effects well-represent the corresponding main effects, which are $A = 4.5$ and $B = 3.5$, and this is due to the large interaction. This could be a major problem if, several months after an analysis was performed, it became desirable or necessary to fix one of the factors at a particular level, with another factor being hard to maintain at a desired level due to physical considerations, so that the factor level is apt to vary within a particular range. If a conditional effects analysis had not been performed, an engineer might have a hard time determining the expected effect on the value of the response variable as the hard-to-maintain factor is varied within its range. This problem might be avoided if all significant interactions are included in the model and the model is used to try to determine the effect of a particular scenario on the response. Interactions can have a disruptive effect on analyses when they are relatively large but not large enough to be declared significant and included in a model.

In this example the interaction effect was positive, which means that the conditional effect at the high level of each factor is greater than the corresponding conditional effect at the low level of each factor. Conversely, if the interaction effect estimate is negative, the conditional effect at the high level of each factor must be less than the conditional effect at the low level of each factor. This result follows from the derivation in chapter Appendix A, since the interaction effect is added to the main effect to obtain the conditional effect at the high level of the other factor and subtracted from the main effect to obtain the conditional effect at the low level of the other factor.

To illustrate the conditional effect at the high level of each factor being less than the conditional effect at the low level, we can simply exchange the last two response values in the example. We then have the following.

The interaction effect estimate is $1/2(14 + 13 - 14 - 22) = -4.5$.

Splitting on $B$, the conditional effects of $A$ are 0 at low $B$ and $-9.0$ at high $B$.
Splitting on $A$, the conditional effects of $B$ are 8.0 at low $A$ and $-1.0$ at high $A$.

Again, none of the conditional effects are representative of the corresponding main effects because of the size of the interaction relative to the main effects, which here are $-4.5$ for $A$ and $B$ is still 3.5.

It appears as though simple/conditional effects and the need for their use in analyzing data from designed experiments has been largely ignored in the applied literature, however. Exceptions include Toews, Lockyer, Dobson, Simpson, Brownell, Brenneis, MacPherson, and Cohen (1997), Stehman and Meredith (1995), and Peeler (1995).

Whatever descriptive label is used, the concept should be applied only in the case of fixed factors, since for random factors variance components are estimated rather than effects.

These conditional effects are easy to see and compute for a $2^2$ design, but it becomes much more involved when there are more than two factors. This is illustrated in Sections 4.6 and 5.1.4, for example.

**IMPORTANT POINT**

Conditional effects should be used routinely because at least some interaction effect estimates are likely to at least moderately differ from zero in practically any application, and even moderate departures from zero will cause the corresponding conditional effects to differ considerably. As with any average, we would like to have conditional effects computed from a reasonable number of observations. If not, and this could happen with small designs, the variance of the conditional effects estimates could be high and possibly be misleading.

The failure to recognize this has resulted in some bad advice in the literature. For example, Emanuel and Palanisamy (2000) discuss sequential experimentation but state "... two-factor interactions containing at least one insignificant main effect are negligible." The same message can be found in Chipman (1996), who referred to the "strong heredity assumption," which holds that an interaction is likely to be significant only if both main effects are significant (see Wu and Hamada, 2000, p. 365 for additional discussion). Many researchers adopt an opposing position, including Bingham (2001) who argues against the strong heredity assumption. The use of the strong heredity assumption overlooks the fact that large two-factor interactions can *cause* insignificant main effects, with both the interactions and the conditional effects being real. Similarly, Box, Meyer, and Steinberg (1996) and Box and Tyssedal (2001) have contended that the "weak heredity assumption," which requires that an interaction cannot be present in a model unless at least one of the factors in the interaction is present, is not valid in every experimental situation.

One obvious question to ask is the following: Should these interaction plots be constructed by having the factor that comes second in alphabetical order plotted on the horizontal axis, or does this matter? It does matter because we tend to associate the magnitude of an interaction with the extent to which the lines deviate from parallelism, and visually we see stronger evidence of extreme interaction when the lines cross than when they don't cross. Whether or not they cross, however, will often depend on which factor is plotted on the horizontal axis, as the reader is asked to show for the data given in Exercise 4.3. Because of this, it is a good idea to construct interaction plots both ways. Of course this is easy to do with a small number of factors—such as two, as we have here—but becomes more of a chore for a moderate-to-large number of factors. (For the data in Figure 4.1, the lines will cross regardless of which factor is plotted on the horizontal axis.)

Fortunately, there is an efficient way to do this with software. For example, the FFINT command in MINITAB can be used to construct a matrix of interaction plots for two-level designs. To reverse the axis on each plot in the scatterplot, the order in

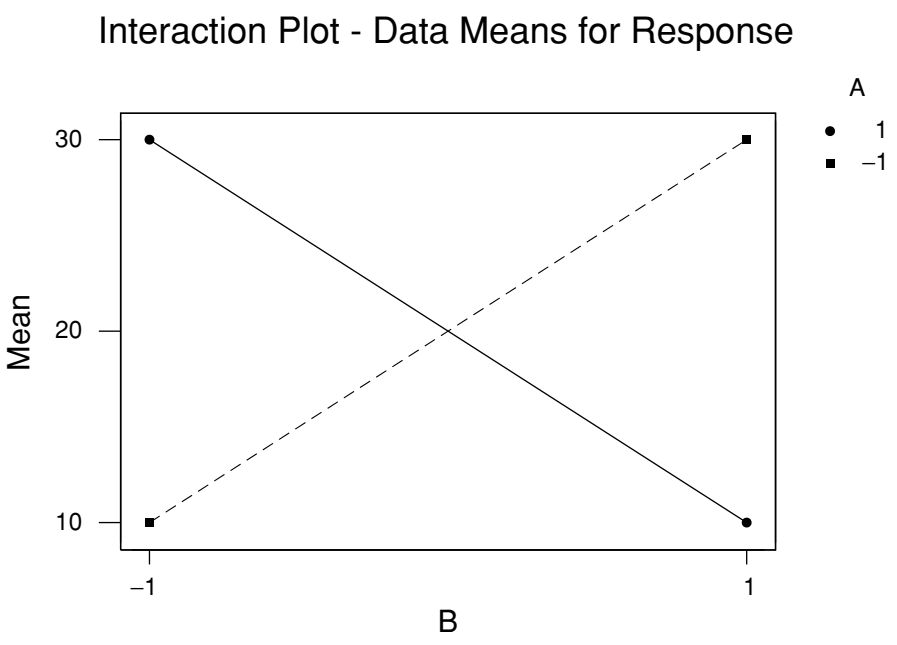



**Figure 4.2** The most extreme interaction possible.

which the factors are listed in the command are simply reversed. For example, FFINT C3 C2 C1 in command mode will give the reversal of the interaction plots produced by FFINT C1 C2 C3.

As a slightly more extreme example, using the "X" configuration mentioned previously, consider Figure 4.2.

It is clear that the value of the response variable varies by 20 units when either $A$ or $B$ is set at one of the two levels and the other factor is varied between its two levels. Yet, when the data are analyzed we would find that the main effect estimates for each of the two factors are exactly zero. This is because the conditional effects differ only in sign and thus add to zero. When this occurs, the main effect estimate will be zero, as the reader is asked to show in Exercise 4.1.

This result shows that a "blind" analysis, such as unthinkingly relying on computer output, would lead an experimenter to conclude that there is neither an $A$ effect nor a $B$ effect, although each clearly has an effect on the response variable when viewed in the proper perspective. It is important that interaction plots such as those in Figures 4.1 and 4.2 be used in addition to other graphical displays in analyzing data from multifactor designs.

As in the graphical analysis of the data in Figure 4.1, we might ask if the configuration of points and the strength of the signal of the interaction depend on which factor is plotted on the horizontal axis. In order to have a perfect "X," there can be only two distinct values of the response variable. The "X" means that both main effect estimates are zero, and the only way this can happen when the other factor is plotted

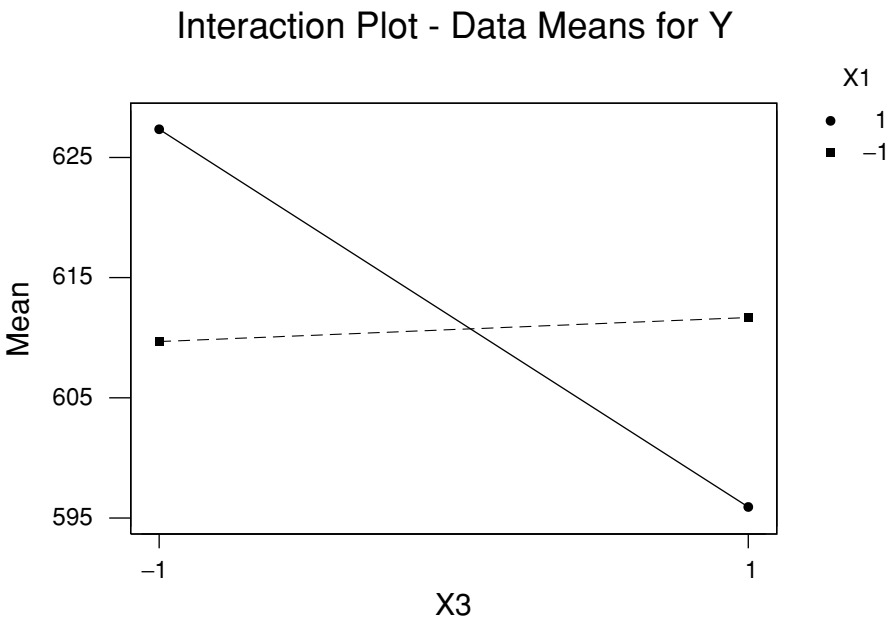



**Figure 4.3** Interaction plot of data from the NIST ceramics experiment.

on the horizontal axis is when the other graph is the same "X," which must occur since there are only two distinct values of the response variable. The practical significance of this is that if we have anything approaching such a configuration of points, then we shouldn't bother to construct the other graph.

In general, however, it is a good idea to construct both graphs, especially if conditional effects are being used.

Although the configuration in Figure 4.2 is an extreme example, and would seem not likely to occur exactly in practice, we should not be surprised to find significant interactions and nonsignificant main effects. We may, however, observe interaction plots that approach the configuration in Figure 4.2, and Figure 4.3 is an interaction plot from an actual experiment. Furthermore, Box (1999–2000) described an actual experiment such that "all the variables were clearly active . . . but our main effects were all essentially zero." So a scenario that is essentially the same as Figure 4.2 *has* occurred in practice and has perhaps occurred in a higher fraction of experiments than we might suspect.

The data graphed in Figure 4.3 are given in the case study in Section 1.4.2.10.1 of the *NIST/SEMATECH e-Handbook of Statistical Methods* (http://www.itl.nist.gov/div898/handbook/eda/section4/eda42al.htm). The data used in that case study were from a larger set of data collected by Said Jahanmir of the NIST Ceramics Division in 1996 for a NIST/industry ceramics consortium. Three factors were used in the case study, from the 12 factors in the study of Jahanmir.

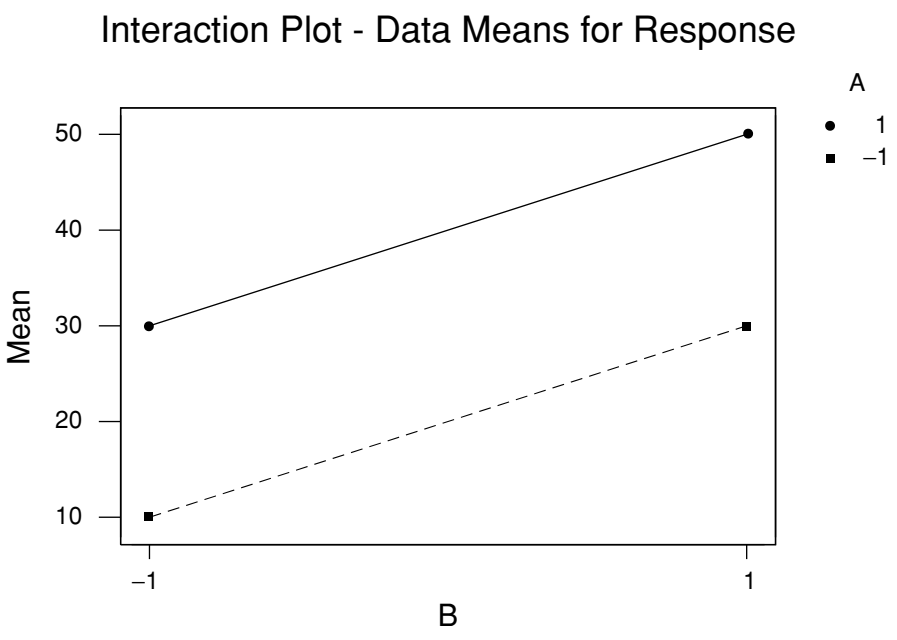



**Figure 4.4** Interaction plot illustrating no interaction.

Two batches were used and the overall objective was to optimize the strength of a ceramic material. The factors denoted as X1 and X3 were table speed and wheel grit, respectively. The former was not identified in the case study as an important factor in the analysis of the data for batch 2, and we can see from Figure 4.3 that the effect estimate must be close to zero, relative to the magnitude of the numbers. (It is actually 0.949.)

If the dotted line in Figure 4.3 had been horizontal and the solid line had crossed that line halfway between $-1$ and 1 for X3, the estimate of the main effect for X1 would have been zero, as the two sides of the graph would have been mirror images. What occurred is very close to this, which illustrates that an extreme interaction plot configuration can occur in a case study of data from an experiment.

Since the X1X3 interaction effect estimate was $-16.711$, the conditional effects for X1 are $0.949 \pm (-16.711) = -15.762$ and 17.660. The conclusion from the case study was that the main effect of X1 was essentially zero, but X1 clearly has an effect, as is indicated by the magnitude of the conditional effects, and the different signs should also be noted.

Another type of extreme interaction plot is given in Figure 4.4. This illustrates no interaction and we would like to see something close to this as then the main effect estimates can be interpreted unambiguously.

To this point in our discussion of the $2^2$ design the main effect estimates have been defined in terms of the average response value at one level of a factor minus the

average response value at the other level of the factor, and the way that the interaction effect estimate is obtained was illustrated in Table 4.1.

**Example 4.4**

Sztendur and Diamond (2002) gave an example in which the coefficients of conditional effects were estimated using a regression approach applied to data from an experiment given in Hsieh and Goodwin (1986).

We will work up to what they did by building on the discussion of conditional effects in this chapter and by using results in the chapter Appendix A.

There is a way of expressing conditional effects in terms of the main effect and interaction effect that is derived in chapter Appendix A and which was used earlier in this chapter. Applying that result, let $A^+$ denote the effect of factor $A$ at the high level of factor $B$, while $A^-$ denotes the effect of factor $A$ at the low level of factor $B$, with $A$ denoting the main effect and $AB$ denoting the interaction effect. Then $A^+ = A + AB$ and $A^- = A - AB$. In chapter Appendix B, the derivation is given for the general result: regression coefficient $= \frac{1}{2}$ (effect estimate).

This suggests that the "regression variables" to be used for the conditional effects of $A$ should be $\frac{(A+AB)}{2}$ and $\frac{(A-AB)}{2}$. If we apply this to the $2^2$ example in Example 4.1 and list the response values, we obtain the following if we elect to use the three degrees of freedom for estimating the conditional effects of $A$ and the $B$ effect.

| $A^+$ | $A^-$ | $B$ | $Y$ |
|---|---|---|---|
| 0 | -1 | -1 | 70 |
| 0 | 1 | -1 | 62 |
| -1 | 0 | 1 | 59 |
| 1 | 0 | 1 | 71 |

If we fit a regression model to this, we obtain $\widehat{Y} = 65.5 + 6A^+ - 4A^- - 0.5B$, which tells us indirectly that the conditional effects of $A$ are 12 and $-8$, respectively, and the $B$ effect is $-1$. Of course we can see these results just by looking at the columns above.

Sztendur and Diamond (2002) used a fitted model that contained a conditional effect estimate (of the effect of factor $B$ at the high level of factor $G$). No additional computing is necessary if $B$ and $BG$ have already been fit in a model since the conditional effect coefficient is simply the sum of the coefficients of $B$ and $BG$ because the conditional effect of $B$ at high $G$ is simply $B + BG$. Since this was done in the context of determining a path of steepest ascent in a response surface approach, we will return to this example in Section 10.10.

#### *4.2.1.1 Sample Sizes for Conditional Effects Estimation*

The size of an effect that can be detected (with high probability) was discussed in Section 1.4.4. Some consideration should be given to the number of observations that each conditional effect estimate will be computed from when $2^k$ designs are used.

Of course this number will be $2^{k-1}$ when an unreplicated design is used and $r2^{k-1}$ when $r$ replicates are used. For an unreplicated design, $2^{k-1}$ will of course be small when $k$ is small and may not be large enough to obtain conditional effect estimates with an acceptably small variance. This is something that should be kept in mind if interactions are anticipated or at least thought to be quite possible, so that a replicated design might be chosen if extra experimental runs are not expensive or otherwise impractical.

### 4.2.2 Can We "Transform Away" Interactions?

Since interactions will usually necessitate a conditional effects analysis and thus complicate the analysis, we might ask whether we should try to transform the data in an attempt to remove an interaction. Sometimes it is possible to transform the data so as to essentially remove an interaction by making the interaction on the transformed scale quite small. However, if the interaction is large on the original scale, it will not be small on the transformed scale.

To illustrate, assume that a two-factor interaction is as shown in Figure 4.1. Common transformations such as square root and logarithmic transformations will preserve the ranking of the numbers. To illustrate, assume that the numbers represented by Figure 4.1 are 70, 62, 58, and 72. The only way the lines in the interaction profile for the transformed numbers would not cross would be if the ranking of the first two numbers is the same as the ranking of the last two numbers when the data in each pair are ranked separately. That is, the ranking of the untransformed numbers is 1, 2, 2, 1. If we use a logarithmic transformation or a square root transformation, for example, the rankings will be maintained and the lines will still cross. If we use a reciprocal transformation, however, we "transform" the rankings so that they become 2, 1, 1, 2. The lines will still cross since we have "flipped" the rankings and the ranking of the pairs is still different.

Similarly, if the interaction is extreme, but not so extreme that the lines cross, transforming the data will not remove the interaction. This is not to suggest that one shouldn't consider removing interactions through transformations, as the analysis of no-interaction models is much simpler than the analysis of models with interactions. Not only is a conditional effects analysis obviated, but hypothesis testing is greatly simplified since interaction terms are the denominators in $F$-tests for experiments with random factors as well as experiments with both fixed and random factors.

## 4.3 EFFECT ESTIMATES

In this section we take a more formal look at effect estimates, complementing what was done in Section 4.1. There it was stated that the effect of a two-level factor would logically be estimated as the average response at the high level of the factor minus the average response at the low level.

Another (mathematically equivalent) way of obtaining a main effect estimate is as follows. Consider Figure 4.4. It would be logical to estimate the $A$ effect by taking

the average of the $A$ effect at the high level of $B$ and the $A$ effect at the low level of $B$. Since the lines are parallel, we have the same number for each, namely, 20, which is the vertical distance between the lines. (The reader is asked to show in Exercise 4.2 that this is mathematically equivalent to what was used in Section 4.1.)

It is comforting that the two numbers are the same because if they differ greatly, they will each differ from the average of the numbers that is used to represent the effect.

If we similarly view the estimate of the $B$ effect in the same way, we see that the estimate of the $B$ effect is the average of the "slopes" of the lines, and that each number is 20. (This must be the case if the numbers that we average to obtain the $A$ effect estimate are the same.)

An obvious way to measure the interaction effect is to use the *difference* between the two vertical distances in Figure 4.4. The reader may wish to verify that this is equivalent to using the coefficients in Table 4.1. Thus, the set of numbers in the $AB$ column has a physical interpretation.

Whereas the scenario in Figure 4.4 is ideal from the standpoint of assessing the effect of the factors $A$ and $B$, the scenarios depicted in Figure 4.2 and 4.3 present major problems. In particular, it is imperative that an analysis using conditional effects be performed for the data shown in Figure 4.2 since both main effect estimates are zero. Similarly, conditional effects must also be used when data plot as in Figure 4.3. The situation illustrated in Figure 4.1 is almost as bad as the conditional effects differ in sign, which will always be the case when the lines cross.

So only for the Figure 4.4 scenario does it make sense to do a traditional analysis. Although we of course use software to perform these analyses, it is useful to see how the effect estimates are computed, and the computing formulas are given below, with "2" denoting the high level of the factor and "1" denoting the low level.

$$\begin{aligned} A &= \tfrac{1}{2}(A_2B_1 - A_1B_1 + A_2B_2 - A_1B_2) \\ B &= \tfrac{1}{2}(A_1B_2 - A_1B_1 + A_2B_2 - A_2B_1) \\ AB &= \tfrac{1}{2}[A_2B_2 - A_2B_1 + (A_1B_2 - A_1B_1)] \\ &= \tfrac{1}{2}(A_2B_2 - A_2B_1 - A_1B_2 + A_1B_1) \end{aligned}$$

## 4.4 WHY NOT ONE-FACTOR-AT-A-TIME DESIGNS?

As we consider statistically designed experiments and consider more than one factor, it seems logical to pose the following question: "Why not study each factor separately rather than simultaneously?" Indeed, this is frequently done. We can use Figure 4.1 to show why this won't work, and indeed it won't work in general, although one-factor-at-a-time designs can sometimes be useful. (See Section 13.1 for a detailed discussion of such designs.)

Assume that a company's engineering department is asked to investigate how to maximize process yield, where it is generally accepted that temperature and pressure have a profound effect on yield. Let factor $A$ denote temperature and factor $B$ denote

pressure. Three of the engineers are given this assignment, and their initials are BW, JC, and LM, respectively. Each engineer conducts his own experiment. Assume that BW and JC each investigate only one factor at a time, whereas LM decides to look at both factors simultaneously. Assume further that Figure 4.1 depicts what can be expected to result when both factors are studied together.

If engineer BW had used the low temperature and varied the pressure from low to high, he would conclude that the best way to increase the yield is to increase the pressure, whereas he would have reached the opposite conclusion if he had used the high temperature. Similarly, if engineer JC had set the pressure at the high level, he would have concluded that the best way to increase yield is to reduce the temperature, whereas he would have reached the opposite conclusion if he had used the low pressure level.

Engineer LM, on the other hand, would be in the proper position to conclude that interpreting a traditional main effects analysis would not be possible because of the interaction effect of the two factors.

This type of feedback is not available when factors are studied separately rather than together. These "one-at-a-time" plans have unfortunately been used extensively in industry. They are considered to have very little value, in general, although Daniel (1973, 1976, p. 25) discusses their value when examining three factors. Qu and Wu (2005) also discuss conditions under which one-at-a-time plans might be used. In particular, the designs can have value when there are hard-to-change factors. These designs are discussed extensively in Section 4.19.

## 4.5 ANOVA TABLE FOR UNREPLICATED TWO-FACTOR DESIGN?

The data from a designed experiment is typically presented, whenever possible, in an ANOVA table, as was done starting in Chapter 2. In this section we show why this won't work for an unreplicated $2^2$ design, and indeed won't work for any unreplicated two-level factorial design.

Let's try to construct the ANOVA table for the data in Figure 4.2. We could compute the sum of squares corresponding to each effect analogous to the way that sums of squares were computed in Chapter 1. Since we already have the effect estimates, however, it would be easier to obtain the sums of squares by squaring the effect estimates. In general, for any two-level factorial, $\text{SS}_{\text{effect}} = r(2^{k-2})(\text{effect estimate})^2$, with $r$ denoting the number of replicates and $k$ denoting the number of factors. Thus, for an unreplicated $2^2$ design this reduces to $\text{SS}_{\text{effect}} = (\text{effect estimate})^2$. The sums of squares for the Figure 4.2 data are thus 0, 0, and 400 for $A$, $B$, and $AB$, respectively.

Since we now have these sums of squares, we might attempt to construct an ANOVA table. Remembering from Section 2.1.3.3 that the degrees of freedom (df) for "Total" is always the total number of observations minus one, and that the df for a factor is always the number of factor levels minus one, we thus have $\text{df}(\text{Total}) = 3$, $\text{df}(A) = 1$, and $\text{df}(B) = 1$. The df for any interaction effect is always obtained as the product of the separate df of each factor that comprises the interaction. Thus, in this case we have $\text{df}(AB) = (1)(1) = 1$.

**TABLE 4.2 ANOVA for the Data in Figure 4.2**

| Source of Variation | df | SS | MS | $F$ |
|---|---|---|---|---|
| $A$ | 1 | 0 | 0 | |
| $B$ | 1 | 0 | 0 | |
| $AB$ (residual) | 1 | 400 | 400 | |
| Total | 3 | 400 | | |

If we add the df for $A$, $B$, and $AB$, we recognize immediately that we have a problem. Specifically, there is no df left for estimating $\sigma^2$. Thus, unless we have an estimate of $\sigma^2$ from a previous experiment (remember that experimentation should generally be thought of as being sequential) we have a case in which the interaction is said to be "confounded" (i.e., confused or entangled) with the "residual," where the latter might be used in estimating $\sigma^2$. We can summarize what we know to this point in the ANOVA table given in Table 4.2.

Notice that the $F$-values are not filled in. It is "clear" that there is no $A$ effect and $B$ effect since the sum of squares for each is zero. (Remember, however, we recently saw that each does have an effect on the response variable; their effect is simply masked by the interaction effect.)

It was stated in the section on one-factor ANOVA that the analysis is not influenced by whether the factor is fixed or random. This is not true when there is more than one factor, however. In general, when both factors are fixed, the main effects and the interaction (if separable from the residual) are tested against the residual. When both factors are random, the main effects are tested against the interaction effect, which, in turn, is tested against the residual. When one factor is fixed and the other random, the fixed factor is tested against the interaction, the random factor is tested against the residual, and the interaction is tested against the residual. (By "tested against" we mean that the mean square for what follows these words is used in producing the $F$-statistic.)

In this example the interaction is not separable from the residual because the experiment has not been "replicated"; that is, the entire experiment has not been repeated so as to produce more than one observation ($r > 1$) per treatment combination. This should be distinguished from *multiple readings* obtained within a *single* experiment, which does *not* constitute a replicated experiment (i.e., the entire experiment is not being replicated). This may seem like a subtle difference, but it is an important distinction that is discussed further in Section 4.7 (see also Box, Hunter, and Hunter, 1978, p. 319).

If a prior estimate of $\sigma^2$ is available, possibly from a previous replicated experiment with perhaps slightly different factor levels, that estimate could be used in testing for significance of the main effects. If a prior estimate is not available, we might still be able to obtain an estimate of $\sigma^2$.

Tukey (1949) proposed a test for detecting interaction of a specific functional form for an unreplicated factorial. The general idea is to decompose the residual into an interaction component and an experimental error component, and perform an $F$-test on the interaction. If the test is not significant, then $\sigma^2$ might be estimated using the

residual. It should be recognized, however, that this test will detect only an interaction that can be expressed as a product of main effects times a constant.

It should be noted that there is a difference, conceptually, between "experimental error" and "residual," and the latter cannot be used, in general, as a substitute for the former. Experimental error should be regarded as the variability that results for a given combination of factor levels in a replicated experiment, and is composed of variability due to factors not included in the experiment, sampling variability, and perhaps variability due to measurement error. A residual (as in residual sum of squares) may consist of various interaction terms that are thought to be not significant, in addition to experimental error. To estimate $\sigma^2$ using $\mathrm{MS}_{AB}$ in Table 4.2 would be to act as if no model fits the data because the only term that has a nonzero sum of squares is the $AB$ term. But we have a perfect fit using that term and there is also much information in the conditional effects for $A$ and $B$.

It is interesting to note that Tukey's test would not detect this interaction, since it can detect only interactions that are a constant times the product of main effects, as stated previously. In this case the test would indicate that the interaction is zero because the main effects are zero. We should remember that the test is not a general test for detecting the presence of interaction, nor can there be such a test for an unreplicated experiment.

This idea of experimental error versus residual is a very important one, and it is indicated in Section 4.6 how we can go wrong by using an interaction to estimate $\sigma^2$ for a particular set of actual data.

Can the analysis begun in Table 4.2 be completed? The analysis was actually completed *before* the (attempted) construction of the ANOVA table, as the data are not appropriate for analysis by an ANOVA table. We have seen that there is indeed a temperature effect and a pressure effect, and the interaction profile in Figure 4.1 clearly shows the strong interaction.

In the absence of a prior estimate of $\sigma^2$, the only course of action that would allow completion of the ANOVA table would be to use the interaction as the residual, and to test the two main effects against it. This, of course, would be sheer folly for these data as it would lead to the conclusion that nothing is significant, whereas in actuality all three effects are of significance.

This example was given for the purpose of illustrating how a routine analysis of data could easily lead to wrong conclusions. This message can also be found in other sources such as Box et al. (1978, p. 329) and Daniel (1976). The reader is referred to Daniel (1976, p. 20) for additional reading on the interpretation of data from a design of this type when a significant interaction effect exists.

It might appear that one solution to the problem of not being able to separate an interaction from a residual is simply to replicate the experiment. Although this is generally desirable, it is not always practical. One possible impediment is, of course, money, the data may be so expensive to collect as to preclude replication.

There are, however, some methods for assessing the possible significance of main effects and interactions in unreplicated experiments. One of these methods is due to Daniel (1959) and consists of plotting effect estimates on normal probability paper. This is illustrated in a later example.

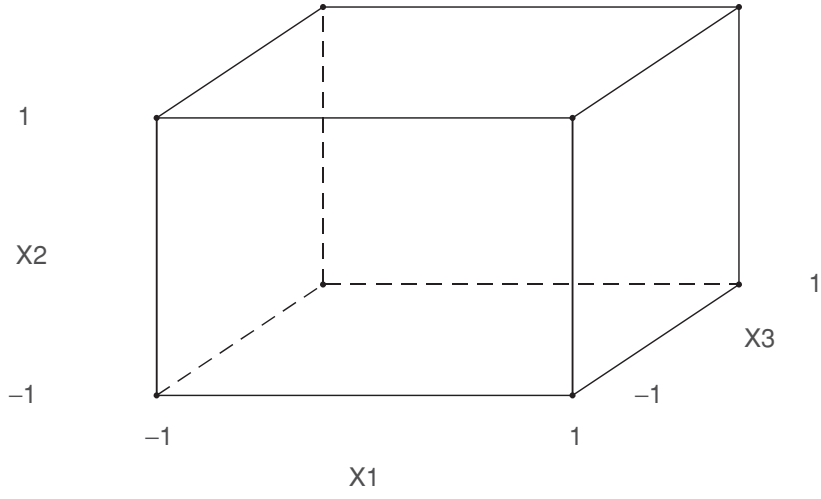



**Figure 4.5** Configuration of points in a $2^3$ design.

## 4.6 THE $2^3$ DESIGN

The $2^3$ design has been used in many applications and George E. P. Box, in lamenting the relative lack of use of experimental designs by engineers, has stated on occasion that just getting them to use a $2^3$ design would be a big step in the right direction. Quite frankly, this would not be a big enough step, however, because 8-point designs do not have good power for detecting differences between means.

The $2^3$ design is just an extension of the $2^2$ design to one additional factor, which would be denoted as factor $C$. There are seven estimable effects, although the $ABC$ interaction generally will not be significant. If it is significant, that could be caused by one or more data points. This is illustrated in Section 4.10. There are various ways in which we can think of the $ABC$ interaction. Specifically, $ABC = AB(C) = AC(B) = BC(A)$. The expression $AB(C)$ means the interaction between the $AB$ interaction and factor $C$ (and similarly for the other expressions). This interaction will be significant if the two $AB$ interaction profiles (as in Figure 4.1, for example) differ considerably over the two levels of $C$.

The points in the design form a cube, as is shown in Figure 4.5.

Recall the discussion in Section 4.1 regarding the magnitude of effects that can be detected with a $2^2$ design. Does an unreplicated $2^3$ design have the same shortcoming? Only a slight problem, because the smallest difference that can be detected is $2.29\sigma$ (using the expression for $\Delta$ given in Example 4.2), with as before, a probability of .90 of detecting the difference, a significance level of $\alpha = .05$, and $\sigma$ assumed to be known. We can relate this to a $z$-value or a $t$-value from a simple hypothesis test in an introductory course for testing $H_0 : \mu_1 = \mu_2$. We would generally reject the hypothesis when the test statistic, $z$ or $t$, has a value much greater than 2, and we would certainly reject the hypothesis for a value greater than 3. Relative to these

**TABLE 4.3 Number of Replicates Needed to Detect a Minimum Detectable Effect of $2.0\sigma$ for $2^k$ Designs**

| $k$ | Number of Replications |
|---|---|
| 3 | 2 |
| 4 | 1 |
| 5 | 1 |
| 6 | 1 |

numbers, we would prefer to see a multiplier less than 2.29, but this is not a serious shortcoming of the $2^3$ design.

A $2^3$ design would have to be replicated in order for the multiplier of $\sigma$ to be clearly acceptable. For example, the multiplier drops from 2.29 to 1.62 for two replicates (i.e., two observations per treatment combination), 1.32 for three replicates, and 1.15 for four replicates. The last two, in particular, are much more reasonable values than the multiplier 2.29 for the unreplicated $2^3$ design.

If a practitioner considers, for example, a multiplier of 2.0 to be acceptable, Table 4.3 could be used to determine how many replications are needed for a $2^k$ design with various values of $k$. Again, this assumes a probability of .90 of detecting the difference and a significance level of $\alpha = .05$. For extensive tables, the reader is referred to Lynch (1993).

Note, however, that the tables of Lynch (1993) assume that $\sigma$ is estimated, and more specifically that it is estimated from a certain number of degrees of freedom. Hence, the numbers given there will not agree with formulas that assume $\sigma$ to be known. Of course, typically $\sigma$ is unknown and the likelihood of detecting effects that are real and of considerable magnitude will depend upon whether or not the design is replicated. This is because whereas the analysis is straightforward for a replicated design, the estimate of $\sigma$ for an unreplicated design is affected by the magnitude and number of the "small" effects, and a poor estimate will result when the number of such effects is small. These methods are discussed and illustrated in Section 4.10.

These methods generally depend upon the assumption that high-order interactions are not significant, but the folly of using the highest-order interaction in an unreplicated $2^3$ factorial design was illustrated by Daniel (1976) and discussed in Ryan (2000).

### Example 4.5

In describing a cement experiment, Daniel (1976) stated that $\sigma$ was known to be about 12, whereas the estimate of $\sigma$ obtained using the *ABC* interaction is 3.54—approximately $1/3$ of the assumed value.

Other problems with those data that indicate the need for a careful analysis include the large interactions, with the effect estimates shown in Table 4.4.

There are two numbers that preclude a customary main effects analysis: the *BC* effect estimate of 47.5 and the *AB* effect estimate of 13.5. That is more than half the absolute value of the *C* effect estimate, which means that the conditional main effects

**TABLE 4.4 Effect Estimates from Example in Daniel (1976)**

| Effect | Estimate |
|---|---|
| *A* | 15.5 |
| *B* | −132.5 |
| *C* | −73.5 |
| *AB* | 13.5 |
| *AC* | 1.5 |
| *BC* | 47.5 |
| *ABC* | 2.5 |

of $C$ will differ considerably. These are easy to compute because, as stated previously, they are simply the main effect estimate plus and minus the interaction effect estimate (see chapter Appendix A for details). That is, $C \pm BC = -73.5 \pm 47.5$, so the $C$ conditional main effect estimates are $-121$ and $-26$. These are the $C$ effect estimates that would be obtained by splitting the data into two halves on factor $B$, using first the low level of $B$ and then the high level of $B$, respectively.

In like manner it can be seen that the $B$ conditional main effects are $-180$ and $-85$, which also differ greatly. Similarly, the conditional main effects of $A$ that are obtained by splitting the data on $B$, since the $AB$ interaction is large relative to the $A$ effect, are 2 and 29.

In this example all of the conditional effects had the same sign. This will often not be the case and they will be of opposite signs for interaction plots of the form shown in Figure 4.1 and anything that resembles such a plot. For less extreme configurations, either or both sets of the two conditional main effects could have different signs. Exercise 4.3 is an example of this, and Wu and Hamada (2000, p. 109) give another example and make an important point that when the conditional main effects differ in sign, there may be a point between the levels in the experiment for which the response is a flat function of the factor. As they point out, this could have important ramifications if robust parameter design (see Chapter 8) were used, as we would then expect the response to not vary much over, say, a manufacturing variable that could not be tightly controlled during production, but which could be controlled within a reasonable range.

Although it is apparent from the list of averages in Table 4.4 that a conditional effects analysis must be used, the same message would come from an ANOM display, with the latter showing not only the magnitude of the effect estimates but also the average response values that determine the numerical value of the effect estimate. Of course practitioners would want to know these average response values, in addition to the effect estimates.

### Example 4.6

Kinzer (1985) described an experiment that utilized a $2^3$ design plus two centerpoints, with the latter used to detect curvature of the response if it existed since curvature

cannot be detected with only two levels. The objective of the experiment was to determine influence of curing conditions on the strength of a composite material product. The development engineers identified three factors that they felt contributed most significantly to the product strength: autoclave temperature, autoclave time, and air cure. Each point was replicated five times, but only the average of the five values was given in Kinzer (1985).

The analysis was performed in the raw units of the factors, something that should not be done if interaction terms are to be fit, as discussed in Section 4.1. This is because an interaction term will not be uncorrelated with the factors in a $2^3$ design when raw form is used, but will be uncorrelated when coded form is used. This can create contradictory results between the two forms, as was illustrated in Section 4.1 and by Ryan (2000, p. 401).

For this example, Kinzer (1985) used stepwise regression to arrive at a model, which had the following terms: X1, X3, and X1 $*$ X3. This model was selected from an initial fit that had all the linear terms, all four of the possible interaction terms, and a single term that represented the sum of the squares of the factors. There are high correlations between some of these terms, however, which can cause problems when a technique such as stepwise regression is used. Nevertheless, when the reported data were used (not the data for all replicates), the same conclusion was reached regardless of whether the analysis was performed using coded form or raw form. Specifically, X1 and X3 are significant main effects, and the model that uses these terms has an $R^2$ value of .934. This is considerably higher than the $R^2$ value of .794 given by Kinzer (1985) for his model with three terms, when all the data are used.

There are two important points here: (1) a different model is obtained when averages are used instead of all the raw data, and (2) $R^2$ is much lower when all the data are used. The latter is generally true and we don't want the selected model to depend upon whether or not averages were used.

## 4.7 BUILT-IN REPLICATION

Although the analysis of unreplicated factorials can be challenging when there is a high percentage of real effects, as will be illustrated starting in Section 4.9.1, it is worth noting that under certain circumstances it is reasonable to proceed as follows. Assume that a $2^5$ design is used and one of the main effect estimates is practically zero, the conditional effects for the factor are also very close to zero, and all interaction effect estimates involving the factor are also close to zero. It would then be safe to drop the factor from the analysis. Doing so then produces a replicated $2^4$ design with two replicates, which could then be analyzed using ANOVA.

The situations under which this approach can be used will be quite limited, however, and indiscriminate dropping of factors can lead one astray. For example, let's assume that an experimenter generally uses a $2^5$ design and drops the factor with the smallest main effect estimate, reasoning that all five factors are not going to be significant. But a small main effect estimate could be caused by a large interaction, with the conditional main effects differing considerably from zero. If so, dropping the factor

would be the wrong thing to do. In this case it would be better to rely on a normal probability plot analysis (see Section 4.9.1).

How often will we be better off dropping a factor or two and performing the analysis using ANOVA as opposed to using a probability plot analysis? Any answer to that question is of course conjectural, but let's consider some possible scenarios. Again assume that a $2^5$ design is used and the main effect estimates are $A$ (16.82), $B$ (24.84), $C$ (31.16), $D$ (19.43), and $E$ (1.16). Looking at only these numbers, it might seem safe to eliminate factor $E$ and analyze the data as a replicated $2^4$, although of course the experiment wasn't designed that way. We can't do this, however, without looking at the magnitude of all interactions involving factor $E$. If any of those are large, then factor $E$ cannot be removed from the analysis because then at least one conditional effect will differ considerably from zero. If all four of the two-factor interactions involving $E$ are close to zero, then there should be enough terms to use in constructing a meaningful pseudo-error term (PSE), since hardly any of the higher-order interaction terms could be expected to be large.

When the importance of examining conditional effects is considered and the discoveries that might be made when all factors are used in the initial analysis are also considered, it is somewhat difficult to imagine scenarios where the probability plot approach will fail and dropping a factor or two and using an ANOVA analysis would be the right thing to do.

## 4.8 MULTIPLE READINGS VERSUS REPLICATES

It seems safe to say that true replicated experiments rarely exist in practice, even though probably every experiment that even resembles a replicated experiment is analyzed as if it were a replicated experiment.

The conditions that must be met for an experiment to be termed a replicated experiment are rather stringent. In general, everything must be reset. For example, if factor levels in an experiment were set by turning knobs, then the knobs must be turned to some neutral position and then set at the desired level before there is an additional run within a set or runs or between replicate runs. To obtain observations on the response variable without resetting any factors is to obtain multiple readings, not replicate values, and the errors will not be independent, as pointed out by Lucas (1999). As stated in Lucas (1999, p. 29) in his 1997 Annual Quality Congress talk, Lucas asked three questions of the over 200 people who attended his talk. All of them raised their hands when he asked them if they were involved in running experiments and if they used randomization. However, when they were asked if they set each factor to a neutral level and then reset it, only four people raised their hands.

Based on the results of this informal survey, it seems safe to assume that very few experiments in which randomization is used actually have complete randomization.

Even if the proper procedures are followed, it still may not be possible to have true replicates. For example, Ryan (2004) described a lead recovery from paint experiment in which true replicates were not possible because particle sizes could not be ground to be an exact size. Since the extent of the departure from true replicates may not

be known under such conditions, nor the effect of that departure on the statistical analysis, it is best to also analyze the data by obtaining an average for each replicated point and performing an analysis using the averages, as was done in that case study.

## 4.9 REALITY VERSUS TEXTBOOK EXAMPLES

Experiments do not always work out as designed, and sometimes practical considerations can result in some unusual designs. One such example is the study described by Ryan (2004). Five factors were examined, each at two levels. There were 112 experimental units (specimens) available, and 7 could be used in each run. Of course, 112 is not an integer multiple of $2^5$, so the only way to produce 112 runs starting from a $2^5$ design is to use something other than an equal number of replicates for each treatment combination. That is what was done, with half of the treatment combinations replicated three times and the other half replicated four times.

The fact that an unequal number of replicates was used creates problems, regardless of how the 32 combinations are split into two halves. The problems can be minimized if the split is performed in an optimal manner, however. This is explained when we return to this example in Section 5.12. (A detailed discussion is appropriate for Chapter 5 because splitting the data in half strongly relates to fractional factorials.)

### 4.9.1 Factorial Design but not "Factorial Model"

Ideally, the design that is used should facilitate the estimation of the coefficients of the model that results from analysis of the data from the experiment. When a factorial design is used, there is the tacit assumption that if all the important factors have been identified and are included in the experiment, a good model will be a linear model that includes some or all the factors used in the experiment. That isn't necessarily true, however. Bisgaard, Vivacqua, and de Pinho (2005) illustrated this by using an example from Brownlee (1953). A $2^4$ design was used and a normal probability plot of the effect estimates (given in Fig. 4.6) shows that no effect is significant. (Lenth's PSE, which is shown on the plot, is explained in detail in Section 4.10.) At first thought, this might seem to suggest that the experimenters did a poor job of selecting the effects to be studied, but we need to look further.

The Pareto chart of effect estimates exhibits some real oddities (Fig. 4.7). In particular, although no effect is even close to being significant, the stairstep appearance of the chart is odd, as is the fact the four-factor interaction is much larger than two of the main effect estimates. (Note that the absolute values of the effect estimates are plotted here on the Pareto chart, whereas the actual values of the effect estimates are plotted on the normal probability plot.)

The determination of the threshold value that is denoted by the vertical line needs some explanation. The line is drawn at the value of $t \times$ PSE with the latter being Lenth's PSE that is explained in Section 4.10, as stated, and $t = t_{.025,\nu}$ with $\nu$ denoting the degrees of freedom defined by MINITAB, which was used to produce this graph, as the total number of effects divided by 3. The value of $t \times$ PSE also determines whether

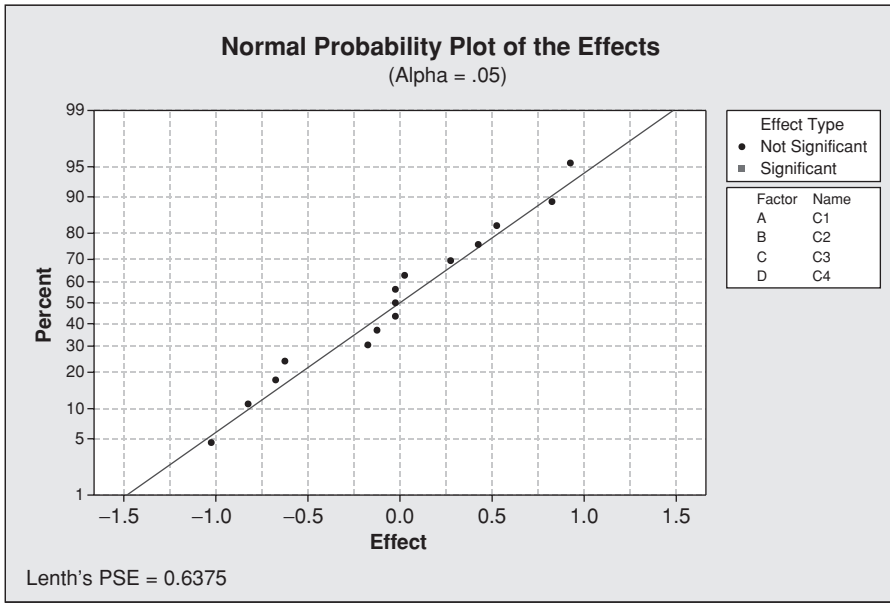



**Figure 4.6** Normal probability plot of effect estimates.

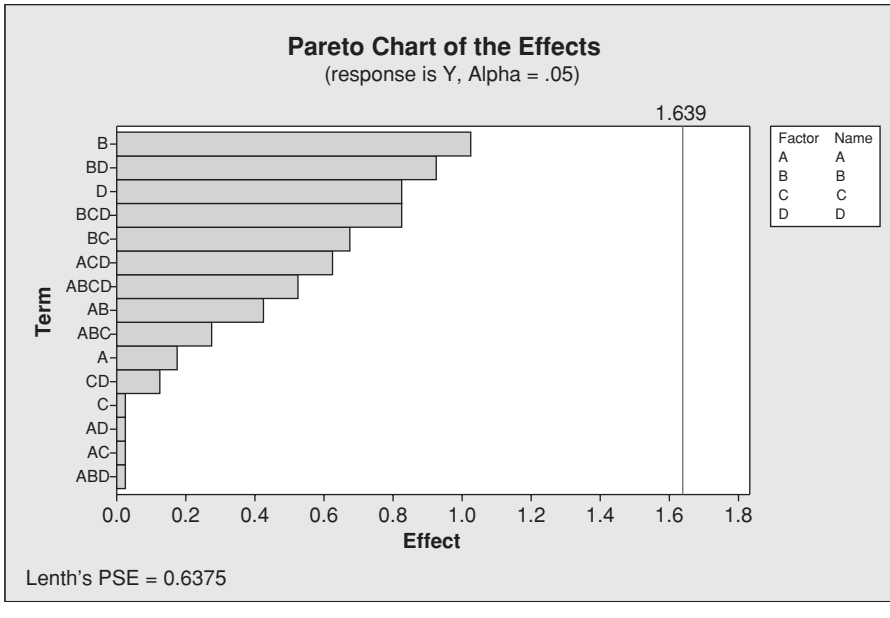



**Figure 4.7** Pareto chart of effect estimates.

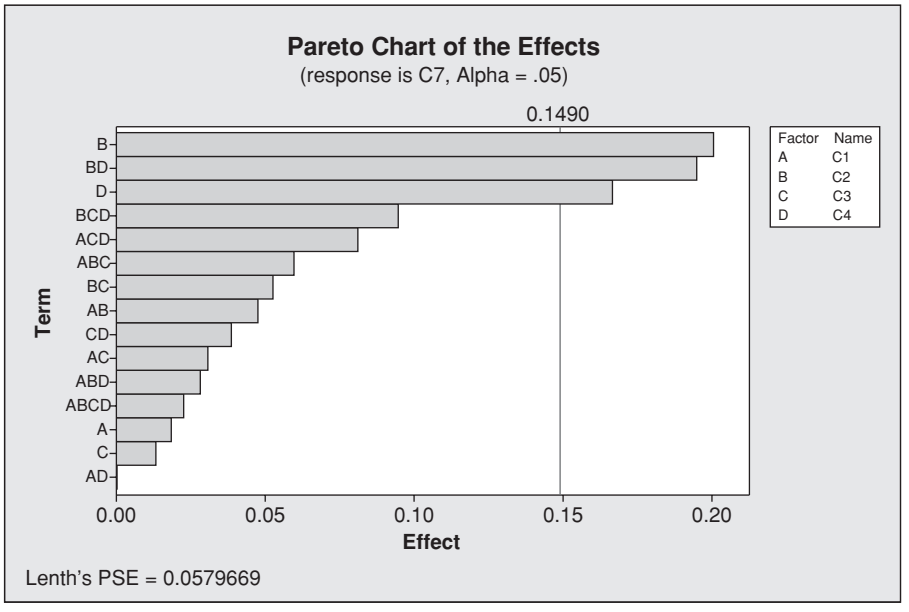



**Figure 4.8** Pareto chart of effect estimates after transformation.

or not any effects in a normal probability plot are declared significant, although the threshold values are $\pm t \times$ PSE, since effect estimates can be negative and some of them are negative in this example. We can see why no effect was declared significant in Figure 4.6 as the PSE value is given as 0.6375 and all but one of the effect estimates is in the range $(-1, 1)$ and the one that is outside the range is just very slightly outside the range. Since $t_{.025,\nu} > 1.96$ for any value of $\nu$, it follows that $0.6375 \times t_{.025,\nu}$ will exceed 1.2, which clearly exceeds the absolute values of any of the plotted effect estimates. (More specifically, $(0.6375)(2.5706) = 1.639$.)

Since the ratio of the largest response value to the smallest response value exceeds 3, there is the potential for improving the fit of the model by transforming the response variable. It should be noted, however, that the response variable is generally transformed for the purpose of trying to meet the model assumptions, and not for improving the fit of the model. Nevertheless, Bisgaard et al. (2005) elected to use the reciprocal transformation, stating that in attempting to find a suitable transformation "... for most practical purposes it is easier to simply proceed by trial and error." A trial-and-error approach is reasonable only if we are willing to restrict consideration to common transformations, of which there is a small number.

The Pareto chart that results from the analysis using the transformation is shown in Figure 4.8.

We see immediately that this is much better, as there are three effects that are clearly larger than the rest of the effects and are judged as being significant. Thus, we might wish to fit a model with the $B$ and $D$ main effects and the $BD$ interaction. We

should not be quick to dismiss the main effect of $C$, however, since the $AC$ interaction is the fourth largest effect. Nevertheless, since the interaction does not show as being significant, we will proceed with these three effects.

If we fit a regression model with the transformed response variable, $Y^{-1}$, we obtain $\widehat{Y}^{-1} = 0.414 + 0.100\,B - 0.0833\,D - 0.0975\,BD$. Bisgaard et al. (2005) took a different approach, however, concluding that this is a "critical mix situation," meaning that the only treatment effect occurs when factor $B$ is at the high level and factor $D$ is at the low level. Both factors have presence/absence levels with the high level of factor $B$ signifying that the crude material used in the experiment was "boiled" and the low level signifying that it was "not boiling"; and the low level of factor $D$ means that the crude material was precipitated from either of two solvents, with the high level signifying that this did not occur.

The response values at 3 of these 4 treatment combinations of the total of 16 do support this view, but the response at the fourth of these treatment combinations (0.53) is not much larger than the next two largest values, both of which are 0.45. Bisgaard et al. (2005) used an indicator variable approach and arrived at the model $\widehat{Y}^{-1} = 0.318 + 0.377\ \mathbf{I}(x_{\mathrm{B}}, x_{\mathrm{D}})$, with $\mathbf{I}(x_{\mathrm{B}}, x_{\mathrm{D}}) = 1$ if $B = 1$ and $S = -1$, and 0 otherwise. Not surprisingly, the standardized residual at the response value of 0.53 is $-2.13$. Although this is not a large value, it is large enough to raise questions about the critical mix assumption. Furthermore, when the first model is fit with the three terms, the fit is slightly better as the residual sum of squares is 0.1136 compared to 0.1163 for the indicator variable approach.

## 4.10 BAD DATA IN FACTORIAL DESIGNS

If factors are chosen judiciously and an experiment run properly, a reasonable number of effects should be significant. Something is generally wrong when we obtain extreme results such as no effects being significant or almost all effects being significant.

The latter will often be caused by bad data. As an example, consider the following data resulting from use of a $2^3$ design, with the data given in Yates order: 16, 22, 18, 24, 19, 23, 20, and 78. Assume that the last observation should have been 28 but was misread as 78. One extreme bad data point such as this will result in none of the effects estimates being small, which is nothing being declared significant by Lenth's (1989) method because a reasonable pseudo-error cannot be constructed. Indeed the normal probability plot for this example is given below and we can see that (1) the plot looks very peculiar, and (2) no effects are declared significant.

This plot was generated using MINITAB; the method that MINITAB uses for determining if an effect is significant is due to Lenth (1989). The latter uses a PSE that is defined as follows. Let $s_0 = 1.5 \times \text{median}\ c_j$, with $c_j$ denoting the estimate of the $j$th estimable effect. Then the PSE is computed as

$$\text{PSE} = 1.5 \times \underset{|c_j|<2.5s_0}{\text{median}}\ |c_j|$$

with the PSE essentially computed from a trimmed median of $|c_j|$ values, as the median is computed using only values of $|c_j|$ that are less than $2.5s_0$.

Lenth's method and alternatives were investigated by Haaland and O'Connell (1995), whose modification was not as simple as Lenth's method. They used the general form

$$\widehat{\sigma}_{\text{PSE}}(q, b) = a_{\text{PSE}}(q, b) \cdot \text{median } \{|\hat{\theta}_i| : |\hat{\theta}_i| \leq b \cdot s_0(q)\} \tag{4.3}$$

with $a_{\text{PSE}}(q, b)$ denoting a consistency constant, $b$ denotes a tuning constant, the $|\hat{\theta}_i|$ are the absolute values of the ordered effect estimates, and $q$ is obtained from the initial robust estimator of scale $s_0(q)$. The latter has the general form

$$s_0(q) = a_0(q) \cdot \text{quantile } \{q; |\hat{\theta}_i| \, i = 1, \ldots, k\} \tag{4.4}$$

with $a_0(q) = 1/\Phi_0^{-1}(q)$ and $\Phi_0^{-1}(q) = \Phi^{-1}[(q + 1)/2]$, with $\Phi$ denoting the cumulative distribution function of the standard normal distribution. Thus, for example, if $q = 0.5$, then $s_0(q) = 1.48 \cdot$ median $(|\widehat{\theta}_i|)$. Lenth's method is essentially equivalent to $\widehat{\sigma}_{\text{PSE}}(0.5, 2.5)$. As explained by Haaland and O'Connell (1995), Daniel (1959) used $q = 0.683$ in $s_0$, and also used this value of $q$ in $\widehat{\sigma}_{\text{PSE}}$. He did not, however, have a trimming threshold of $b \cdot s_0(q)$ but rather trimmed large effects by inspection. We will return to the paper by Haaland and O'Connell (1995) in Section 5.3.1, since their paper was on unreplicated, 16-point fractional factorials, but the results also apply to unreplicated full factorials.

Of course the objective with any PSE and with any method for identifying significantly large effects in an unreplicated factorial (or fractional factorial, as in Chapter 5) is to have a method that works well when there are few real effects and also when there are many real effects, yet has a tolerable false alarm rate.

Consider the form of $s_0(q)$. A single bad data point (as in this example) can cause none of the effects to be small without any of the effects being particularly large, as will be seen in Figure 4.9. When this occurs, $s_0(q)$ will break down regardless of what value of $q$ is used. Assume that $q = 0.5$ since this is a common choice. As is stated in Section 4.11, Daniel (1976, p. 75) stated that seven significant effects is about average for a $2^5$ design. Notice that this is far fewer than half of the estimable effects. In general, less than half of the estimable effects should be real, so the median of the $|\widehat{\theta}_i|$ should generally be estimating an effect that is zero.

Notice that when $s_0(q)$ is inflated, the value of $b \cdot s_0(q)$ in $\widehat{\sigma}_{\text{PSE}}(q, b)$ will also be inflated. The consequence of this is that certain large effects will not be trimmed in determining median $(|\widehat{\beta}_i|)$, with the consequence that certain large effects will not be trimmed in computing $\widehat{\sigma}_{\text{PSE}}(q, b)$. The consequence of this would be that $\widehat{\sigma}_{\text{PSE}}(q, b)$ is inflated, which would reduce the number of effects that are identified as real. If the inflation of $\widehat{\sigma}_{\text{PSE}}(q, b)$ is severe, then multiple real effects may not be identified as such.

For the present example, with 28 being recorded as 78, we can regard the effects computed using 28 as the truth and compare the results with those obtained using 78. Using the value 28, Lenth's PSE = 1.5 and $A$ is identified as a real effect. As

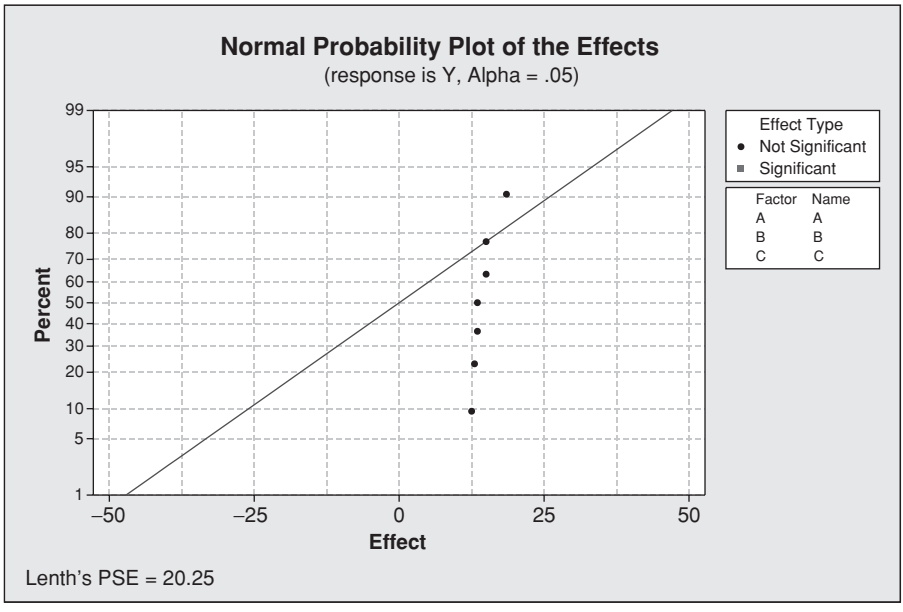



**Figure 4.9** Effect of a single bad data point.

is shown in Figure 4.9, PSE = 20.25 when 78 is used and no effect is identified as significant. This shows how bad data can result in an erroneous conclusion when a normal probability plot of effects estimates is used.

There is an important point that should be made regarding this plot. Before hypothesis test statistics became a part of probability plots of effects, the recommended use of such plots was to construct a line through the majority of the points. In a typical experiment most effects will *not* be significant, so points that lie more than slightly off the line would represent significant effects. Notice that most of the points do practically lie on a line, but the line in the plot is not drawn through the center of the points. When the axes are labeled in this manner, the line must go through the point (0, 50 percent), but notice that there is no way to draw a line through this point and have it lie even close to the majority of the points.

We might ask why none of the effects are identified as being significant since most of the points lie well off the line. As indicated previously, Lenth's method, and indeed any method for obtaining a pseudo error from an unreplicated factorial, implicitly uses small effects in obtaining the pseudo error, but here practically all the effects are of virtually the same size, as can be seen from Figure 4.9. This causes none of the effects to be identified as significant, despite the fact that one of the effect estimates is larger than two of the observations, and the average effect estimate is more than half of the average of the seven good observations. Thus, the effects are actually large relative to the magnitude of the observations but the error causes none of the effects to be identified as significant.

In this example the bad data point was obvious just from inspection of the data, but bad data points won't always be so obvious. In general, whenever, the results differ from what might be expected, it is important that the data be checked carefully. If no bad data can be detected, it would be wise to perform a follow-up experiment and compare the results with the original experiment.

A follow-up experiment may be necessary even when bad data are detected and discarded, as discarding a bad data point will cause the orthogonality of an orthogonal design to be lost. For example, if the bad data point in this example is deleted, the correlations between the columns of the design matrix are all $-.167$. These are not large correlations, but the definition of an effect estimate becomes somewhat shaky as there will be an unequal number of observations at the high and low levels of each factor, and there will also not be a direct relationship between the effect estimates and the regression coefficients, as there is when all the data are available. Furthermore, the *ABC* interaction could not be estimated because of the loss of one degree of freedom.

Bad data that are detected and discarded present the same general type of problem as is caused by data that are simply missing (one or more runs that were not performed because of equipment that broke, etc.). Missing data and incomplete data that result from bad data being detected and discarded are discussed in the following sections.

Box (1990–1991b) illustrated a simple method, due to Daniel (1976), for finding bad data in factorial designs. The method consists of constructing a normal probability plot of effect estimates, which would normally be done anyway for an unreplicated factorial design. For illustration, Box used a $2^4$ design. When we think about the configuration of $+1$ and $-1$ components in the table for a $2^4$ design, analogous to Table 4.1, it should be apparent that each row of the table has eight plus signs and eight minus signs. Therefore, if a number like 10.3 is erroneously recorded as 103, eight of the effect estimates will be greatly overestimated and the other eight will be greatly underestimated, perhaps resulting in two distinct lines of plotted points on a normal probability plot.

Assume that the data, in Yates order, are supposed to be 4.8, 5.9, 10.3, 6.4, 11.6, 12.6, 9.8, 12.4, 6.2, 8.4, 4.5, 7.8, 7.0, 8.1, 10.5, and 10.3 but the 6.2 is erroneously recorded as 62. Of course this number is so much bigger than the other numbers that we would know something is wrong just by looking at the numbers. With such a large error, this should cause two distinct lines on the normal probability plot, which is what we see in Figure 4.10.

Of course the signal from the plot would not be as strong if the bad data value was considerably smaller, as the reader is asked to show in Exercise 4.34.

An interesting case study that illustrates the steps that led to the detection of bad data in a designed experiment is described by Mark Anderson and Pat Whitcomb of Stat-Ease, Inc., in their article "How to use graphs to diagnose and deal with bad experimental data," which can be viewed at http://www.statease.com/pubs/baddata.pdf. The authors make an important point that an outlier is an outlier relative to the fitted model, which might not be an appropriate model. They illustrate this with the first of two datasets from Box (1990) that they use for illustration. Specifically, they show that two outliers disappear when a log transformation of the response variable is used.

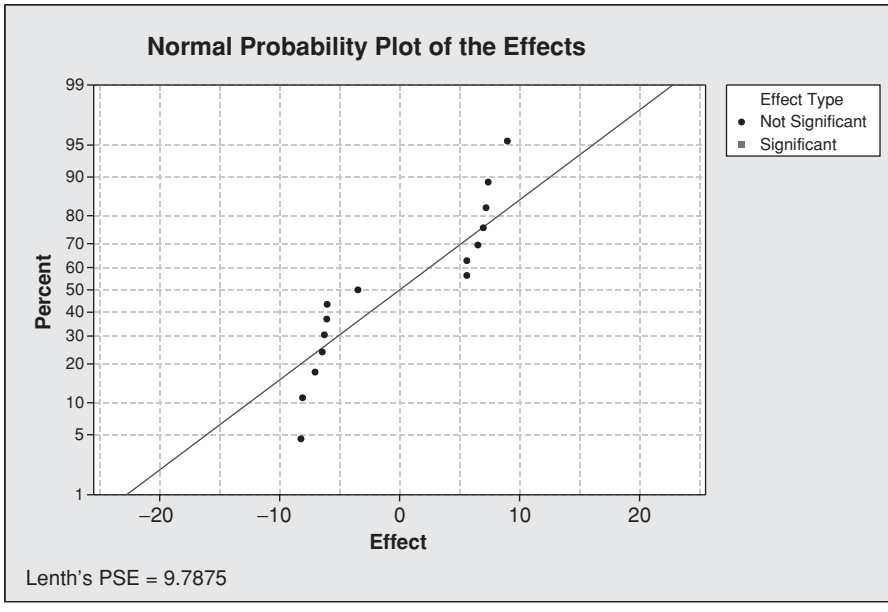



**Figure 4.10** Effect of a bad data point in a $2^4$ design.

### Example 4.7

Yin and Jillie (1987) described the application of a $2^4$ design in developing a nitride etch process. Although the authors apparently did not identify any bad data, suspicions arise when we examine the data. In particular, the absolute values of the effect estimates given in Table 4.5 exhibit some peculiarities.

What is most disturbing about this list is the presence of a two-factor interaction as the second largest effect, and quite surprisingly, the four-factor interaction being the fifth largest effect. Indeed seven of the nine largest effects are interaction effects. This is quite surprising and greatly complicates the analysis. Since the *ABCD* interaction effect is an order of magnitude larger than the seven smallest effects, we should not be surprised if it is declared significant in a normal probability plot assessment. The plot is given in Figure 4.11.

Seven of the 15 effects are judged significant by Lenth's method that is used in this output produced by MINITAB, with the default value of $\alpha = .10$ used. (The *ABCD* effect is also significant when $\alpha = .05$ is used.) It is disturbing that five of the seven significant effects are interaction effects. One or more extreme observations could cause such an anomaly, so it is desirable to search for extreme observations. Accordingly, a dotplot of the response values is given in Figure 4.12.

Obviously there are some extreme points, in addition to small groups of points that are clearly separated from other groups. The four largest values, which seem to be outliers, all occurred at the four treatment combinations of the high level of factor $D$ and the low level of factor $A$. Since the estimate of the $AD$ effect is obtained as

**TABLE 4.5 Rank Order of Effect Estimates by Largest Absolute Value**

| Effect | Estimate |
|---|---|
| *D* | 303.1 |
| *AD* | 153.6 |
| *A* | 101.6 |
| *BC* | 43.9 |
| *ABCD* | 40.1 |
| *BCD* | 25.4 |
| *AC* | 24.9 |
| *ABC* | 15.6 |
| *AB* | 7.9 |
| *C* | 7.4 |
| *ACD* | 5.6 |
| *ABD* | 4.1 |
| *CD* | 2.1 |
| *B* | 1.6 |
| *BD* | 0.6 |

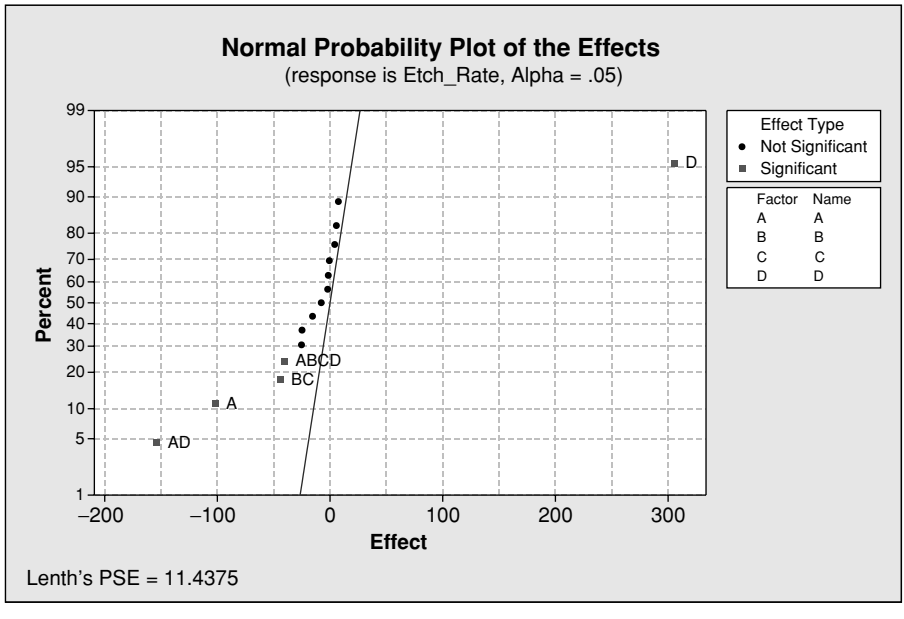



**Figure 4.11** Normal probability plot of effects for Yin and Jillie (1987) data.

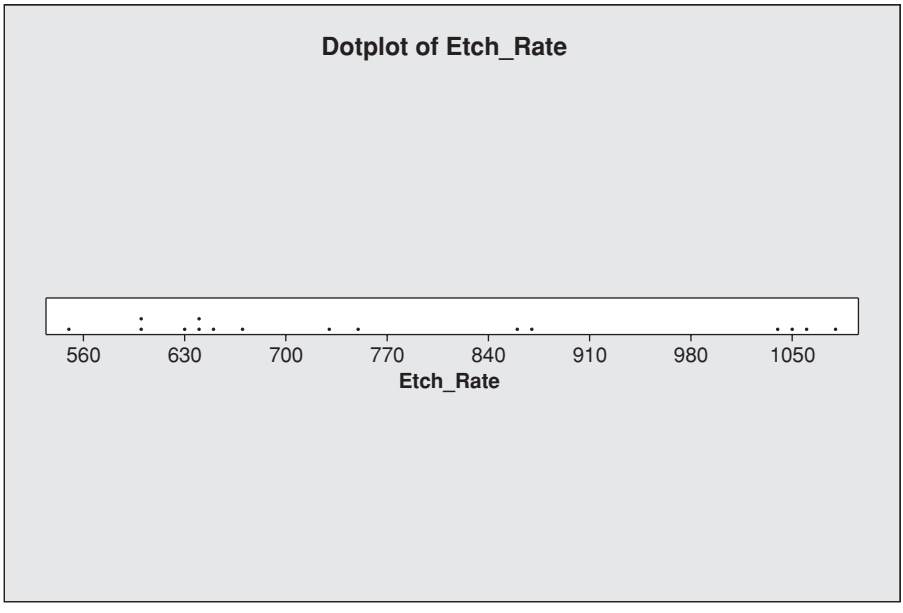



**Figure 4.12** Dotplot of etch rate data.

the average of the eight observations for which, using the coded levels, $A$ times $D$ is positive minus the average of the other eight observations for which $A$ times $D$ is negative, having four large observations occur at the treatment combinations for which $A$ times $D$ is negative is likely to cause a large negative value, as occurs here.

Clearly the authenticity of these four values should be verified. It may simply be that the specific levels of these two factors "interact" as, say, in a chemical experiment to produce extreme values. Or perhaps there are some major measurement issues for this combination of levels.

If these four values are valid, then a conditional effects analysis *must* be performed, especially since the $AD$ interaction is considerably larger than the $A$ effect. The conditional main effects of $A$ are $-52$ and 255.2. Clearly the unconditional effect estimate of 101.6 does not well represent either of these two numbers. It is especially troublesome that they differ in sign. A conditional main effect being negative and not close to zero while the unconditional main effect is a large positive number could result in an erroneous conclusion being drawn, although this sign difference could be of value in robust parameter design, which is covered in Chapter 8.

The significance of the $ABCD$ interaction cannot be explained by the four large values because two of them occur at treatment combinations for which the product of the four factor levels is positive and two at treatment combinations for which the product is negative. This makes the significance of the interaction rather perplexing. The significance of that interaction is due in large part to the fact that the response value that is by far the smallest occurs at a treatment combination for which the product is positive, as does the second smallest value.

If all the observations were valid data points—and that seems highly questionable—it would be necessary to do a full conditional effects analysis, starting with the three-factor interactions, which are affected by the large four-factor interaction. In particular, the *BCD* effect was judged significant in the normal probability plot analysis and the two conditional effects differ in sign because of the large four-factor interaction.

In essence, large high-order interactions create a domino effect because they render meaningless the lower-order interaction effect estimates and main effect estimates. Unfortunately, the analysis of conditional effects is complicated by the fact that conditional effects are correlated across different effects, and conditional effects and unconditional effects have different variances. Nevertheless, if these were valid data, such an analysis would have to be attempted, but will not be attempted here since the significant four-factor interaction casts aspersions on the data.

Not only can there be bad data with experimental designs, but of course there can also be errors in the listing of a design layout, as is illustrated in Section 13.4.1.2, along with a method for detecting the error and how it should be corrected.

### 4.10.1 ANOM Display

ANOM displays for a single factor were used in Chapter 2. When there are multiple factors, and especially when there are four or more factors, some thought must be given to how the display should be constructed. We want to see the magnitude of interaction effects relative to the corresponding main effects, and of course we want to determine which effects are significant. There are some obvious options. One possibility would be to display the effect estimates in decreasing order of magnitude (as in Table 4.5) without regard to the order of the effects regarding main effects and interactions. Another possibility would be to use Yates order (see Section 4.1). A third possibility would be to list all main effects in decreasing order of magnitude, followed by all two-factor interactions ordered in the same way.

When there are only two factors, the appropriate order is obvious: Show the main effects followed by the single interaction, as is shown in Figure 4.13.

The vertical lines for factors $A$ and $B$ are self-explanatory as each line connects the average responses at the low and high levels of the factor. The line representing the interaction requires some explanation, however. To facilitate this explanation, we will let "1" represent the low level and "2" the high level. Then Lbar $= (1/2)(\overline{A_1B_1} + \overline{A_2B_2})$ and Ubar $= (1/2)(\overline{A_1B_2} + \overline{A_2B_1})$, with the bar over the treatment combination designating the average response value for that treatment combination. Obtaining the interaction effect estimate as Lbar $-$ Ubar is in accordance with the expression for the $AB$ effect given in Section 4.3. The Lbar/Ubar notation is essentially the same notation that is used in Ryan (2000) and was originally used by Ellis Ott.

The display shows that although no effect is significant at the $\alpha = .05$ level, the interaction is large relative to the main effects, as the interaction dwarfs the $B$ effect estimate and is almost the same size as the $A$ effect estimate. The fact that the $A$ effect is not far from being significant means that the conditional effects of $A$ will

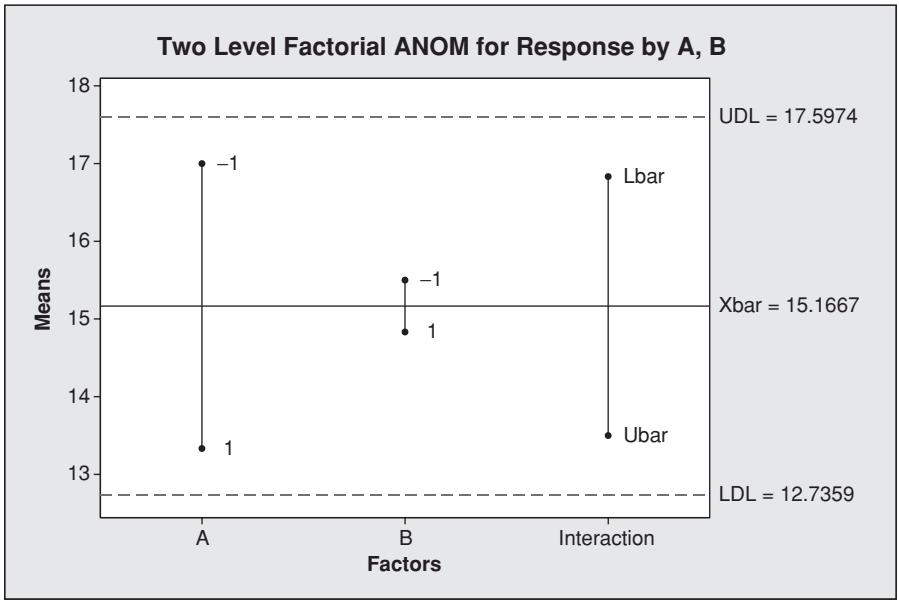



**Figure 4.13** ANOM display for two factors.

differ considerably and thus be "of significance." It can be shown that the conditional effects are $-7.00$ and $-0.33$, compared to the main effect of $-3.667$. Thus, one of the conditional effect estimates is almost double the main effect estimate, whereas the ratio of the standard deviations of the effect estimates is 1.4:1. Thus, the conditional effect estimate of $-7.00$ looks "more significant" than the main effect estimate. Of course, the conditional effects of $B$ also differ considerably, whereas the main effect estimate is obviously quite small.

It is important to note that such displays for factorial designs can be constructed only if $\sigma$ is assumed to be known or if the design is replicated. We may also note that even though this information is available from ANOVA output, it is not available in quite as convenient a form as it is with an ANOM display. In particular, software with design of experiments capability does not produce an ANOVA equivalent to an ANOM plot. Dataplot, developed primarily by Jim Filliben at the National Institute of Standards and Technology and now available as freeware, does come close, however, as it can be used to produce a matrix plot that includes both the main effects scatterplots and the two-factor interaction scatterplots (see the matrix scatterplot at http://www.itl.nist.gov/div898/handbook/pri/section5/pri594.htm). MINITAB can be used to produce a matrix scatterplot for main effects with the FFMAIN command, whereas the FFINT command will produce the matrix scatterplot for two-factor interactions.

Although Figure 4.13 shows the general form of an ANOM display for two factors, there are occasions when a different type of display would be more appropriate. For

example, if the objective is to seek the best combination of two factors to maximize or minimize the response, then plotting each combination would be useful, acting as if there was a single factor with the number of levels given by the product of the levels of the two factors. This is illustrated by Nelson, Coffin, and Copeland (2003, p. 313), with the decision limits obtained from Eq. (2.4). See also Figure 5.24 in Nelson, Wludyka, and Copeland (2005, p. 102), which clearly identifies the best factor-level combination for two 3-level factors and shows that combination to be statistically different from the other eight treatment combinations.

## 4.11 NORMAL PROBABILITY PLOT METHODS

In order for normal probability plot methods such as the one due to Lenth (1989) to be effective in assessing effect significance, it is essential that effect sparsity exists. That is, the number of significant effects should be less than half of the estimable effects. When this condition is met, the various methods that have been proposed (which were reviewed and compared by Berk and Picard (1991) and Hamada and Balakrishnan (1998)) will have approximately the same degree of success in identifying significant effects. When most effects are significant, or at least seem to be significant, all the methods can be expected to fail.

Should this be of concern? Assume that five factors are being investigated and a $2^5$ design is being used. How many of the 31 effects that could be estimated should be significant? As stated previously, Daniel (1976, p. 75) stated that seven significant effects is about average for a $2^5$ design. What I strongly suspect has never been investigated is the variability of the number of significant effects over a large number of datasets for a specified design, such as a $2^5$ design, after the bad data have been identified and removed. If the experimenters are very good at identifying factors to study in an experiment, we should not be surprised that most, if not all, of them turn out to be significant (the main effects, that is). The possible breakdown of the methods due to Lenth (1989) and others will then depend on how many interactions are significant.

Loughin and Noble (1997) proposed a permutation test as an alternative to these normal probability plot methods, and compared the performance of their test with Lenth's test over 36 datasets in the literature. Of course one weakness in making comparisons over actual datasets is that "the truth" is never known, so it isn't possible to say that one method gives the correct answer whereas the other one does not. Nevertheless, the comparison is of interest. Loughin and Noble (1997) found that their test performed the same as Lenth's test for the vast majority of datasets. There were some significant differences, however. In particular, there was a huge difference for an example in Montgomery (1991, p. 314), as the permutation test identified nine significant effects in the $2^4$ design whereas Loughin and Noble (1997) stated that the properly calibrated Lenth's test did not come close to identifying any effects as being significant. This is in reference to the calibration of Lenth's test given by Loughin (1998).

This result is disturbing, so we will examine the data. An interesting feature of this dataset, which may or may not be real data, is that this is a replicated design,

so ANOVA can be applied and the results compared with the results obtained using Lenth's test and the permutation test. Of course ANOVA also isn't "the truth," but it does provide a useful benchmark. Loughin and Noble (1997) gave the ANOVA results for this example, which showed 10 effects as being significant at the .05 level, including the replication that was treated as a factor.

This might at first appear to be a somewhat amazing result since effect sparsity obviously does not exist for this dataset, as we assume that it does in general. This seems to be much ado about nothing, however, at least for this dataset, since it seems as though this probably isn't actual data since the four-factor interaction is significant in the ANOVA table given by Loughin and Noble (1997), with a $p$-value of .0001. Such a highly significant four-factor interaction will occur only very, very rarely. Furthermore, Lenth's method is for unreplicated factorials, as the title of the paper states, whereas the dataset under discussion is a replicated factorial.

An enhancement to Lenth's method, although it was not directly presented as such, was given by Tripolski, Benjamini, and Steinberg (2005). They considered large factorial experiments, for which there is potentially a large number of real effects. They proposed the use of the false discovery rate (FDR) for such experiments and showed how to combine the control of FDR with methods for obtaining a PSE, such as Lenth's method. The FDR, which was introduced by Benjamini and Hochberg (1995) with an adaptive version given by Benjamini and Hochberg (2000), is defined as the number of inert effects falsely identified as real effects divided by the total number of effects identified as real effects.

Let $q$ denote the nominal level at which the FDR is to be controlled, $k$ denotes the number of effects that are tested, with $k_0$ denoting the number of effects that are inert. Let $p_{(1)} \le p_{(2)} \le \cdots \le p_{(k)}$ denote the $p$-values for the $k$ effects. (Of course there are no $p$-values, as such, available for an unreplicated factorial; this will be addressed shortly.) Benjamini and Hochberg (1995) showed that the FDR is controlled at $q(k_0/k) \le q$ when the $l$ largest effects (i.e., those with the smallest $p$-values) are declared active with

$$l = \max\left\{i : p_{(i)} \le \tfrac{i}{k}q\right\} \tag{4.5}$$

with no effects being declared active if the condition is not met for any value of $i$. This rule might be applied to replicated factorial designs for which $p$-values are available. For unreplicated factorial designs (i.e., designs for which Lenth's method would be used), a "$p$-value" of sorts can be computed by using the pseudo-error to compute the value of a $t$-statistic for each effect estimate, with the $p$-value being obtained from the value of that statistic. With the $p$-values thus obtained arranged in ascending order, all effects with a $p$-value smaller than the largest $p$-value that satisfies $p_{(i)} \le \frac{i}{k}q$ would result in the corresponding effects being declared active, which of course is the same condition to be met for a replicated factorial.

Tripolski et al. (2005) recommended using Lenth's procedure in conjunction with the rule for controlling the FDR, with their recommendation based on simulations with 16-run and 32-run designs. They found that this improved the power of Lenth's procedure by more than a small amount. That is, Lenth's procedure would be used

to obtain the pseudo-error, then Eq. (4.5) would be applied. That reduces the role of Lenth's procedure in the decision-making process. Although practitioners might accept the simplicity of the decision rule, they might still want to see the normal probability plot of effect estimates and perhaps a Pareto chart of the effect estimates as well. Although the decision rule might seem to be somewhat arbitrary, the effectiveness of the rule is supported by the authors' simulation results.

A method that is competitive with normal probability plot methods is a Bayes plot (Box and Meyer, 1986). The objective with this plot, which assumes normality for the real effect estimates and the effects that are just noise, is to separate those two sets of effects using a Bayesian analysis.

## 4.12 MISSING DATA IN FACTORIAL DESIGNS

Missing data are a fact of life. Statisticians and lawmakers have focused considerable attention on data missing from the decennial U.S. census. Missing data also occur when designed experiments are run (due to botched runs, etc.), but what to do about data missing from designed experiments has been much less debated.

Missing data can occur in a variety of ways. Sometimes bad parts can result when undesirable treatment combinations are used, thus resulting in no measurement being taken. This occurred in an experiment described by Lin and Chananda (2003), as the treatment combination consisting of the low level of each of four factors in a replicated $2^4$ design resulted in defective parts. Of course this results in a slight loss of orthogonality as all the pairwise correlations are $-.071$. (A general result on correlations that are created by a single missing observation is given later in this section.)

Missing data do destroy the orthogonality of a design, assuming the design was originally orthogonal. For a design with a large number of runs, such as a $2^6$, a single missing observation will not be a major problem as the design will be only slightly nonorthogonal. Furthermore, although orthogonality is a desirable property, it is not absolutely essential. Supersaturated designs (see Section 13.4.2), for example, are not orthogonal.

One obvious consequence of missing observations is that fewer effects can be estimated. For example, if a single observation is missing from a $2^3$ design, one effect less can be estimated and that would logically be the *ABC* interaction. That generally isn't a problem, but the other consequences are more serious. The missing observation causes the correlations between the factors to all be .167 in absolute value, regardless of which observation is missing.

This can be easily explained, as follows, as well as generalized. The correlation between any pair of factors will have a denominator of $n(n-2)/(n-1)$, regardless of the observation that is missing. If the missing observation is in a position that has opposite signs ($+1$ and $-1$ or the reverse) for the pair of factors for which the correlation is computed, the numerator will be of the form $1-(-1)/(n-1) = n/(n-1)$. If the signs are the same, the numerator is $-1-(1)/(n-1) = -n/(n-1)$. Thus, the correlation is either $1/(n-2)$ or $-1/(n-2)$. If $n$ is at least of moderate

size (e.g., 32), this correlation should not be of any great concern, but could result in interpretation problems when $n$ is small.

Another possible concern is the imbalance between the number of observations of a factor and the number of observations at the low level. As with the correlations between factors, this will be a problem for designs with a small number of runs. Specifically, there will be a 4–3 imbalance between high and low levels with an 8-run design with one missing observation, whereas a 16–15 imbalance should not be a major problem.

Probability plots of effect estimates can, strictly speaking, be constructed only when the variances of the effect estimates are equal and the effect estimates are independent. The variances of the effect estimates will often be the same when there is missing data, but that won't always be the case. For example, when data for the *bc* and *abc* treatment combinations are missing when a $2^3$ design is used, the variances are not all equal. Furthermore, regardless of whether or not the variances are equal, the effect estimates will not be independent because missing observations cause correlations between the factors. If those correlations are quite small, however, one might still use a probability plot.

Draper and Stoneman (1964) gave a simple method for dealing with this problem in two-level factorial and fractional factorial designs (the latter are discussed in Chapter 5). Box (1990–1991a) discussed and illustrated this method, using a missing observation from a $2^4$ design as an example. The method consists of setting the *ABCD* interaction equal to zero and solving for the missing value. This is a reasonable approach if the highest-order interaction is at least a four-factor interaction. It would not be a good idea to use that method for, say, a $2^3$ design because the *ABC* interaction could be significant.

### 4.12.1 Resulting from Bad Data

Of course missing data also result when bad data are discovered and subsequently removed before most of the data analysis is performed. Anderson and Whitcomb (1997) used an example of a $2^{5-1}$ design given by Box et al. (1978) to illustrate the deleterious effects of missing data. Box et al. (1978) first analyzed all ($2^5$) of the data points and concluded from a probability plot analysis that $B$, $D$, $E$, $BD$, and $DE$ were significant. Using the same type of analysis, but using Lenth's method, which of course had not been invented by 1978, we additionally identify $ACE$ as a significant effect, although it is clearly borderline.

The term is significant at the .05 level ($p = .03$) when it is added to the model, however. Justifying the term might be difficult, however, because using it in the model creates a nonhierarchical model since the model does not include the main effects of $A$ and $C$, nor does it include the $AC$ interaction. (See the discussion of hierarchical versus nonhierarchical models in Section 4.18, however.) This is not necessarily justification for discarding this interaction term, although there is obviously no strong statistical support for including it.

In terms of the analysis, missing data have the same effect as bad data that are subsequently discarded. It is important that not very many observations be missing,

however, as severe imbalance could result and effect estimates could be misleading. For example, if the last two observations in Yates order are missing when an unreplicated $2^3$ design is used, the correlation between the estimates of the $B$ and $C$ main effects is $-.50$. This disturbs the relationship between the effect estimates and the regression coefficients, as the 2:1 ratio for effect estimate : regression coefficient does not exist when the orthogonality property has been disturbed. Furthermore, with these two missing values the $B$ and $C$ main effect estimates would have to be computed with four observations at one level and two observations at the other level. Clearly such imbalance is highly undesirable.

### 4.12.2 Proposed Solutions

Various solutions have been proposed for the missing data problem. If a single observation is missing, one proposal is to solve for the missing value to minimize the residual sum of squares (Hicks and Turner, 1999), which of course requires a model assumption that might be difficult to make until the data have been analyzed. This approach is also suggested by Montgomery (1997, p. 189) for a randomized complete block design, in addition to recommending that the data simply be analyzed as unbalanced data, as in Searle (1987). The suggestion of Giesbrecht and Gumpertz (2004) is the same, except stated differently, as they suggest, for a randomized complete block design, imputing a value that leaves the error sum of squares unchanged. Nelson (2003) suggested a trial-and-error iterative approach that is based on using either a single interaction or pooled interactions to estimate the error variance with an unreplicated factorial. Of course this could be risky unless four-factor interactions and higher are used for this purpose.

Although these may seem to be reasonable solutions to the problem, phony data are being used with the first three suggested approaches, which of course is objectionable unless the phony data well represent the missing data.

Of course, the simplest approach is to not impute the missing value and to proceed with the data that are present, just as Lin and Chananda (2003) did.

## 4.13 INACCURATE LEVELS IN FACTORIAL DESIGNS

It isn't always possible to maintain intended levels when a design is carried out. We will consider a few possible scenarios. Assume that temperature was to be set at 350 and 400°F but the actual settings were 360 and 390°F. This may affect the outcome of the experiment, but it does not affect the analysis, regardless of whether the actual settings are known or not. If the actual settings were not known, the analysis would be performed in the usual way and conclusions would be drawn, based on the assumption that the actual temperature settings were the nominal settings. If the range of the actual settings is too small to show if the factor has an effect, the conclusion may differ from the conclusion that would have been reached if the nominal settings had been used. Furthermore, if the regression equation were converted to raw form, the equation would be assumed to be valid for 350 and 400°F, whereas the equation does not strictly apply for those values.

If the mistakes were discovered during the course of the experiment but not soon enough to correct the problem, the coded-form analysis would be unaffected, except for the fact that the inference would apply to 360 and 390°F rather than to 350 and 400°F. Somewhat similarly, the raw-form regression equation would be applicable only for 360 and 390°F.

A bigger problem occurs when the settings vary uncontrollably during the course of the experiment. Orthogonality of the design will be lost (almost certainly), and the conclusions could be erroneous, depending upon the degree of departure of the actual settings from the nominal settings. The method of analysis will certainly have to change, as will the interpretation. For example, assume that a $2^4$ design is being used but it is learned during the experiment that the level of the fourth factor cannot be held steady. Consequently, the actual number of levels of that factor might turn out to be, say, 8 or even 16, which would preclude analyzing the data as having come from the use of a factorial design. Instead, regression or a generalized linear model approach would have to be used, and the term "effect estimate" would not have any meaning for the fourth factor. If the departure from the nominal levels is not great, however, the analysis will probably agree with the factorial design analysis that would have been performed if the levels of the fourth factor had been held steady.

Donev (2004) considered the properties of experimental designs where the factor levels cannot be set precisely and found that the properties of the design with the inaccurate levels could be better or worse than the properties of the design without the inaccuracies and gave recommendations for selecting a design region so as to minimize the risk of losing observations.

## 4.14 CHECKING FOR STATISTICAL CONTROL

It was stated in Section 1.7 that processes should be in a state of statistical control when experiments are run. The extent to which this is important depends upon the likelihood of factors extraneous to the experiment (e.g., temperature) going outside of acceptable boundaries and affecting the values of the response variable.

As stated in Section 1.7, one way to check for this is to make runs at the current operating conditions at the beginning, in the middle, and at the end of an experiment. An obvious question is "Can the data from these runs be utilized in some manner, and if so, how?" Specifically, since there are no degrees of freedom for estimating the error variance when an unreplicated factorial is used, can this problem be resolved by using these data points to estimate the error variance? The answer is "yes," provided that the experimenter is willing to make the somewhat strong (and perhaps untenable) assumption that the variance at each design point is the same as the variance at the standard conditions, and provided that the replicates at the standard conditions are true replicated points.

As an aside, it should be noted that failure to maintain a state of statistical control is often what motivates a designed experiment, with the experiment designed to identify the factors that are causing the problem. An example of this is given in Hare (1988), where the objective was to conduct an experiment to identify the factors that were causing unacceptable variation.

## 4.15 BLOCKING $2^k$ DESIGNS

It may not be possible to make all the $2^k$ runs under the same conditions. When faced with this reality, a blocked $2^k$ factorial can be constructed. Because of the analogy with fractional factorials, this topic is covered in more detail in Section 5.6, but we will also make some comments about it here.

Let's consider a simple scenario in which a $2^5$ design is to be used, but only 16 runs can be made per day. Consequently, the 32 runs would have to be split up in some way, and the most obvious option would be to use two blocks with 16 units per block. Doing so would create a block component in an ANOVA table, just as is the case when a randomized block design is used. With two blocks, there would be $2 - 1 = 1$ degree of freedom for the block effect. This means that one of the factorial effects could not be estimated since a $2^5$ design is a saturated design with all of the 31 degrees of freedom available for estimating the 31 effects.

The obvious choice for the interaction to relinquish, which will be confounded (i.e., confused with the block effect), is the five-factor interaction *ABCDE*, since there is virtually no chance that a five-factor interaction will be significant, and interpreting it would be quite difficult even if it were significant.

The usual assumption when blocking is used is that there is no interaction between the blocking factor(s) and the factors under study. Assuming something to be true doesn't make it true, however, so it is a good idea to check this assumption. One way to do so is to simply treat blocks as a regular factor and see if any interactions involving blocks are significant.

### Example 4.8

An example of a $2^3$ factorial that was blocked out of necessity was given by Chapman and Roof (1999–2000). Three factors were studied and the primary objective was to find a method to prevent or reduce potential hypoglycemia rather than treat its effects after it occurs. The authors sought to accomplish this by trying to minimize the variation in glucose levels that results from exercise.

What made the study quite unusual is that the second author, a diabetic, was the experimental unit—the *only* experimental unit! A single experimental unit, although quite unusual in an experimental setting, is what often happens when student experimental design projects are performed (as in Hunter, 1977), as the experimental unit is often the student, or the student's dog, and so on. The concern, of course, when this is done is that there can be carryover effect (see Section 11.4). If there is no carryover effect, there is still the problem of the inference not being extendible beyond the person who was the experimental unit. Both of these issues are discussed later relative to this experiment.

The response variable was blood glucose level and the factors were (1) volume of juice intake before exercise (4 ounces and 8 ounces), (2) amount of exercise on a Nordic Track cross-country skier (10 min and 20 min), and (3) delay between the time of juice intake and the beginning of the exercise period (none and 20 min). The

runs were made in the morning and in the evening, and since blood glucose levels were suspected of varying between morning and evening under constant conditions, it was decided to use time of day as the blocking factor. (*Note*: Blood glucose levels are considered to be higher in the morning than during the rest of the day; see http://www.diabetic-lifestyle.com/articles/feb01_healt_1.htm.)

The data from the experiment are given below, with the runs listed in standard order rather than the randomized order that was used (but not given in the article).

| A | B | C | Block | | | |
|---|---|---|---|---|---|---|
| JUICE (oz) | EXERCISE (min) | DELAY (min) | TIME OF DAY | PRE | POST | AVG |
| 4 | 10 | 0 | pm | 78 | 65 | 71.5 |
| 8 | 10 | 0 | am | 101 | 105 | 103 |
| 4 | 20 | 0 | am | 96 | 71 | 83.5 |
| 8 | 20 | 0 | pm | 107 | 145 | 126 |
| 4 | 10 | 20 | am | 128 | 123 | 125.5 |
| 8 | 10 | 20 | pm | 112 | 147 | 129.5 |
| 4 | 20 | 20 | pm | 111 | 79 | 95 |
| 8 | 20 | 20 | am | 83 | 103 | 93 |

There are some important questions that were not addressed by Chapman and Roof (1999–2000). In particular, no information was provided regarding the length of time covered by the experiment. Perhaps all the runs were made in one day, but such information was not provided. The authors used the average blood glucose level for each treatment combination, with the "PRE" reading being before exercise and the "POST" reading being after exercise. Thus, the reading *before* exercise has a weight of 0.5 in determining the (average) response value that is used *after* exercise.

It would have been very helpful to know whether or not all of the runs were performed in one day and to know the sequence of the runs because carryover effect might be a real problem.

Curiously, although the authors stated that the experiment was blocked, it was not analyzed as such when using average glucose level as the response variable. We can observe, after writing down the treatment combinations that are in each block, that the *ABC* interaction is confounded with blocks. In their analysis, Chapman and Roof (1999–2000) created an error term with two df using Blocks(*ABC*) and the *AB* interaction. Looking at the data, the latter seems to be a reasonable assumption but when we use blocking we would like to see evidence that the blocking was necessary. The authors did not use a normal probability plot. This could be done although such plots applied to small designs may not give reliable results. Indeed, the plot constructed using MINITAB shows no significant effects, whereas the ANOVA table given by the authors shows that all of the estimable effects except Exercise (factor $B$) is significant at the .05 level. As stated in Section 4.11, normal probability plot methods fail when there is no effect sparsity. Here the authors' ANOVA table shows that four of the seven estimable effects are significant at $\alpha = .05$. That is *not* effect sparsity!

Their regression model was constructed using uncoded units, so the results in terms of $p$-values for effects do not agree with the $p$-values in their ANOVA table. Doing the regression analysis on the uncoded units induces correlations between main effects and interaction effects and essentially renders meaningless the regression analysis $p$-values. All analyses should be performed with coded units when interaction terms are used, as was illustrated in Example 4.1 (see also Ryan, 2000, Section 13.11).

Since hypoglycemia can result from exercise, a major goal was to minimize the variation in blood glucose levels due to exercise. Accordingly, the authors did the following. They computed the standard deviation (of two numbers), $s$, at each treatment combination, used the variance-stabilizing transformation $\ln(s)$ because of the large variation in $s$, and then used $\ln(s)$ as the response variable.

The ANOVA table for $\ln(s)$ showed that Exercise and Time of Day were significant at the .05 level. Again, however, there was a discrepancy between the regression analysis and the ANOVA results since the former was performed on the uncoded units. The authors concluded that the best strategy was to drink 8 ounces of juice immediately before exercise (i.e., no delay) and exercise for 10 minutes. This treatment combination does have the smallest value of $\ln(s)$.

This conclusion is highly sensitive to the assumption of no carryover effects, however, and then there is the problem of not being able to apply these results to other diabetics. The second author recognized this with the statement, "Co-author Roof suspects that the optimum conditions achieved by this experiment might not be universally applied at all times nor to all diabetics."

Blocking factorial designs is also discussed in Section 5.8.

## 4.16 THE ROLE OF EXPECTED MEAN SQUARES IN EXPERIMENTAL DESIGN

This section is placed near the end of the chapter because the topic is one that many experimenters would prefer to skip. The way that expected mean squares determine the manner in which hypothesis testing is performed in replicated designs is practically hidden in computer output. The user of experimental design software indicates whether each factor is fixed or random, with fixed generally being the default, and the software performs the appropriate analysis for the indicated classification of factors.

We will look at a simple example to see what effect that classification can have. Assume that we have a $2^2$ design with three replications and the data are as follows.

| Treatment Combination | Data |
|---|---|
| (1) | 12.1, 11.8, 12.6 |
| a | 10.7, 11.1, 11.0 |
| b | 12.0, 12.0, 11.7 |
| ab | 11.9, 12.3, 12.6 |

We will first assume that both factors are fixed. With this assumption, we obtain the output given below.

```
          ANOVA:     Y versus  A, B

    Factor    Type    Levels     Values
      A       fixed     2        -1, 1
      B       fixed     2        -1, 1

              Analysis of Variance for Y

   Source      DF       SS        MS         F         P
     A          1   0.5633    0.5633     6.26     0.037
     B          1   0.8533    0.8533     9.48     0.015
     A*B        1   1.9200    1.9200    21.33     0.002
     Error      8   0.7200    0.0900
     Total     11   4.0567

S = 0.3    R-Sq = 82.25%     R-Sq (adj) = 75.60%
```

Notice that all three effects are significant at the $\alpha = .05$ level.

Now we will assume that both factors are random. This results in the output given below.

```
           ANOVA:  Y versus   C1,  C2

         Factor    Type    Levels     Values
           A      random     2        -1,  1
           B      random     2        -1,  1

                  Analysis of Variance for Y

    Source     DF       SS        MS        F         P
      A         1    0.5633    0.5633    0.29     0.684
      B         1    0.8533    0.8533    0.44     0.626
      A*B       1    1.9200    1.9200   21.33     0.002
      Error     8    0.7200    0.0900
      Total    11    4.0567

  S = 0.3     R-Sq =  82.25%     R-Sq(adj) =  75.60%
```

Now the $A$ and $B$ effects are not even close to being significant, as the $p$-value for testing each effect is quite large. The data have not changed, however, and if the model coefficients had been part of this output, it would be apparent that they do not change. So the only changes are in the $F$-statistics and $p$-values.

The difference is that, assuming replication, in the fixed factor case all effects are tested against the error term, meaning that each of the three mean squares is divided by the mean square error in producing the $F$-statistics, whereas in the random factor case

the main effects are tested against the interaction term and the interaction term is tested against the error. The reason for this is given in Appendix C at the end of the chapter. Since the interaction term in this example is large, the results thus differ greatly.

The user of software for experimental design does not have to be concerned with how expected mean squares are obtained because these are produced with software. For the interested reader, however, Lenth (2001) gave a simple method for determining expected mean squares. For each effect in a restricted model (e.g., a restriction that components of an interaction sum to zero over the levels of a fixed factor, with the other factor being random), the terms that are part of the expected mean square for a given effect are terms that contain the effect in question and that involve no interactions with other fixed factors. The coefficient for a variance component is the number of observations at each distinct level of that component. (Of course, $\sigma^2$ is part of every expected mean square.)

For example, assume that there are three random factors, $A$, $B$, and $C$, with each factor having two levels, and two replicates are used. The model with all possible interaction terms is fit. The expected mean square for the $AB$ interaction is then $\sigma^2 + 2\sigma^2_{ABC} + 4\sigma^2_{AB}$. In accordance with Lenth's (2001) method, $\sigma^2_{ABC}$ is one of the terms because it contains $AB$ and the coefficient is 2 because the design is replicated. The coefficient of the $\sigma^2_{AB}$ term is 4 because there are two levels of $C$ at each $AB$ combination and there are also two observations because the design is replicated. Thus, $2 \times 2 = 4$.

For the unrestricted model, the expected mean square contains all terms that contain the effect and at least one other random factor.

We may note that although expected mean squares can be produced with MINITAB, this must be done with the ANOVA command, not the FFACT command.

## 4.17 HYPOTHESIS TESTS WITH ONLY RANDOM FACTORS IN $2^k$ DESIGNS? AVOID THEM!

It may be disturbing that the classification of factors as fixed or random can have such a dramatic effect on the results. We will show why hypothesis tests with random factors in $2^k$ designs should be avoided. The reason for this is that an $F$-test should never be performed when the denominator degrees of freedom is 1, as will be the case whenever any effect, main effect or interaction, is tested against an interaction effect for two-level designs, as in the example above with both factors random.

Since $F_{1,1,.05} = 161.44$, an effect estimate would have to be extremely large in order to have significance declared at the .05 level. This can be seen as follows. Since the mean squares are the same as the sum of squares when the degrees of freedom is 1, this means that the numerator sum of squares divided by the denominator sum of squares must exceed 161.44.

We can express this relationship in terms of effect estimates. This can be done by using the fact that for a replicated (or unreplicated) $2^k$ design, the relationship is

$$\text{Effect estimate} = \sqrt{\frac{\text{SS}_{\text{effect}}}{r(2^{k-2})}}$$

Thus, for an unreplicated $2^2$ design, the relationship is effect estimate $= \sqrt{\text{SS}_{\text{effect}}}$. (This last result is the general form of the result in Exercise 4.17 that the reader is asked to derive.) Thus, the ratio of the effect estimates would have to be at least 12.71:1 in order for significance to be declared.

Consider the modified data given below, with the numbers in the first two rows the same as previously given but the numbers in the last two rows increased so that the average of those numbers is almost twice the average of the numbers in the first two rows.

| Treatment Combination | Data |
|---|---|
| (1) | 12.1, 11.8, 12.6 |
| a | 10.7, 11.1, 11.0 |
| b | 20.0, 20.0, 19.5 |
| ab | 19.8, 20.5, 21.0 |

Since the difference of these two averages estimates the $B$ effect, there is obviously such an effect, but the analysis using ANOVA results in a $p$-value of .068, so the ANOVA result doesn't make any sense, and this is because of the low power to detect significant differences because of the use of an $F$-test with one degree of freedom in the denominator of the $F$-statistic.

It is worth noting that there is no "Daniel rule" violation with the modified data since the ratio of the effect estimates is almost 10:1. Thus, it is strictly a problem with the denominator degrees of freedom; a problem that exists regardless of the number of replications since the latter does not affect the degrees of freedom for the effects, which will be 1 for each effect, for every two-level design.

## 4.18 HIERARCHICAL VERSUS NONHIERARCHICAL MODELS

A *hierarchical model* is one in which all factors appear as main effects in a model when those factors appear in higher-order terms. For example, a model with a three-factor interaction would contain two-factor interaction terms in the factors that comprise the three-factor interaction, and a model that contains the $AB$ interaction would also have to contain a main effect term in $A$ and one in $B$. A model that does not meet this requirement is a nonhierarchical model.

It was stated in Section 4.12.1 that a nonhierarchical model could be difficult to justify. However, Montgomery, Myers, Carter, and Vining (2005) take a different position on the matter, stating that "Most industrial experiments involve systems driven by physical or chemical mechanisms. These underlying mechanisms, while unknown, are unlikely to be hierarchical." Furthermore, they point out that it is not uncommon to encounter two-factor interactions without both main effects being significant when one factor is quantitative and the other factor is qualitative, and especially when both factors are qualitative. At the other extreme, there are undoubtedly many users who will be influenced by software packages that give warning messages about nonhierarchical models and are also influenced by articles such as Franks (1998).

Recall the discussion in Section 4.2.1 that a signal to use a nonhierarchical model is also a signal to do a conditional effects analysis. In other words, we should not

simply settle for a model that does not contain a main effect but does contain an interaction term in that factor, but rather should do a conditional effects analysis.

## 4.19 HARD-TO-CHANGE FACTORS

Assume that a factor can be varied, with great difficulty, in an experimental setup (such as a pilot plant), although it cannot be freely varied during normal operating conditions. Assume further that it is imperative that the number of changes of the hard-to-change factor be minimized, each factor has two levels, and the design is to have factorial structure.

Again assume that we have two factors. We can minimize the number of level changes of one factor simply by keeping the level constant in pairs of consecutive runs. That is, either the high level is used on consecutive runs and then the low level on the next two runs, or the reverse. This means that we have *restricted randomization*, however, as there are six possible run orders without any restrictions, but with the restriction there are only two possible run orders (high, high, low, low, or the reverse). Thus, there is restricted randomization in regard to the run orders. Restricted randomization increases the likelihood that extraneous factors (i.e., not included in the design) could affect the conclusions that are drawn from the analysis. Furthermore, this will also cause bias in the statistics that are used to assess significance, as shown by Ganju and Lucas (1997). Although restricted randomization and the problems caused by it might seem to have been seriously considered only since the late 1990s, the issue was discussed in the literature much earlier, dating at least from Youden (1964, 1972). (See also Bailey (1985, 1987), White and Welch (1981), Monod and Bailey (1993), and Bowman (2000)).

The need to address the issue of hard-to-change factors dates at least from Joiner and Campbell (1976). Design plans in the presence of such factors have been given more recently by Webb and Lucas (2004). See also Ju and Lucas (2002), who considered designs when there is one easy-to-change factor or one hard-to-change factor.

Although hard-to-change factors have not been discussed extensively in the journal literature or in textbooks, it seems safe to assume that such factors occur very frequently in practice. Some examples of hard-to-change factors described in the literature are as follows. Czitrom (2003) described an experiment originally given in Czitrom, Mohammadi, Flemming, and Dyas (1998) for which one of the three factors used in the experiment was a heat plug in a furnace. Time-consuming hardware changes were necessary whenever the plug was removed and subsequently put back in. Consequently, the first four experimental runs were made with the plug in and the next four were made with the plug out, which created a split-plot structure relative to that factor. (Split-plot and related designs are discussed in Chapter 9, and Example 9.1 illustrates the effect of ignoring the randomization restriction relative to the results using the proper analysis.)

Other examples of hard-to-change (or impossible to change) factors in experiments described in the literature include an experiment described by Inman, Ledolter, Lenth, and Niemi (1992), as it was not possible to change temperature between individual

runs, so temperature was used as a blocking variable. This is an example of an impossible-to-change factor creating an "unnatural" blocking factor, because if there were a time or day effect it would be confounded with the temperature effect. This required that care be exercised to ensure that laboratory conditions did not vary from day to day. The authors stated, "In the analysis, we ignored the restriction on randomization with respect to temperature." Undoubtedly this was also done for countless unpublished experiments during that era, but in light of recent research, analysts should now assess the possible consequences in terms of bias in ignoring hard-to-change or impossible-to-change factors.

Eibl, Kess, and Pukelsheim (1992) described a sequence of experiments with the response variable being paint coat thickness. All factors could be easily changed except paint viscosity. Consequently, the levels of paint viscosity were kept constant as long as possible in one of the three experiments. The authors recognized that this probably caused the experimental error to be underestimated, but they decided not to adjust for the lack of randomization when they noticed that the error standard deviations were practically the same in each experiment.

As a final example of a hard-to-change factor, Prat and Tort (1989) described an experiment conducted in a pet food manufacturing company that had four factors and one of them, compression zone in die, was hard to change. This experiment is described in considerable detail in Section 5.9.1.2.

Simpson, Kowalski, and Landman (2004) found from their consultation with engineers in the aerospace and defense industries that there is a strong need for something less than complete randomization. They show, using MINITAB, how to analyze data from a design that has four factors with two being hard to change. Similarly, there is a good tutorial on using MINITAB with one hard-to-change factor at www.minitab.com/support/docs/OneHardtoChangeFactor.pdf.

If a cost can be assigned to the changing of levels of hard-to-change factors, a design with minimum cost might be sought, although such a design might of course be viewed as undesirable in other ways. Taihrt (1971) and Taihrt and Weeks (1970) gave run orders for two-level factorial designs that require a change in only one factor between successive runs. More recently, Joseph (2000) discussed the construction of minimum cost designs for hard-to-change factors, the objective being to determine the experimental sequence that minimizes the cost of adjusting the factor levels during experimentation. Of course this requires that the cost of adjusting each hard-to-change factor is obtainable, or at least that factors can be ranked in terms of difficulty of adjustment. Joseph (2000) gave an example for which it was difficult to obtain the adjustment costs from speaking to the operators, so it was necessary to settle for a rank order of the factors in terms of the difficulty of adjustment. Goos and Vandebroek (2004) showed that factorial designs with hard-to-change factors run in a split-plot arrangement can be superior, in terms of the design efficiencies that they use, to factorial designs run without a split-plot arrangement.

Using an experimental run sequence determined from cost considerations is of course a form of restricted randomization and thus makes the use of hypothesis tests using ANOVA rather shaky, although they still might be used, recognizing their limitations.

### 4.19.1 Software for Designs with Hard-to-Change Factors

Since the consideration of hard-to-change factors is of somewhat recent origin, the way in which the designs are generated and analyzed with popular statistical software can best be described as a workaround with some software. For example, although the designs can be constructed and analyzed with MINITAB, doing so is rather involved (e.g., see http://www.minitab.com/support/docs/OneHardtoChangeFactor.pdf for information on how to proceed with one hard-to-change factor).

## 4.20 FACTORS NOT RESET

In addition to experiments involving hard-to-change factors, there are many experiments for which at least one factor is not reset when the same level is to be used for the next run. Resetting a factor that is to have the same level for the next run (e.g., the same temperature) would seem to be a waste of time and money. Therefore, it is not surprising that a sizable fraction of industrial experiments have at least one factor that is not reset and in many experiments none of the factors are reset (Lucas, 1999).

There would seem to be no harm in not resetting a factor if the same level is to be used in the next run. For example, if temperature is one of the factors, why lower it to some level only to raise it back to the level at which it was set? This would seem to be rather impractical, but the argument in favor of resetting is that not resetting the temperature could induce a correlation between consecutive values of the response variable, which would invalidate the outcome of statistical tests. The question that must be addressed, however, is whether or not the bias and prediction variance inflation (see Webb, Lucas, and Borkowski, 2004) that results from an incorrect analysis is more than offset by the cost of resetting factors. See Webb et al. (2004) for a detailed discussion of these issues.

## 4.21 DETECTING DISPERSION EFFECTS

The detection of dispersion effects can be at least as important as the detection of location effects. Designs for detecting both are discussed in Chapter 8; in this chapter we simply wish to emphasize the importance of identifying dispersion effects and also to point out some problems in attempting to do so. In particular, Schoen (2004) explains how the detection of dispersion effects can be hampered by unidentified location effects; McGrath and Lin (2001) showed that the presence of dispersion effects may make it difficult to detect other dispersion effects and to estimate those effects. Bisgaard and Fuller (1995–1996) showed how to identify factors that affect dispersion by using a $2^4$ design in conjunction with the logarithm of the sample variance as the response variable. Fuller and Bisgaard (1996) compared methods for identifying dispersion effects using two-level factorial designs.

## 4.22 SOFTWARE

Software for factorial designs is quite plentiful and all of the leading statistical software packages have this capability. Design-Expert is unusual, however, in that commentary on analyses is provided, as long as the "Annotated ANOVA" option is selected, and there is also a "Tips" button that can be used to obtain general advice on two-level factorial designs. For example, when the cement dataset from Daniel (1976) is analyzed with the software, the following direction and advice is shown after the saturated model is fit.

> Proceed to Diagnostic Plots (the next icon in progression). Be sure to look at the:
>
> **1)** Normal probability plot of the studentized residuals to check for normality of residuals.
> **2)** Studentized residuals versus predicted values to check for constant error.
> **3)** Externally Studentized Residuals to look for outliers, i.e., influential values.
> **4)** Box-Cox plot for power transformations.
>
> If all the model statistics and diagnostic plots are OK, finish up with the Model Graphs icon.

This has the flavor of expert systems software for design of experiments, which, although developed and used internally by certain companies for many years, has not been a part of the best known, general purpose statistical software.

Another software package that is exclusively for design of experiments, D. o. E Fusion Pro, also has some features of expert systems software. The software is not well known but received a rating of "Excellent +" in the comparison study of Reece (2003). One feature of this software that might raise some eyebrows, however, is that one of the four modules is Data Mining/Analysis. We generally don't think of the analysis of data from a designed experiment as data mining since the latter term is usually reserved for the analysis of large datasets.

D. o. E. Fusion Pro users can follow one of two paths, the "Design Navigation Wizard" or the "Design Menu Wizard," with the latter used when a user-specified design is to be constructed, and the former selected when the user does not know the most appropriate design to use. One interesting feature is that the number of design points to be replicated can be selected, but the points cannot be specified unless centerpoint replication is specified or all the points are to be replicated.

The comparison study of Reece (2003) was mentioned in Section 2.3. It is also mentioned here since the software packages that were included in the study were rated on factorial and fractional factorial designs, with the latter covered in Chapter 5.

## 4.23 SUMMARY

Although they have been used extensively, full factorial designs should be used only under certain conditions. When there is a small number of factors, such as 3, the design should be replicated so as to have a reasonable chance of detecting something

less than large effects. Designs with at least 16 runs are more useful (see Box, 1992). When there is at least a moderate number of factors, a full factorial design may be too expensive to run, depending on the field of application. (An exception is computer experiments, which are generally inexpensive to conduct.)

This is not to suggest that one-factor-at-a-time designs should be used instead of factorial designs, as the former will generally be inferior to the latter. There are conditions under which well-constructed, one-factor-at-a-time designs can be useful, but such conditions do not occur very often. An extended discussion of these designs is deferred to Section 13.1.

Unless interaction effects are quite small, a conditional effects analysis should be performed. Unfortunately, such an analysis could become complicated when there is at least a moderate number of factors and more than a few moderate-to-large interactions. A step-by-step guideline for handling such scenarios has not been developed, however.

It is important to properly identify random factors and fixed factors as the analysis of factorial designs is determined by the declaration of fixed and random factors, as was seen in Section 4.16.

The importance of having processes in statistical control, or at least approximate statistical control, cannot be overemphasized, as spurious results can be obtained when processes are badly out of control.

Analysis of Means (ANOM) is a statistical tool that should undoubtedly be used more often. It has the advantage over ANOVA of the plotted points being in the original units, and there is also the advantage of the procedure being inherently graphical.

## APPENDIX A

### Derivation of Conditional Main Effects

We wish to show that $\delta_C = \delta \pm \delta_{int}$, with $\delta$ denoting a main effect, $\delta_{int}$ a two-factor interaction that contains the factor represented by $\delta$ plus another factor, and $\delta_C$ the conditional effects that result from splitting the data on the other factor in the two-factor interaction represented by $\delta_{int}$.

Let $\delta^+$ denote the conditional main effect of the factor of interest at the high level of the other factor (i.e., the factor that the data are being split on), and similarly define $\delta^-$ for the low level of the other factor. Using the result $\delta_{int} = (\delta^+ - \delta^-)/2$, which follows from the definition of a two-factor interaction (see, e.g., Wu and Hamada (2000, p. 106, Eq. 3.10) and/or consider how the interaction effect would be computed from an interaction plot), and $\delta = (\delta^+ + \delta^-)/2$, which follows from the definition of an effect with a plus sign being attached to the high level of the factor and a minus sign to the low level, we obtain $\delta_C = \delta^+$ and $\delta_C = \delta^-$ when these substitutions are made in the postulated expression. This establishes the result since $\delta^+$ and $\delta^-$ designate the two conditional effects.

The ramification of this result is that the data should be split on the factor whose interaction with the factor of interest is the largest, as this will give the greatest difference in the conditional effects for that factor and thus give the maximum insight.

## APPENDIX B

### Relationship between Effect Estimates and Regression Coefficients

It was stated in Section 4.1 that the OLS regression coefficients are half the effect estimates. This can be explained as follows. Define a matrix $\mathbf{X}$ such that the first column is a column of 1s (which provides for estimation of $b_0$, the constant) and the other columns contain the coded values (i.e., $+1$ and $-1$) of the effects to be estimated. Since all but the first column sum to zero and pairwise dot products of all columns are zero, $\mathbf{X}'\mathbf{X}$ is a diagonal matrix with $n$ as each element on the main diagonal. Thus, $(\mathbf{X}'\mathbf{X})^{-1}$ is a diagonal matrix with each diagonal element $1/n$. The regression coefficients are obtained from $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ with $\mathbf{Y}$ a vector that contains the response values. The vector obtained from $\mathbf{X}'\mathbf{Y}$ then has the sum of the response values as the first element and every other element of the form $\sum_{j=1}^{n}(X_{ij}Y_j)/n$ for each effect $X_i$ that is being estimated. Since each $X_{ij}$ is either $+1$ and $-1$, each sum is thus of the form $(\sum y_{(+)} - \sum y_{(-)})/n$, with $y_{(+)}$ denoting a value of $Y$ for which $X_{ij}$ is positive, and similarly $y_{(-)}$ denotes a value of $Y$ for which $X_{ij}$ is negative. Note that the divisor is $n$ instead of $n/2$, with the latter being the number of terms in each of these last two sums. Thus, this simplifies to $\frac{1}{2}(\overline{y}_{(+)} - \overline{y}_{(-)}) = \frac{1}{2}$ (effect estimate).

## APPENDIX C

### Precision of the Effect Estimates

Since the regression coefficients are obtained from $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, it follows that the variance of each one is $\text{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\text{Var}(\mathbf{Y})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. As noted in the previous section, $(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}$ is a diagonal matrix with each diagonal element $1/n$. Thus, each regression coefficient is estimated with a variance of $\sigma^2/n$, so that the standard deviation is $\sigma/\sqrt{n}$. Since an effect estimate is twice the corresponding regression coefficient, it follows that the standard deviation of an effect estimate is $2\sigma/\sqrt{n}$. (This result could also be obtained without using matrices, by writing each effect estimate as the appropriate function of the $y's$, converting that to the appropriate function of the error terms and then computing the variance of that expression (see, e.g., Bisgaard and de Pinho, 2004).

Of course $\sigma$ must be estimated, which would lead to the standard error of an effect estimate expression of $2\widehat{\sigma}/\sqrt{n}$. Thus, each effect is estimated with the same precision with a $2^k$ design and of course the precision increases with larger $n$, which might result either from replication, or from increasing the value of $k$.

## APPENDIX D

### Expected Mean Squares for the Replicated $2^2$ Design

The determination of expected mean squares in general is a laborious task and there is no need for an experimenter to do so as these can be produced with software,

including MINITAB. Design-Expert, however, does not give expected mean squares. Nevertheless, experimenters who wish to "see for themselves" how the expected mean squares are obtained could proceed in one of two ways: either derive the results directly (not recommended routinely although we will do so here for illustration), or use general rules for obtaining the results, such as those given by Dean and Voss (1999, p. 616) or Lenth (2001).

(The derivations sketched out and given in the rest of this section require a knowledge of expected value, variance, and covariance. Readers without this knowledge can obtain it from any introductory book on mathematical statistics or statistical theory, such as Casella and Berger, 2001.)

When derived from scratch rather than by using general rules, expected mean squares are first obtained by computing the corresponding expected sum of squares, since the mean square and sum of squares of course differs by only a constant. We consider the model

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijk} \quad i = 1, 2 \;\; j = 1, 2 \;\; k = 1, 2, \ldots, r \qquad \text{(D.1)}$$

corresponding to a $2^2$ design with $r$ replicates, with, as before, "1" denoting the low level of each factor and "2" denoting the high level, with both factors assumed to be random.

We will use the result stated in Exercise 4.17, $\text{SS}_A = r(\overline{A}_2 - \overline{A}_1)^2$, for a design with $r$ replicates, which can be easily derived, as the reader is asked to do in that exercise. Here $\overline{A}_2$ is the average of the $Y$-values at the high level of factor $A$.

For an unreplicated $2^2$ design, this would be the average of two numbers, whereas for the replicated design it is the average of $2r$ numbers.

Assume $r = 2$, for the sake of illustration. Then there will be four observations at the high level of $A$ and four observations at the low level of $A$. From Eq. (D.1), the former are $Y_{211}, Y_{212}, Y_{221}$, and $Y_{222}$, and the latter are $Y_{111}, Y_{112}, Y_{121}$, and $Y_{122}$. The $E(\text{SS}_A)$ is then $E[2(\overline{A}_2 - \overline{A}_1)]^2 = (1/8)E(Y_{211} + Y_{212} + Y_{221} + Y_{222} - Y_{111} - Y_{112} - Y_{121} - Y_{122})^2$. By definition, $E(W^2) = \text{Var}(W) + [E(W)]^2$, for any random variable $W$. Since $E(Y_{ijk}) = \mu$ for all $i, j, k$, which follows from the assumption that the effect terms in Eq. (D.1) are assumed to have a mean of zero, $E(Y_{211} + Y_{212} + Y_{221} + Y_{222} - Y_{111} - Y_{112} - Y_{121} - Y_{122}) = 0$. Thus, the desired expected value is given by $(1/8)[\text{Var}(Y_{211} + Y_{212} + Y_{221} + Y_{222} - Y_{111} - Y_{112} - Y_{121} - Y_{122})]$.

Obtaining the variance expression without the use of rules or dot notation (which can be confusing) is both tedious and unorthodox, but that is the only way to literally "see" how the final result is obtained. The variance of each $Y_{ijk}$ is the same, namely, $\sigma^2 + \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2$, which follows directly from Eq. (D.1) and the independence of the effects listed in the model. (Note that the interaction term is random because both factors are random, thus there is a variance component for the interaction term.) There are 28 covariance terms since there are eight $Y$-values and $\binom{8}{2} = 28$, and eight of these are zero. The 20 nonzero covariance terms are determined by what is common to each pair of $Y$-values. For example, $\text{Cov}(Y_{211}, Y_{221}) = \sigma_A^2$ because $A$ is at the high level in

each of the two expressions so that $A_2$ is common to each $Y$-expression. Similarly, $\text{Cov}(Y_{211}, Y_{212}) = \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2$ since both $A$ and $B$ are at the same level in each $Y$-value, and hence $AB$ has the same combination of values.

Combining the 20 nonzero covariance expressions with the variance part, which is $8(\sigma^2 + \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2)$ and then multiplying by the constant $(1/8)$, given two paragraphs earlier, produces the final result, which is $\sigma^2 + 4\sigma_A^2 + 2\sigma_{AB}^2$, which the reader is asked to verify in Exercise 4.22.

The other mean squares could be similarly obtained, and it can be shown that $E(\text{MS}_{AB}) = \sigma^2 + 2\sigma_{AB}^2$. The $A$ effect would thus be tested against the $AB$ effect because the two mean squares differ only by $4\sigma_A^2$, and $\sigma_A^2 = 0$ is the hypothesis that would be tested. Since $E(\text{MS}_{\text{Error}}) = \sigma^2$, the $AB$ effect would thus be tested against the error term.

## APPENDIX E

### Expected Mean Squares, in General

The expected mean square for an effect depends upon the classification of factors as fixed or random. This applies not only to any factor involved in an effect whose expected mean square is being computed, such as the $A$ effect in Appendix D, but it also applies to the other factors in the design and the effects in the model.

To see this, recall from the preceding section that when we start from scratch, we see that expected mean squares depend upon $E(Y)$, which in turn depends partly on the model and partly on the classification of factors. When we have a replicated $2^2$ design, with both factors being random, we have four variance components: $\sigma^2$, $\sigma_A^2$, $\sigma_B^2$, and $\sigma_{AB}^2$. If we had an unreplicated $2^2$ design, then we could only have three variance components because we could not separate $\sigma^2$ from $\sigma_{AB}^2$.

If both factors are fixed, then we cannot have a variance component associated with those factors because the levels were not chosen at random from a population of levels. To illustrate the difference, let's consider a simple example. Assume that there are two students who perform the following experiment. One student randomly generates two integers in the interval 1–10 and for each random integer so generated, generates a value for $Y$ using a simple linear regression model with known parameter values *and without an error term*. The other student picks two integers in that interval, but not at random, and uses the same regression model to generate the $Y$-values. If the second student were to do this again, using the same two integers, then the same pair of $Y$-values would be generated. Thus, there would be no variance component; there would simply be the difference between two $Y$-values, and that difference is caused by the different results when each of the two integers is used in the model. If the first student randomly generates another pair of integers and solves for the corresponding $Y$-values through the model, the new pair of $Y$-values will almost certainly be different from the first pair since the second pair of integers will almost certainly be different.

Hence, the first student would generate a *variance* of $Y$ through repeated experimentation, whereas the second student would be forever generating the same two values, with the *effect* of the two integers represented by the difference of the $Y$-values.

Now think about doing this *with* an error term used in the simple linear regression model, with the errors randomly generated. The second student will now be generating different pairs of $Y$-values, due only to the random errors in the regression model, whereas the first student will be generating different $Y$-values that are due to two variance components: the variance due to the variability in the factor levels and the second due to the variance of the error term.

Thus, there will be an "effect" associated with a fixed factor, but there will not be a variance component. Consider again the replicated $2^2$ design, but this time we will assume that both factors are fixed. We can proceed with the same general approach that was used in the random effects case, for which $E(Y_{211} + Y_{212} + Y_{221} + Y_{222} - Y_{111} - Y_{112} - Y_{121} - Y_{122}) = 0$ and the variance required tedious but straightforward calculations. We have the reverse in the fixed effects case in that the variance is a simple expression but the expected value is nonzero and requires some work.

Since the only random component in Eq. (D.1) in the fixed effects case is the error term, all of the covariances in the linear combination of $Y$-values are zero, so the variance of the linear combination is simply $8\sigma^2$. Unlike the random effects case, the expected value of the linear combination is not zero, and this is because the fixed effects translate into constants and there are different constants involved. For example, $E(Y_{211}) = \mu + A_2 + B_1 + (AB)_{21}$, whereas $E(Y_{221}) = \mu + A_2 + B_2 + (AB)_{22}$. Only when the first two subscripts are the same are the expected values the same. When we examine the linear combination for estimating the $A$ effect, $Y_{211} + Y_{212} + Y_{221} + Y_{222} - Y_{111} - Y_{112} - Y_{121} - Y_{122}$, we note that the levels of the other factor add to zero when the appropriate sign of each term is used. All of the $(AB)_{ij}$ are different, but when the side condition $\sum_{i=1}^{2}(AB)_{ij} = \sum_{j=1}^{2}(AB)_{ij} = 0$ is imposed, we see that the interaction term also drops out of the computation, as of course does $\mu$ since the coefficients add to zero. So all that is left from the expected value expression is $(4A_1 - 4A_2)^2$. The correct answer for two replicates is $E(\mathrm{SS}_A) = \sigma^2 + 4\sum_{i=1}^{2} A_i^2$, so it would seem as though we are about to obtain an incorrect answer because, for one thing, $(4A_1 - 4A_2)^2$ will have a cross-product term. We recall from Section 2.1.1, however, that we also have the side condition $\sum_{i=1}^{2} A_i = 0$. From this it follows that $A_1 = -A_2$. We may thus write our result as $(1/8)(8\sigma^2 + 64A_2^2) = \sigma^2 + 8A^2$, which can be seen to be equivalent to the conventional way of writing the result.

In the mixed effects case with one random factor and one fixed factor, two models have been proposed: the restricted model and the unrestricted model. The difference between the two models is that the interaction term for the two-factor design is treated differently in each case. Assume that factor $A$ is random and factor $B$ is fixed. If an interaction term includes at least one random factor, then the interaction term, $(AB)_{ij}$, is considered to be random. More specifically, $(AB)_{ij} \sim N(0, \sigma_{AB}^2)$. This is the *unrestricted* model.

The other approach is to impose the restriction that the sum of the interaction effects over the fixed factor is zero, thus treating part of the interaction as being fixed. For this example that means the restriction $\sum_{j=1}^{2}(AB)_{ij} = 0$. If this approach is adopted, then $E(\mathrm{MS}_B)$ will not contain a term in $\sigma_{AB}^2$, so that $B$ would be tested for significance using the statistic $\mathrm{MS}_B/\mathrm{MS}_\mathrm{E}$ instead of $\mathrm{MS}_B/\mathrm{MS}_{AB}$. Thus, the approach

that is adopted will affect the results in a way similar to the effect of classifying a factor as fixed or random.

Opinions vary on the approach that should be adopted. Dean and Voss (1999, p. 629) prefer the unrestricted model; Montgomery (1997, p. 480) states, "Most statisticians tend to prefer the restricted model, and it is the most widely encountered in the literature." Kuehl (2000, p. 243) points out that the restricted model cannot be used with unbalanced data and concludes that the choice between the two models depends on the scenario (see also Cobb, 1998, Section 13.3).

There is also a difference in the approaches adopted by the major statistical software. For example, MINITAB fits the unrestricted model by default but will fit the restricted model if specified by the user. Both JMP and SAS use only the unrestricted approach and Design-Expert similarly does not give a user the choice between a restricted and unrestricted model analysis.

All things considered, including software, the unrestricted model seems preferable since regarding part of an interaction as fixed is somewhat impractical.

It is important to recognize that all of these expected value computations are based on the assumption that the fitted model is the correct model, which is almost certainly not going to be true. Recall that $E(Y_{ijk})$ underlies the expected mean squares computations, and this expected value will not, in general, be $\mu$ when the wrong model is used. Some model must be assumed if a parametric approach is used, however, and we would hope that the postulated model is a good representation of the true, unknown model.

## REFERENCES

Anderson, M. and P. Whitcomb (1997). How to analyze two-level factorials with missing data. Slide presentation. Stat-Ease, Inc.

Bailey, R. A. (1985). Restricted randomization versus blocking. *International Statistical Review*, **53**, 171–182.

Bailey, R. A. (1987). Restricted randomization: A practical example. *Journal of the American Statistical Association*, **82**, 712–719.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300.

Benjamini, Y. and Y. Hochberg (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, **25**, 60–83.

Berk, K. N. and R. R. Picard (1991). Significance tests for saturated orthogonal arrays. *Journal of Quality Technology*, **23**, 79–89.

Bingham, D. (2001). Discussion of "Factor screening and response surface exploration" by Cheng and Wu. *Statistica Sinica*, **11**, 580–583.

Bisgaard, S. and A. L. S. de Pinho (2004). Quality Quandaries: The error structure of split-plot experiments. *Quality Engineering*, **16**(4), 671–675.

Bisgaard, S. and H. T. Fuller (1995–1996). Reducing variation with two-level factorial experiments. *Quality Engineering*, **8**(2), 373–377. (This article is available as Report No. 127,

Center for Quality and Productivity Improvement, University of Wisconsin-Madison and can be downloaded at http://www.engr.wisc.edu/centers/cqpi/reports/pdfs/r127.pdf.)

Bisgaard, S., C. A. Vivacqua, and A. L. S. de Pinho (2005). Quality Quandaries: Not all models are polynomials. *Quality Engineering*, **17**(1), 181–186.

Bohrer, R., W. Chow, R. Faith, V. M. Voshi, and C.-F. Wu (1981). Multiple decision rules for factorial simple effects: Bonferroni wins again! *Journal of the American Statistical Association*, **76**, 119–124.

Bonett, D. G. and A. J. Woodward (1993). Analysis of simple main effects in fractional factorial experimental designs of Resolution V. *Communications in Statistics A*, **22**, 1585–1593.

Bowman, D. T. (2000). TFPlan: Software for restricted randomization in field plot design. *Agronomy Journal*, **92**, 1276–1278.

Box, G. (1990). George's Column: Do interactions matter? *Quality Engineering*, **2**(3), 365–369.

Box, G. (1990–1991a). A simple way to deal with missing values in designed experiments. *Quality Engineering*, **3**(2), 249–254. (This is also Report No. 57, Center for Quality and Productivity Improvement, University of Wisconsin-Madison and can be downloaded at http://www.engr.wisc.edu/centers/cqpi/reports/pdfs/r057.pdf.)

Box, G. (1990–1991b). Finding bad values in factorial designs. *Quality Engineering*, **3**(3), 405–410.

Box, G. (1992). What can you find out from sixteen experimental runs? *Quality Engineering*, **5**(11), 167–178. (This article is also Report No. 78, Center for Quality and Productivity Improvement, University of Wisconsin-Madison and can be downloaded from http://www.engr.wisc.edu/centers/cqpi/reports/pdfs/r078.pdf.)

Box, G. (1999–2000). The invention of the composite design. *Quality Engineering*, **12**(1), 119–122.

Box, G. E. P. and N. R. Draper (1969). *Evolutionary Operation.* New York: Wiley.

Box, G. E. P., W. G. Hunter, and J. S. Hunter (1978). *Statistics for Experimenters.* New York: Wiley.

Box, G. E. P. and R. D. Meyer (1986). An analysis of unreplicated fractional factorials. *Technometrics*, **28**(1), 11–18.

Box, G. E. P., R. D. Meyer, and D. Steinberg (1996). Follow-up designs to resolve confounding in multifactor experiments. *Technometrics*, **38**, 303–332.

Box, G. E. P. and J. Tyssedal (2001). Sixteen run designs of high projectivity for screening. *Communications in Statistics—Simulation and Computation*, **30**(2), 217–228.

Brownlee, K. A. (1953). *Industrial Experimentation.* New York: Chemical Publishing.

Casella, G. and R. L. Berger (2001). *Statistical Inference.* Belmont, CA: Brooks/Cole.

Chapman, R. E. and V. Roof (1999–2000). Designed experiment to stabilize blood glucose levels. *Quality Engineering*, **12**(1), 83–87.

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, **24**, 17–36.

Cobb, G. W. (1998). *Introduction to Design and Analysis of Experiments.* New York: Springer-Verlag.

Czitrom, V. (2003). Guidelines for selecting factors and factor levels for an industrial designed experiment. In *Handbook of Statistics*, Vol. 22, Chap. 1 (R. Khattree and C. R. Rao, eds.). Amsterdam: Elsevier Science B. V.

Czitrom, V., P. Mohammadi, M. Flemming, and B. Dyas (1998). Robust design experiment to reduce variance components. *Quality Engineering*, **10**(4), 645–655.

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, **1**, 311–341.

Daniel, C. (1973). One-at-a-time plans. *Journal of the American Statistical Association*, **68**, 353–360.

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.

Donev, A. N. (2004). Design of experiments in the presence of errors in factor levels. *Journal of Statistical Planning and Inference*, **126**, 569–585.

Draper, N. R. and D. M. Stoneman (1964). Estimating missing values in unreplicated two-level factorial and fractional factorial designs. *Biometrics*, **20**(3), 443–458.

Eibl, S., U. Kess, and F. Pukelsheim (1992). Achieving a target value for a manufacturing process: A case study. *Journal of Quality Technology*, **24**(1), 22–26.

Emanuel, J. T. and M. Palanisamy (2000). Sequential experimentation using two-level fractional factorials. *Quality Engineering*, **12**(3), 335–346.

Franks, J. (1998). The importance of hierarchy in design of experiments. *Quality in Manufacturing*, March–April issue. Available at: http://www.manufacturingcenter.com/qm/archives/0398/398doe.htm.

Fuller, H. T. and S. Bisgaard (1996). A comparison of dispersion effect identification methods for unreplicated two-level factorials. Report No. 132, Center for Quality and Productivity Improvement, University of Wisconsin-Madison. This can be downloaded at http://www.engr.wisc.edu/centers/cqpi/reports/pdfs/r132.pdf.)

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.

Giesbrecht, F. G. and M. L. Gumpertz (2004). *Planning, Construction, and Statistical Analysis of Comparative Experiments*. Hoboken, NJ: Wiley.

Glasnapp, D. R. and J. Sauls (1976). Comparative magnitude of simple effects as an interpretive index in factorial ANOVA interactions. *Journal of Experimental Education*, **45**(2), 42–46.

Goos, P. and M. Vandebroek (2004). Outperforming completely randomized designs. *Journal of Quality Technology*, **36**(1), 12–26.

Haaland, P. D. and M. A. O'Connell (1995). Inference for contrast-saturated fractional factorials. *Technometrics*, **37**, 82–93.

Hamada, M. and N. Balakrishnan (1998). Analyzing unreplicated factorial experiments: A review with some new proposals. *Statistica Sinica*, **8**, 1–41.

Hare, L. B. (1988). In the soup: A case study to identify contributors to filling variability. *Journal of Quality Technology*, **20**(1), 36–43.

Hicks, C. R. and K. V. Turner, Jr. (1999). *Fundamental Concepts in the Design of Experiments*, 5th ed. Oxford, UK: Oxford University Press.

Hinklemann, K. and O. Kempthorne (2005). *Design and Analysis of Experiments. Vol. 2: Advanced Experimental Design.* Hoboken, NJ: Wiley.

Hsieh, P. I. and D. E. Goodwin (1986). Sheet molded compound process improvements. In *Fourth Symposium on Taguchi Methods*, pp. 13–21. American Supplier Institute, Dearborn, MI.

Hunter, W. G. (1977). Some ideas about teaching design of experiments with $2^5$ examples of experiments conducted by students. *The American Statistician*, **31**(1), 12–17.

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.

Joiner, B. L. and C. Campbell (1976). Designing experiments when run order is important. *Technometrics*, **18**, 249–260.

Joseph, V. R. (2000). Experimental sequence: A decision strategy. *Quality Engineering*, **12**, 387–393.

Ju, H. L. and J. M. Lucas (2002). $L^k$ factorial experiments with hard-to-change and easy-to-change factors. *Journal of Quality Technology*, **34**(4), 411–421.

Kao, L.-J., W. I. Notz, and A. M. Dean (1997). Efficient block designs for estimating main effects contrasts. *Journal of the Indian Society of Agricultural Statistics*, Special Golden Jubilee Issue, **49**, 249–258.

Kinzer, G. R. (1985). Application of two-cubed factorial designs to process studies. In *Experiments in Industry: Design, Analysis, and Interpretation of Results* (R. D. Snee, L. B. Hare, and J. R. Trout, eds.). Milwaukee, WI: Quality Press.

Kirk, R. E. (1995). *Experimental Design: Procedures for the Behavioral Sciences*, 3rd ed. Belmont, CA: Duxbury Press.

Kuehl, R. O. (2000). *Design of Experiments: Statistical Principles of Research Design and Analysis*, 2nd ed. Pacific Grove, CA: Brooks/Cole.

Lenth, R. V. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics*, **31**, 469–473.

Lenth, R. V. (2001). Review of book by Weber and Skillings. *The American Statistician*, **55**(4), 370.

Lin, T. and B. Chananda (2003). Quality improvement of an injection-molded product using design of experiments: A case study. *Quality Engineering*, **16**(1), 99–104.

Loughin, T. M. (1998). Calibration of the Lenth test for unreplicated factorial designs. *Journal of Quality Technology*, **30**, 171–175.

Loughin, T. M. and W. Noble (1997). A permutation test for effects in an unreplicated factorial design. *Technometrics*, **39**, 180–190. (A SAS program for performing this test can be found at http://www-personal.ksu.edu/∼loughin/permtest.sas.)

Lucas, J. M. (1999). Comparing randomization and random run order in experimental design. In *AQC Annual Quality Congress Transactions*, pp. 29–35. American Society for Quality, Milwaukee, WI.

Lynch, R. O. (1993). Minimum detectable effects for $2^{k-p}$ experimental plans. *Journal of Quality Technology*, **25**(1), 12–17.

McGrath, R. N. and D. K. J. Lin (2001). Testing multiple dispersion effects in unreplicated fractional factorial designs. *Technometrics*, **43**, 403–414.

Monod, H. and R. A. Bailey (1993). Valid restricted randomization for unbalanced designs. *Journal of the Royal Statistical Society, Series B*, **55**, 237–251.

Montgomery, D. C. (1997). *Design and Analysis of Experiments*, 4th ed. New York: Wiley. (6th ed. in 2004).

Montgomery, D. C., R. H. Myers, W. H. Carter, Jr., and G. G. Vining (2005). The hierarchy principle in designed industrial experimentation. *Quality and Reliability Engineering International*, **21**, 197–201.

Natrella, M. (1963). *Experimental Statistics*, National Bureau of Standards Handbook 91. Washington, DC: United States Department of Commerce.

Nelson, L. (2003). Designed experiments with missing or discordant values. *Journal of Quality Technology*, **35**(2), 227–228.

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*. Philadelphia: American Statistical Association and Society for Industrial and Applied Mathematics.

Peeler, D. F. (1995). Shuttlebox performance in BALB/CBYJ, C57BL/6BYJ, and CXB recombinant inbred mice—environmental and genetic-determinants and constraints. *Psychobiology*, **23**(2), 161–170.

Prat, A. and X. Tort (1989). Case study: Experimental design in a pet food manufacturing company. Report No. 37, Center for Quality and Productivity Improvement, University of Wisconsin-Madison. (This can be downloaded at www.engr.wisc.edu/centers/cqpi.)

Qu, X. and C. F. J. Wu (2005). One-factor-at-a-time designs of Resolution V. *Journal of Statistical Planning and Inference*, **131**, 407–416.

Reece, J. E. (2003). Software to support manufacturing systems. In *Handbook of Statistics*, Vol. 22, Chap. 9 (R. Khattree and C. R. Rao, eds.). Amsterdam: Elsevier Science B. V.

Ryan, T. P. (1997). *Modern Regression Methods*. New York: Wiley.

Ryan, T. P. (2000). *Statistical Methods for Quality Improvement*, 2nd ed. New York: Wiley.

Ryan, T. P. (2004). *Case Study of Lead Recovery Data*. Gaithersburg, MD: National Institute of Standards and Technology, Statistical Engineering Division. (see http://www.itl.nist.gov/div898/casestud/casest3f.pdf)

Schabenberger, O., T. G. Gregoire, and F. Kong (2000). Collections of simple effects and their relationship to main effects and interactions in factorials. *The American Statistician*, **54**, 210–214.

Schoen, E. D. (2004). Dispersion-effects detection after screening for location effects in unreplicated two-level experiments. *Journal of Statistical Planning and Inference*, **126**(1), 289–304.

Searle, S. R. (1987). *Linear Models for Unbalanced Data*. New York: Wiley.

Simpson, J. R., S. M. Kowalski, and D. Landman (2004). Experimentation with randomization restrictions: Targeting practical implementation. *Quality and Reliability Engineering International*, **20**, 481–495.

Stehman, S. V. and M. P. Meredith (1995). Practical analysis of factorial experiments in forestry. *Canadian Journal of Forest Research—Revue Canadienne de Recherche Forestiere*, **25**(3), 446–461.

Sztendur, E. M. and N. T. Diamond (2002). Extensions to confidence region calculations for the path of steepest ascent. *Journal of Quality Technology*, **34**(3), 289–296.

Taguchi, G. (1987). *System of Experimental Design*, Vol. 1. White Plains, NY: UNIPUB.

Taihrt, K. J. (1971). Randomization for $2^{n-p}$ factorials in sequential experimentation. *Journal of Quality Technology*, **3**, 120–128.

Taihrt, K. J. and D. L. Weeks (1970). A method of constrained randomization for $2^n$ factorials. *Technometrics*, **12**, 471–483.

Toews, J. A., J. M. Lockyer, D. J. G. Dobson, E. Simpson, A. K. W. Brownell, F. Brenneis, K. M. MacPherson, and G. S. Cohen (1997). Analysis of stress levels, among medical students, residents, and graduate students at four Canadian schools of medicine. *Academic Medicine*, **72**(11), 997–1002.

Tripolski, M., Y. Benjamini, and D. M. Steinberg (2005). The false discovery rate for multiple testing in large factorial experiments. In *Proceedings of the Fifth Annual European Network of Industrial and Applied Statistics*, Newcastle upon Tyne, UK. (submitted for publication)

Tukey, J. W. (1949). One degree of freedom for non-additivity. Biometrics, **5**(3), 232–242.

Webb, D. F. and J. M. Lucas (2004). Blocking strategies for factorial experiments with hard-to-change factors. In *Proceedings of the Joint Statistical Meetings*, pp. 2181–2188. American Statistical Association, Alexandria, VA.

Webb, D. F., J. M. Lucas and J. J. Borkowski (2004). Factorial experiments when factor levels are not necessarily reset. *Journal of Quality Technology*, **36**(1), 1–11.

White, L. V. and W. J. Welch (1981). A method for constructing valid restricted randomization schemes using the theory of D-optimal design of experiments. *Journal of the Royal Statistical Society, Series B*, **43**, 167–172.

Winer, B. J., D. R. Brown, and K. M. Michels (1991). *Statistical Principles in Experimental Design*, 3rd ed. New York: McGraw-Hill.

Wisnowski, J. W., G. C. Runger, and D. C. Montgomery (1999–2000). Analyzing data from designed experiments: A regression tree approach. *Quality Engineering*, **12**(2), 185–197.

Woodward, A. J. and D. G. Bonett (1991). Simple main effects in factorial designs. *Journal of Applied Statistics*, **18**, 255–264.

Wu, C. F. J. and M. Hamada (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization.* New York: Wiley.

Yin, G. Z. and D. W. Jillie (1987). Orthogonal design for process optimization and its application in plasma etching. *Solid State Technology*, **30**, May, 127–132.

Youden, W. J. (1964). Inadmissible random assignments. *Technometrics*, **6**, 103–104.

Youden, W. J. (1972). Randomization and experimentation. *Technometrics*, **14**, 13–22.

## EXERCISES

**4.1** Show that when the conditional effect estimates for one of the factors in a $2^2$ design differ only in sign, the corresponding main effect estimate must be zero.

**4.2** Show that the approach for obtaining a main effect estimate given in Section 4.3 is equivalent to the method used in Section 4.1.

**4.3** Consider the following data, with the data ordered in accordance with Table 4.1: 70, 62, 65, and 63. Show that the lines cross in the interaction for one of the two ways to construct the plot but not for the other way. What does this suggest about how interaction plots might be used?

**4.4** An experiment is run using a $2^3$ design. The eight response values are as follows.

| | $C_{low}$ | | $C_{high}$ | |
|---|---|---|---|---|
| | $B_{low}$ | $B_{high}$ | $B_{low}$ | $B_{high}$ |
| $A_{low}$ | 21 | 23 | 27 | 29 |
| $A_{high}$ | 24 | 19 | 26 | 32 |

**(a)** Construct the *BC* interaction plot.
**(b)** Based solely on this plot, would you recommend that the *B* and *C* main effects be reported? Why, or why not?

**4.5** Consider a $2^2$ design with three observations per treatment combination. If the *AB* interaction is very close to zero, what will be the relationship between the conditional effects of factor *B*?

**4.6** Critique the following statement: "The coefficients obtained from use of a $2^3$ design can sometimes be unreliable because of multicollinearity."

**4.7** Bill Hunter believed that statisticians should "do statistics." Similarly, students and others who want to learn how to design experiments should "practice" designing experiments and carrying them out. A manufacturing plant is not the place to learn, however, as costly mistakes could be made. Instead, practice should be gained in scenarios where mistakes will do no harm. Many examples of innocuous $2^3$ experiments performed by students were given in a paper by Hunter (1977, *The American Statistician*, **31**(1), pp. 12–17). Conduct an actual experiment using a $2^3$ design and do a thorough analysis of the data. Ideally, this should be in a subject area that interests you; so consider one of the 32 examples in the paper by Hunter, only if nothing else comes to mind. If data points are both inexpensive and easy to obtain, then conduct a replicated experiment.

**4.8** Critique the following statement: "I'm not worried about the effect of large interactions in data that results from my use of factorial designs because if I had such interactions, I would simply include them in the model along with the factors that comprise large interactions."

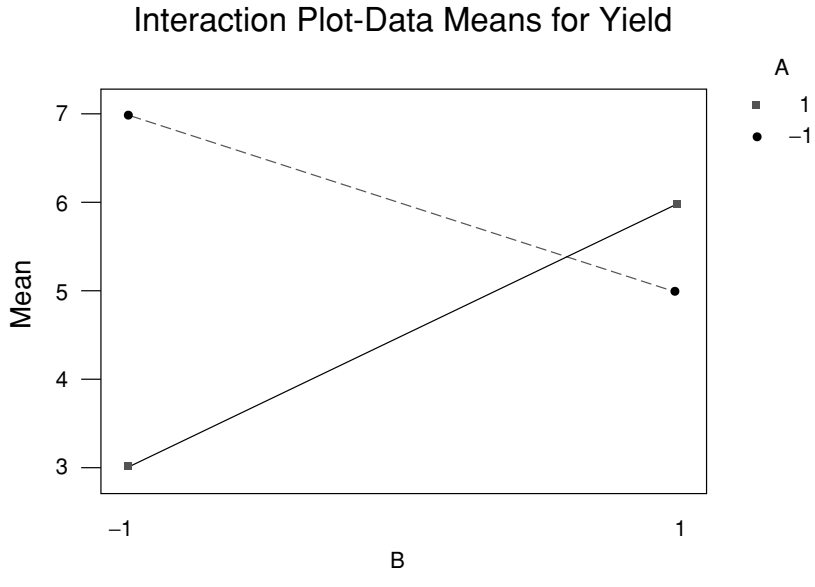
**4.9** Assume that Analysis of Means is to be applied to data from a $2^3$ design with four replicates. The estimate of $\sigma$ using $s$ thus has 24 degrees of freedom. If any main effect or interaction effect is to be significant using $\alpha = .01$, the *difference* between the two means for any of the effects must exceed _____ when $s = 4$.

**4.10** Analyze the following data from a $2^2$ design with three replicates, assuming that *A* and *B* are fixed factors.

| | | | A | |
|---|---|---|---|---|
| | | Low | | High |
| | Low | 18 14 17 | | 8 9 12 |
| B | | | | |
| | High | 13 18 16 | | 11 13 14 |

**4.11** Assume that ANOM is used for data from a $2^4$ design. How would you suggest that the effects be arranged (ordered) in a single ANOM display so as to permit an easy visual comparison of interactions relative to main effects?

**4.12** Consider the following interaction plot.




**(a)** Assuming that this is for an unreplicated $2^2$ design, what is the estimate of the $B$ effect? How would you report the effect of factor $B$ to management?

**(b)** If this plot were for an unreplicated $2^4$ design, each point in the plot would actually be the average of how many observations?

**4.13** Natrella (1963, p. 12–14, references) gave part of the data from a larger experiment that was designed to evaluate the effect of laundering on certain fire-retardant treatments for fabrics. The 16 observations given are from a $2^4$ factorial design with the factors being $A$ – fabric, $B$ – treatment, $C$ – laundering condition, and $D$ – direction of the test. The response variable is inches burned, measured on a standard size sample after a flame test. The data are as follows: (1) (4.2), $a$(3.1), $b$(4.5), $ab$(2.9), $c$(3.9), $ac$(2.8), $bc$(4.6), $abc$(3.2), $d$(4.0), $ad$(3.0), $bd$(5.0), $cd$(4.0), $abd$(2.5), $acd$(2.5), $bcd$

(5.0), $abcd(2.3)$. Analyze the data, paying particular attention to the possible need to compute conditional effects.

**4.14** Assume that the levels of a factor are 250 and 375 (degrees Fahrenheit). What is the coding transformation that would be used to convert the levels of the factor to $+1$ and $-1$?

**4.15** Explain the difference between a $2^3$ design and a $3^2$ design.

**4.16** Assume that an experiment with a $2^3$ design is used and only the main effects are significant, and (only) those terms are to be used in the model. Indicate the relationship between the model coefficients with the coded data and those for the raw data, using whatever symbolism you prefer.

**4.17** Show that for a $2^2$ design with $r$ replicates, $SS_A$ can be written as $r(\overline{A}_{\text{high}} - \overline{A}_{\text{low}})^2$, with of course "high" and "low" denoting the high and low levels, respectively.

**4.18** Under what conditions, if any, would it be safe to use high-order interactions in estimating $\sigma^2$?

**4.19** Consider the interaction plot in Exercise 4.12. Construct the plot with factor $A$ plotted on the horizontal axis. (All of the plotted points in the graph in Exercise 4.12 are integers, which may not be obvious from the graph.)

**4.20** Considering the nature of the design that was used to produce the data given in Exercise 4.4, what would be necessary before an ANOM display of the effect estimates could be constructed? Assuming that condition is met, if all of the effect estimates are shown in an ANOM display, what will be the length of the line segment that connects the two plotted points for the $ABC$ effect?

**4.21** Assume that a data analyst has constructed an ANOM display for data from an experiment in which a $2^3$ design was used, with two of the factors fixed and the other factor random. Does this present a problem? If so, explain what the problem is in regard to testing hypotheses.

**4.22** Verify $E(MS_A)$ that was given in Appendix C.

**4.23** Designed experiments have been performed by very young students for class projects. Eric Wasiloff and Curtis Hargitt, a pair of ninth grade students, used a $2^3$ design to determine factors that affect AA battery life. The results of their experiment were reported in their article "Using DOE to determine AA battery life" (*Quality Progress*, March, 1999, pp. 67–71). The second page of their

article contains the all-too-familiar-words "... the typical one-factor-at-a-time method." The response variable that they used was time to discharge in minutes and the three factors were battery type (Durall high-cost alkaline batteries and Panasonic low-cost dry cell batteries), connector design type (gold-plated and standard), and battery temperature (ambient and cold). The data are given below.

| Battery type | Connector design type | Temperature | Time to Discharge |
|---|---|---|---|
| High | Gold-plated | Ambient | 493 |
| High | Gold-plated | Cold | 490 |
| High | Standard | Ambient | 489 |
| High | Standard | Cold | 612 |
| Low | Gold-plated | Ambient | 94 |
| Low | Gold-plated | Cold | 75 |
| Low | Standard | Ambient | 93 |
| Low | Standard | Cold | 72 |

The students claimed that the "battery type" factor was significant (which is obvious from looking at the data, and concluded that the other two factors are not significant). They obtained $F$-statistics of 170, 0.7, and 0.4 (rounded off) for these three effects, which led to their conclusion.

**(a)** Since this is an unreplicated factorial, how were those $F$-statistics obtained? What advice would you give the students/authors?

**(b)** Do you feel that a conditional effects analysis of these data is necessary? Explain.

**4.24** Although a $2^2$ design is seldom used in practice, it was used in an experiment described by D. F. Aloko and K. R. Onifade in their paper, "The stability of the adsorption of some anions on chemical manganese dioxide," *Chemical Engineering and Technology*, **26**(12), 2003, pp. 1281–1283. There were three response variables, so three regression models were developed, with the two main effects and interaction effect used in each model. The design was unreplicated, so there were four values for each response variable. The value of $R^2$, which was termed the "regression coefficient," was given for each model and the values were .54, .67, and .84. Something is wrong here. What is it? If possible, read the paper and try to determine the cause of the problem.

**4.25** In the paper cited in the preceding problem, the authors gave the model with both main effects, interaction term and intercept term, which they referred to as Eq. (3) and stated, "The $2^2$ arrangement, when applied to only two variables, permits uncorrelated, low variance estimates of the four coefficients indicated in Eq. (3)." Apart from the fact that the word "estimates" should have been "estimators," is there anything else that is erroneous or suspect in the quote?

In particular, what is the relationship between the expression for the variances of the estimators when a $2^2$ design is used relative to the variances of the appropriate estimators when a $2^3$ design is used?

**4.26** Various books and journal articles have listed the two levels of a factor as (0, 1) or (1, 2) rather than (−1, 1). How would you code a temperature factor with levels of 450 and 470 so as to produce (a) the first pair and (b) the second pair. Does the use of (0, 1) or (1, 2) present any special problems in the analysis of data?

**4.27** Usher and Srinivasan (*Quality Engineering*, **13**(2), 2000–2001) describe a $2^4$ experiment for which there were six observations per treatment combination, but the data were not analyzed as having come from a replicated experiment. Specifically, the authors stated, "Note that these six observations do not represent true replicates in the usual sense, because they were not manufactured on unique runs of the process." Consequently, the average of the six values at each treatment combination was used as the response value, and the data were thus analyzed as if it had come from an unreplicated experiment. Explain how analyzing the data as if the data had come from a replicated experiment could produce erroneous results.

**4.28** Somewhat similar to the scenario in Exercise 4.27, Ryan (2004, references) analyzed data from a case study using both averages and individual observations and compared the results, with both approaches necessary because the replicates weren't quite true replicates. Read Ryan (2004), which is available on the Web, and comment.

**4.29** Assume that you have data from a $2^5$ design with two replicates. What is the first step that you will take in analyzing the data? Would your answer be the same if the data in hand were from an unreplicated $2^2$ design? Explain.

**4.30** (Requires MINITAB). The sample data file YIELDPLT.MTW that comes with MINITAB contains data from a $2^3$ design with two response variables: yield of a chemical reaction and cost. The factors are reaction time, reaction temperature, and type of catalyst. There were enough resources for 16 runs, but only 8 could be made in a day. Therefore, two replicates of the $2^3$ design were used, with replicates being days. Analyze the data, using yield as the response variable. (You will notice that a two-factor interaction is significant. Does this necessitate the use of conditional effects? Why, or why not?)

**4.31** Photolithography is the process of transferring a pattern from a photomask onto the surface of a silicon wafer or substrate. In photolithography it is important to achieve a certain line width of the etched grid. One of the key chemicals of this photolithographic process is polyvinyl alcohol (PVA). In addition, the other components of the emulsion, which we will label as $A$, $B$, and $C$, are

thought to be important factors in controlling line width. In order to study the effects of these three components on line width, a $2^3$ factorial was chosen as the design. The levels of each factor used in the experiment were as follows.

| Factor | Low | High |
|---|---|---|
| A | 0% | 6% |
| B | 8% | 16% |
| C | 0% | 3% |

The line widths for the eight runs, written in the usual (Yates) order were as follows: 6.6, 7.0, 9.7, 9.4, 7.2, 7.6, 10.0, and 9.8. Of course the order of the runs was randomized, and assume that the factors were reset after each run. Determine the effects that are significant.

**4.32** Assume that you have data from a $2^4$ design with three replicates. Do you need to assume homogeneity of variance if you use ANOM or does that apply only to ANOVA? Explain.

**4.33** Referring to Exercise 4.32, will the decision limits for assessing each effect if ANOM is used be computed using a $t$-value? If a $t$-value is appropriate, what will be the degrees of freedom? If a $t$-value is inappropriate, how should the decision limits be computed?

**4.34** Consider the data for the $2^4$ design in Section 4.10. It was shown in that section that a badly misrecorded value can result in two distinct lines on a normal probability plot, thus providing a signal that something is wrong. Does this same phenomenon occur when the number is recorded as say, 26.2, or 36.2 or 46.2 instead of 62.0? What does this suggest about using normal probability plots to detect bad data?

**4.35** Consider Example 4.5, in which the *ABC* interaction was small. Does this mean that the two *AB* interaction graphs for each level of *C* will be similar? Construct the displays and comment.