CHAPTER 6

# Experimental Designs

Designed experiments are often costly and consequently there has long been interest in determining adequate sample sizes, with such interest dating at least from Harris, Horvitz, and Mood (1948).

In examining sample size determination for experimental designs, we need to consider replicated and unreplicated designs separately. Included in the latter category are fractional factorial designs, which are generally not replicated. In this chapter we will question whether they should be replicated, and what would be gained by using replication. We will also consider whether certain types of blocking designs, such as Latin square designs, should be replicated.

Another thing to keep in mind is that a large number of experimental runs, which may result if the experimenter wants to detect a somewhat small effect, could, depending on the application, generally not be performed in a short period of time. This could create a time effect if the results did vary over time. An experimenter might not anticipate this, however, with the consequence that a time effect may not be designed into an experiment so that time can be separated from the error term. If not, the error term may be considerably inflated and the power of hypothesis tests reduced. This should be kept in mind in determining the sample size for experiments that cannot be performed quickly.

Since sample sizes will be determined so that effects of specified magnitudes can be detected with high probability, we will almost totally restrict attention to fixed factors until Section 6.13, since effect sizes are not used with random factors (variance components are used instead). Sample size determination for designs with random factors is discussed briefly in Section 6.13.

We should also keep in mind that experimentation should ideally be sequential so that more than a single experimental design should be used. For example, a two-level screening design might be used in the first stage to identify the important factors, followed by a design with more levels that would be used to try to identify

*Sample Size Determination and Power*, First Edition. Thomas P. Ryan.

optimum levels of the important factors as well as to identify interaction terms that might be needed in the model.

Daniel (1976) recommended that 33 to 50% of the resources be spent on the first experiment, and Box, Hunter, and Hunter (1978) recommended that at most 25% of the resources be used for the first experiment. (The latter did not discuss this topic in the 2nd edition, 2005, of their book.) This means that sample size (i.e., design size) must be determined for each design that is part of the sequence. It is especially important that this be planned carefully so that the budget is not exceeded. (Recall the discussion in Section 2.7 on the cost of sampling.) For some experimentation objectives it may not be readily apparent that sequential experimentation should be performed, whereas for other objectives, such as determining optimum operating conditions, the use of at least two experimental designs is routine and common practice. This is the case in response surface experimentation, which is discussed in Section 6.6. In general, once the linear effect of important factors is identified, it is desirable to see if any of those factors additionally have any nonlinear effects, such as quadratic effects.

## 6.1 ONE FACTOR—TWO FIXED LEVELS

Consider the simplest case of one factor with two fixed (i.e., predetermined) levels, with the data independent between the two levels. The data could be analyzed using either a pooled-$t$ test in Eq. (1.1) or by using analysis of variance, with the results being equivalent.

Assuming that the number of observations to be made at each of the two levels is the same and the variances at each level are equal and known (admittedly a strong assumption), the formula for the sample size at each level, $n/2$, assuming a two-sided test, is given by

$$\frac{n}{2} = \frac{(z_{\alpha/2} + z_{\beta})^2 (2\sigma^2)}{\Delta^2} \tag{6.1}$$

(See the chapter Appendix for the derivation.) Here, as in previous chapters, $Z_\alpha$ is the standard normal variate for the selected value of $\alpha$, $Z_\beta$ is the standard normal variate for $1 - \text{power} = \beta$, and $\Delta$ is the difference between the two means. If $\alpha = .05$, $\beta = .20$ (corresponding to a desired power of .80), and we express $\Delta$ generally as $k\sigma$, then

$$\frac{n}{2} = \frac{(1.96 + 0.84)^2 (2\sigma^2)}{(k\sigma)^2} = \frac{15.68}{k^2} \tag{6.2}$$

Thus, $n/2$ is a decreasing function of $k$ as a greater sample size is needed to detect a small difference than to detect a large difference.

Especially when human subjects are involved, the experimenter will at times have a fixed number of subjects available to participate in a study. In that case, it is useful to solve for $Z_\beta$ and ultimately determine power as a function of $n/2$ and either $\Delta$ or $k\sigma$. Doing so produces

$$Z_\beta = \frac{\Delta\sqrt{\frac{n}{2}} - Z_{\alpha/2}\sigma\sqrt{2}}{\sigma\sqrt{2}}$$

If we let $\Delta = k\sigma$, this simplifies to

$$Z_\beta = \frac{k\sqrt{n}}{2} - Z_{\alpha/2}$$

Then $\beta = 1 - \Phi(k\sqrt{n}/2 - Z_{\alpha/2})$, with $\Phi(\cdot)$ denoting the normal cumulative distribution function. Thus,

$$\text{Power} = 1 - \beta = \Phi\left(\frac{k\sqrt{n}}{2} - Z_{\alpha/2}\right) \tag{6.3}$$

To illustrate, for the current example and using $k = 1$,

$$\text{Power} = 1 - \beta = \Phi\left(\frac{\sqrt{31.36}}{2} - 1.96\right) = \Phi(.84) = .80,$$

as was specified.

When data are not independent between the two levels, such as when a paired-$t$ test is used, the two samples are collapsed into one set of differences ("before" minus "after," say), and sample size determination then proceeds the same as when there is only one sample (i.e., as in Section 3.3).

Although the computations are relatively simple when there are two levels, software can be used, if desired. For example, when MINITAB is used, the power and sample size routine for comparing two means assumes that a $t$-test is being used, even though the user must specify $\sigma$. Of course, if we knew $\sigma$, then we wouldn't use $t$! We can't replace $Z$ by $t$ in Eq. (6.1) and obtain an expression that could be used to solve for the sample size because a $t$-variate depends on the degrees of freedom for the $t$-statistic in the hypothesis test, which in turn depends on the sample size. Thus, an iterative approach must be used for determining sample size whenever a $t$-statistic is used in hypothesis testing. MINITAB does not provide the capability for determining sample size when $\sigma$ is either assumed to be known or estimated without using data from the two samples that are to be taken in a two-sample procedure. That is, there is not a two-sample $Z$ procedure

whereas the user does have the option of determining sample size for either a one-sample $t$ or a one-sample $Z$.

The Power Pack applet developed by Russell Lenth (`http://homepage.stat.uiowa.edu/~rlenth/Power/`) is one of many other available software that could be used. It works the same way as does MINITAB for this test. That is, $\sigma$ for each population must be specified, but the computation is performed using the $t$-distribution. Thus, if the pooled-$t$ test is selected, the user must specify either equal or unequal sigmas; then the computation of power for specified sample sizes or for "true difference of the means," whichever is selected, is based on the use of the $t$-distribution.

For example, if $\sigma_1 = \sigma_2 = 1$, a two-tailed test is to be performed, the minimum true difference between the means that one wishes to detect is also 1 (i.e., $k\sigma = 1$ in the notation of this section), equal sample sizes for the two groups are specified, and a desired power of .80 is used, then the applet produces $n_1 = n_2 = 17$, with a power of .807. (Of course, the applet does not give fractional sample sizes, so the power will almost certainly not be equal to the specified value.)

We can verify this sample size by proceeding analogously to the example used at the start of this section. That is, if we proceed as in Eq. (6.3) but use $t$ instead of $z$, we have

$$\text{Power} = 1 - \beta = \Phi^* \left( \frac{k\sqrt{n}}{2} - t_{\alpha/2,((n_1+n_2)/2)} \right)$$

If we use $n_1 = n_2 = 17$ and $n = 34$, with $\Phi^*$ denoting the cdf of a $t$-distribution with 32 degrees of feedom, we obtain Power = .807, which is the value given by the applet.

### 6.1.1 Unequal Sample Sizes

We might expect that it would be logical to have $n_1 = n_2$. This would indeed be reasonable if $\sigma_1 = \sigma_2$, but not if they differed more than slightly. The logic is as follows. If one population had considerable variability, we would expect to need a large sample size in order to have an estimator of the population mean (i.e., the sample mean), with acceptable variability, whereas a smaller sample size would suffice for a population with less variability.

To illustrate this point, consider the following simple example. Let one (very small) population consist of the numbers 12, 23, 29, 32, and 34, and let the other population consist of the numbers 13, 15, 18, 19, and 20. For the first population, no sample of size three would have a sample mean that is close to the population mean of 26 (the closest would be 28, the average of the middle three numbers), whereas the average of the middle three numbers of the second population differs from the population mean by only 0.3. Because the population standard deviations differ considerably (9.91 versus 3.26), the sample sizes should also differ.

Lenth's applet can be used to determine either equal or unequal sample sizes for the two-sample $t$-test. The user has three options: (1) equal sample sizes, (2) choose $n_1$ and $n_2$ independently, and (3) optimal allocation, which sets $n_1/n_2 = \sigma_1/\sigma_2$, with this allocation minimizing the standard error of the difference of the sample means. To illustrate, continuing with the example in which the two population standard deviations are 9.91 and 3.26, respectively, if a difference in means of 4 is to be detected with power of .80, using a two-sided test with $\alpha = .05$, the applet gives $n_1 = 66$ and $n_2 = 21$, with power = .7997, if we try to come as close to .80 as possible, and $n_1 = 66$ and $n_2 = 22$ with power = .8043 if we require that the power be at least .80.

Such software options are clearly desirable because there will be many applications in which it will not be plausible to assume $\sigma_1 = \sigma_2$. Although a pooled-$t$ test then cannot be used, an approximate $t$-test (often called the Welch $t$-test) can be used. If both sample sizes are large, then a $z$-test could be used as each standard deviation would then be a good estimator of the corresponding population standard deviation.

The Power and Precision software, for example, does not allow for unequal sample sizes, nor does MINITAB, and problems ensue when one attempts to use either nQuery Advisor or PASS. The former has a Satterthwaite $t$-test procedure for unequal variances but the routine can be used only to solve for power, not sample sizes. If $n_1 = 66$ and $n_2 = 22$ are entered in addition to the other input, the power is given as .80, in general agreement with the Lenth applet solution. The solution given by PASS is also in general agreement, as selecting the procedure "Tests for Two Means (Two-Sample T-Test) [Differences]" and using an allocation ratio of 0.329 (= 3.26/9.91) results in 66 and 22 for the two sample sizes, respectively, with the power given as 0.80453.

See Luh and Guo (2010) for sample size determination with unequal variances.

## 6.2 ONE FACTOR—MORE THAN TWO FIXED LEVELS

When a factor has more than two levels, analysis of variance (ANOVA) is used to analyze the data. We will assume that the design is a completely randomized design (CRD). That is, the levels are assigned at random to the experimental units and the design is run with random order. For example, if there were six experimental units for each of three levels, the 18 experimental runs would be made in a random order.

The sample size hand computation is naturally more involved when there are more than two levels, but it can be done without too much difficulty, if hand computation is desired to gain understanding.

With more than two levels, one obvious question is: What "effect" is used in determining sample size? That is, should it be the largest pairwise effect, or should it be the $F$-statistic for testing the equality of the means, or something

else? The user of Lenth's applet, for example, can determine sample size based on either the $F$-test or on multiple comparisons such as Tukey's Honestly Significant Difference [see, e.g., `http://en.wikipedia.org/wiki/Tukey's_test` or Gravetter and Wallnau (2009)] or the set of $t$-tests. It might also be of interest to solve for $n$ based on the smallest difference of the effects of two levels of interest. This is discussed, for example, in Sahai and Ageel (2000, p. 63).

When an $F$-statistic is the criterion, power is computed using the value of the *noncentrality parameter*. Unfortunately, it is not possible to give a simple definition of a noncentrality parameter, and this is due in large part to the fact that noncentrality parameters for various tests in experimental design are defined differently by different writers. For example, in one-way ANOVA with $k$ means, the noncentrality parameter is often given as $\sum_{i=1}^{k}[n_i(\tau_i - \bar{\tau})^2/\sigma^2]$, where $n_i$ denotes the number of observations in level $i$, $k$ is the number of means, $\tau_i = \mu_i - \mu$, with $\mu_i$ the mean for the $i$th level and $\mu$ the overall mean (i.e., the expected value of each individual observation under the null hypothesis that the treatment means are all equal, and $\bar{\tau} = \sum_{i=1}^{k} n_i\tau_i / \sum_{i=1}^{k} n_i$) Of course, $\sigma^2$ is the variance of the individual observations under the null hypothesis. Note that the expression simplifies to $n\sum_{i=1}^{k}(\tau_i^2/\sigma^2)$ when there are an equal number of observations for each level. Some authors (e.g., Sahai and Ageel, 2000, p. 57) have given $\sum_{i=1}^{k}[n_i(\tau_i - \bar{\tau})^2/2\sigma^2]$ as the form of the noncentrality parameter, and other authors have used the square root of this expression. [See., for example, the discussion of this in Giesbrecht and Gumpertz (2004, p. 61).] These different, conflicting expressions do impede understanding somewhat.

One very general and simple explanation of a noncentrality parameter given by Minitab, Inc. at `http://www.minitab.com/en-US/support/answers/answer.aspx?log=0&id=733&langType=1033` is that noncentrality parameters "reflect the extent to which the null hypothesis is false." More specifically, it is a standardized measure of the difference between the hypothesized value of a parameter and the actual (assumed) value. We can see this if we think about the noncentrality parameter for a $t$-test, although it isn't necessary to use the noncentrality parameter for that test to compute power and to solve for a sample size, as was shown in Section 3.11. Nevertheless, in the present context it is helpful to do so. If the null hypothesis for such a test is $H_0$: $\mu = 72$, the noncentrality parameter is $(\mu - 72)/(\sigma/\sqrt{n})$, with $\mu$ denoting the true mean. Thus, the noncentrality parameter is the difference between the true mean and the hypothesized mean divided by the standard deviation of the point estimator of $\mu$, which is $\bar{x}$, with its standard deviation being $\sigma/\sqrt{n}$.

For simplicity of exposition, we will now assume that the $n_i$ are equal, so that the (first) noncentrality parameter expression simplifies to $n\sum_{i=1}^{k}\tau_i^2/\sigma^2$. If we think about that explanation relative to the noncentrality parameter, under the null hypothesis the $\tau_i = 0$, which is equivalent to saying that the $k$ means are all equal to the overall mean, $\mu$. Although mathematically unnecessary, it is somewhat useful to think of the noncentrality parameter as $n\sum_{i=1}^{k}(\tau_i - 0)^2/\sigma^2$. This is

useful because the hypothesized value won't necessarily be zero in all types of sample size determination problems. This view of the form of a noncentrality parameter is intuitive because it is the difference between the true state of nature and the hypothesized state, divided by a standardization factor, $\sigma^2$, and multiplied by the number of observations sampled from each of the $k$ populations. Of course we are continuing to assume fixed factors, whereas as will be seen in Section 6.13, expressions for noncentrality parameters with random factors are not so simple and intuitive.

Although noncentrality parameters were not discussed in previous chapters, the presentation, such as in Chapter 3, could have been along these lines. Equations (3.2), (3.4), and (3.6) show the sample size expression in each equation being a function of the squared difference of the actual parameter value and the hypothesized value, just as is being discussed here. For a pooled-$t$ test, the noncentrality parameter is $\delta/\sqrt{2\sigma^2/n}$ or the square of this quantity, if preferred, with $\delta$ denoting the difference in the population means.

For practitioners who have an interest in seeing how the power is determined rather than just relying on sample size and power software, there are a few options. One option would be to work directly with the noncentral $F$-distribution, which is available in the *R* programming language, MATLAB, and Mathematica, as well as in Power and Precision.

## ■ EXAMPLE 6.1

To illustrate, let's assume that there are three fixed levels so that the null hypothesis is that the three population means are equal. In order to compute power for a specific example, we need to specify the true state of nature. That is, what is the effect of each level of the factor? The estimate of each effect is given by $\hat{\tau}_i = \hat{\mu}_i - \hat{\mu}$, with $\hat{\mu}$ the average of all the observations and $\hat{\mu}_i$ the average of the observations for the $i$th factor level. It is conventional in Analysis of Variance (ANOVA) with a fixed factor to assume that $\sum_{i=1}^{k} \hat{\tau}_i = 0$ for $k$ levels of the factor. This is logical because the sum of the observations from the overall mean must be zero. Let's assume that $\tau_1 = 1$, $\tau_2 = 0$, and $\tau_3 = -1$. Then, assuming an experiment with ten observations obtained at each of the three levels, $\sigma^2 = 1$, and $\alpha = .05$, the noncentrality parameter is $n \sum_{i=1}^{3} \tau_i^2/\sigma^2 = 10[(1)^2 + (-1)^2]/1 = 20$. The tail probability of a noncentral $F$-distribution with 2 degrees of freedom for the numerator and 12 for the denominator (i.e., the power), as given by Power and Precision, is .9733. Thus, there is a high probability that the $F$-test will reject the hypothesis of equal population means for these values of the $\tau_i$. And similarly if some other combination of $\tau_i$ values had produced $n \sum_{i=1}^{3} \tau_i^2 = 10(2) = 20$. Of course, in practice it may be very difficult to even determine a value for $n \sum_{i=1}^{3} \tau_i^2$, much less values for the individual $\tau_i$. This general type of problem occurs with *all* types of sample size determination problems, however, as certain parameter values must of course be specified in order to solve for $n$. ■

An obvious problem with computations of this type is that if we knew the effect of each level of a factor, we wouldn't need to conduct an experiment! The more means that are involved, the more difficult it will be to think in terms of a value for the noncentrality parameter. Accordingly, it will often be easier to compute power for specific values of contrasts, such as $\mu_1 - \mu_2$. Software can also be used when this approach is taken.

Similarly, software can be used when specific methods of multiple comparisons are used, such as Tukey and Scheffé. The general textbook recommendation, however, has been that specific comparisons or contrasts should not be made unless the overall $F$-test for the equality of all the means is rejected. Because of this and because of the fact that the choice of a comparison procedure can depend on whether or not the experimenter wants to be conservative and also on whether or not the comparisons to be made are determined before looking at the data, we will not give much attention to sample size determination for multiple comparisons in this chapter, with the topic treated somewhat briefly in Section 6.2.1.

To illustrate the computation of sample size for this example such that the power will be .80, if we use, say, either Lenth's applet or Power and Precision, we need to compute $\sigma_\tau^2 = \sum_{i=1}^{3} \tau_i^2/d.f.$, where here *d.f.* (degrees of freedom) is $3 - 1 = 2$ since there are three levels. Thus, $\sigma_\tau^2 = 2/2 = 1$ for the current example. Using Lenth's applet, we find that 6 observations per level are necessary to provide power of approximately .80 (actually .8053). Of course, we know that it must be less than 10 observations per level since that produced a power of .9733.

Since power of .80 is an arbitrary choice, it is usually a good idea to do at least a small-scale sensitivity study to see how power changes with the sample size. For this example, if we use 7 observations per level instead of 6, the power increases to .877. That is a sizable jump for just a one-unit increase in the number of observations, so depending on cost and resources, at least 7 observations may be preferred. Figure 6.1 shows the power values for $n_i$ from 6 to 10.

Of course one problem with these calculations – which is common to all of the designs covered in this chapter – is that it is necessary to specify the error variance, which of course is generally unknown.

At this point it is worth noting how the case of one factor with more than two levels was handled by Cohen (1988), who defined a measure $f$ as $f = \sigma_m/\sigma$, with $\sigma$ the (assumed equal) standard deviation of each population, and $\sigma_m$ the Standard deviation of the population means" defined, using the author's notation, as

$$\sigma_m = \sqrt{\frac{\sum_{i=1}^{k}(m_i - m)^2}{k}}$$

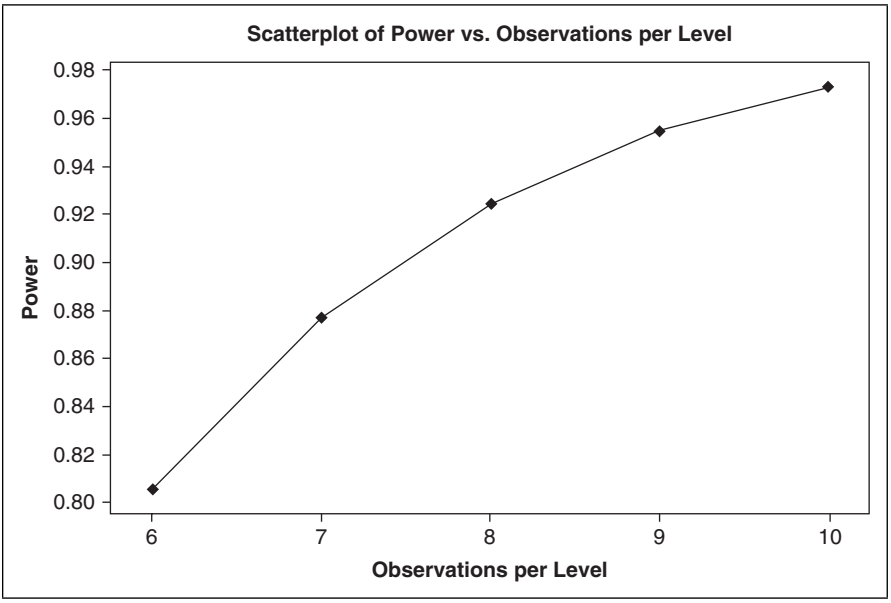


**Figure 6.1** Power for various numbers of observations per level relative to Example 6.1.

for $k$ means (levels of the factor), with $m_i$ denoting the $i$th population mean and $m$ denoting the means of the $k$ population means, assuming equal sample sizes. This, of course, is a rather loose definition since population means are fixed and do not have a standard deviation. Thus, this should be viewed as simply a measure of spread of the population means.

Given this definition of $f$, Cohen (1988, pp. 284–285) considered small, medium, and large values of $f$ to be .10, .25, and .40, respectively, which is the same set of designations that the Power and Precision software uses for Analysis of Variance with a single factor. Cohen was obviously influenced by what happens in the behavioral sciences and stated (p. 284) that "values of $f$ as large as .50 are not common in the behavioral sciences."

Comparing this set of numbers with the three corresponding numbers (.10, .30, and .50) given in Section 2.5 shows that the three designations given by Cohen depend on the type of test that is being performed. There has been some confusion about this in the literature.

It was stated in Section 2.5 that such designations have been criticized by Lenth (2001) and others. One obvious deficiency is that practical significance is ignored with such a classification. The following illustration of this should be helpful.

Of course, $\sigma$ and the population means will generally be unknown, so for illustration we will let them be estimated by their natural sample estimators

for the following example of three levels of a factor and seven observations per level.

| 1 | 2 | 3 |
|---|---|---|
| 4.62 | 4.64 | 4.65 |
| 4.61 | 4.63 | 4.63 |
| 4.63 | 4.62 | 4.63 |
| 4.62 | 4.63 | 4.64 |
| 4.62 | 4.63 | 4.65 |
| 4.63 | 4.64 | 4.63 |
| 4.61 | 4.62 | 4.65 |

Let $f$ be estimated by

$$\hat{f} = \sqrt{\frac{\sum_{i=1}^{3} (\bar{x}_i - \bar{\bar{x}})^2}{3}} \Big/ \hat{\sigma}$$

with $\hat{\sigma}$ denoting the square root of the pooled variance from an ANOVA table and $\bar{\bar{x}}$ denoting the average of the three means. Since the means are 4.62, 4.63, and 4.64, respectively, so that $\bar{\bar{x}} = 4.63$, we thus obtain $\hat{f} = \sqrt{(0.0002/3)}/0.00882 = 0.926$. Thus, the value of $\hat{f}$ is *more than double* what is considered to be a "large" value, yet the means differ by only 0.01. Thus, it is not practical to try to extend Cohen's (1988) effect sizes to general applications, and, in fairness to him, he did not try to do so. It is judicious to ignore Cohen's designations, except perhaps in behavioral science applications.

### 6.2.1 Multiple Comparisons and Dunnett's Test

When there are more than two levels of a fixed factor, there is often interest in performing multiple comparison tests involving subsets of the population means. For example, if a study involves three population means, the comparison of the first two means might be of particular interest, as well as the comparison of the second and third means. The question then arises as to how knowing that multiple comparisons are to be performed affects the determination of sample size. Witte, Elston, and Cardon (2000) addressed this issue, with the adjustment they gave being a (conservative) Bonferroni adjustment. Specifically, if $m$ comparisons are to be performed, Witte et al. (2000) gave the Bonferroni adjustment as replacing $z_d$ by $z_{d/m}$ in Eq. (3.6). They concluded: "the increase in sample size required for multiple comparisons (or tests) is not as great as might perhaps be expected. In particular, the relative sample size—allowing for the Bonferroni adjustment—is

approximately linearly related to the logarithm of the number of comparisons made." Bang, Jung, and George (2005) gave a simple method for determining sample size and power for a simulation-based multiple testing procedure that they claimed is superior to a Bonferroni adjustment. Pan and Kupper (1999) determined optimal sample sizes for four multiple comparison procedures: Scheffé, Bonferroni, Tukey, and Dunnett. Schwertman (1987) presented a method for determining sample size to detect the largest pairwise difference in means for a single factor.

Although some multiple comparison tests require equal sample sizes, it is worth noting that unequal sample sizes are preferred when Dunnett's test is used. This is a test in which all of the treatments are compared against a control and these are the only comparisons that are made. Since the data for the control are thus being "Worked more" than the data for any one of the treatments, we might suspect that it would be desirable to use more observations for the control.

We can use PASS to show that this is indeed what should be done. If the minimum detectable difference is specified as $1.5\sigma$ and $\alpha = .05$, the power is .664 with equal sample sizes of 20 per group for two groups (treatments) plus the control, compared to a power of .883 when the sample sizes are 10, 20, and 30, with 30 being used for the control group. Thus, the total number of subjects is 60 in both cases, but the power figures differ greatly.

Generally, we want to have equal sample sizes for designs so that the data are balanced, but this is an exception.

### 6.2.2 Analysis of Means (ANOM)

Although ANOVA has been the assumed method of analysis to this point in the chapter, an alternative method that is inherently graphical is Analysis of Means (ANOM), for which the same assumptions must be met as in ANOVA. The latter and ANOM have different null hypotheses, as with ANOVA the null hypothesis is that all of the means are equal, whereas with ANOM there is a null hypothesis for each mean, with that null hypothesis being that the mean is equal to the average of all of the means. With ANOVA, the alternative hypothesis is that at least one of the means differs from the others, whereas with ANOM the objective is to see whether or not one or more means differ from the average of all of the plotted means. Thus, what is being tested is different for the two procedures, so the results will not necessarily agree. In particular, when $k - 1$ sample averages are bunched tightly together but the $k$th sample average (i.e., the other one) differs considerably from the $k - 1$ averages, the $F$-value in ANOVA would likely be relatively small (thus indicating that the population means are equal), whereas the difference would probably be detected using ANOM. Conversely, if the differences between adjacent sample averages are both sizable and similar,

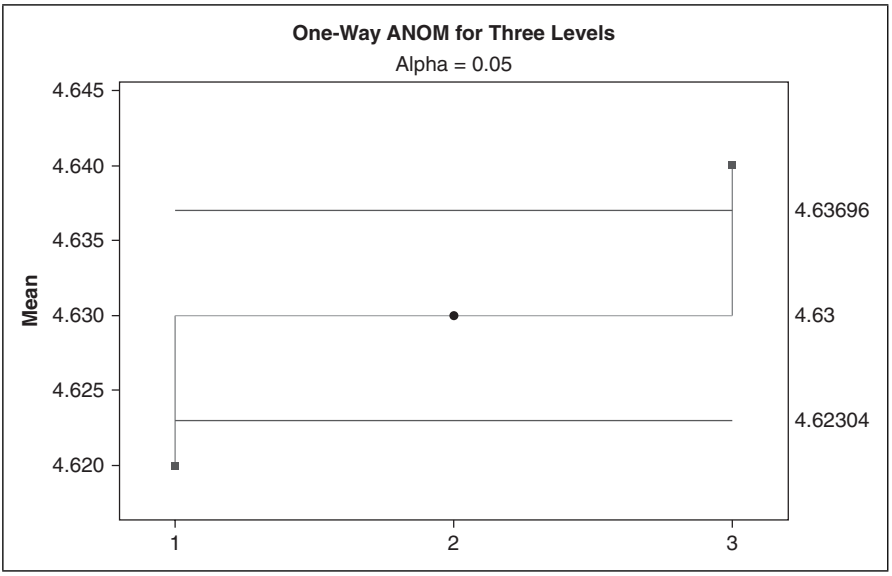


**Figure 6.2** ANOM display for one factor and three levels.

the (likely) difference in the population means is more apt to be detected with ANOVA than with ANOM.

ANOM applied to the data given in Section 6.2 results in Figure 6.2.

The display shows that, with $\alpha = .05$, the first and third means differ significantly from the overall average.

There is no software available for determining sample sizes for ANOM, but some tables and power curves are available for that purpose at the back of Nelson, Wludyka, and Copeland (2005), which are also given in Nelson (1983). (See also P. R. Nelson, 1985.)

To use those tables and figures, the user must specify a value for $\Delta$, which is defined as

$$\Delta = \max_{i,j} \frac{|\mu_i - \mu_j|}{\sigma}$$

In words, this is the largest standardized difference between two means that a user wishes to detect. For this example, $\Delta = 2.27$, and interpolating in Table B.2 of Nelson et al. (2005) shows that a sample size of 7 (which was used in this example), gives a power of approximately .90 of detecting the difference.

Nelson (1983) compared ANOM and ANOVA in terms of the sample sizes required for each to attain a certain power for a specified value of $\Delta$ and found that ANOM compared very favorably to ANOVA. For example, for power $=$ .90, $\alpha = .05$, and $\Delta = 2$, the required sample size for ANOM is the same as for ANOVA for $3 \leq k \leq 7$, and is either the same or one less for $8 \leq k \leq 16$.

### 6.2.3 Unequal Sample Sizes

As when there are two levels, it may be desirable to have unequal sample sizes when there are more than two levels. Although ANOVA with assumed equal variances is the standard approach, it will not always be appropriate to assume equal variances, as stated previously. Of course, in both cases there must be a way to obtain estimates of the either equal or unequal variances. If the variances are believed to be unequal and the resultant data corroborate this belief, then heteroscedastic ANOVA (Bishop and Dudewicz, 1978) would have to be used. Cohen (1988, p. 394), however, gave examples of conditions under which it would be desirable to have unequal sample sizes (such as highly unequal population sizes) when the variances are assumed to be equal.

It should be noted, however, that this should be an issue only if populations are small, as different population sizes of, say, 10,000, 20,000, and 30,000 are of no concern if sample sizes normally encountered in practice are to obtained from each population. This is because the standard error of a mean for a finite population is

$$\hat{\sigma}_{\bar{x}} = \hat{\sigma}/\sqrt{n}\left(\sqrt{\frac{N-n}{N-1}}\right)$$

with the parenthetical expression being the *finite population correction factor*. If samples of size 10 are to be taken from each of three populations of sizes 10,000, 20,000, and 30,000, the values of the correction factor will vary only slightly for the three population sizes.

Unlike the case of two means, Lenth's applet cannot be used to solve for unequal sample sizes, nor can Power and Precision, which, as stated previously, does not have that capability for two means. Similarly, MINITAB and nQuery cannot be used to solve for unequal sample sizes based on unequal variances.

Although Cohen (1988) does illustrate how to use his tables for the unequal sample size case, these tables are based on values of $f$ that will not be suitable for many types of application outside the behavioral sciences.

PASS does have the capability for unequal sample sizes, however, as the user enters the desired relationship between the sample sizes. To illustrate, assume that a factor has three levels, $\sigma = 10$, and the user enters 0.167, 0.334, and 0.501 for the "group allocation ratios" in the "One Way Analysis of Variance" procedure, which of course is the same as the pattern 1 2 3 using integers. The three sample sizes will then have the same ratios as the ratios of these three numbers, so if the user enters hypothesized means of 10, 20, and 20, respectively, with a desired power of .90 and $\alpha = .05$, the software gives the three sample sizes as 5, 10, and 15, respectively, which is in agreement with the specified sample size pattern. The power is .9421. The pattern 1 2 3 produces the same sample sizes for this example and it is best either to use a pattern with the first number being a 1 or to use decimal fractions that add to 1. This is because the sample sizes must

obviously be integers and the algorithm searches for the smallest multiple of the first number that will meet the power requirement, with the sample sizes for the other groups determined by the specified pattern. The user should NOT enter a pattern with the first integer greater than one because that can result in sample sizes that are larger than necessary to give the specified power. For this example, if the pattern 2 4 6 had been entered, the software gives 6, 12, and 18 as the group sample sizes, with the power of .9759 far exceeding .90. Thus, the wrong answer is obtained because the first sample size is being increased in increments of 2 and thus misses the optimal sample size of 5.

If the power is well above the specified power, the user might want to experiment and specify a lower desired power, realizing that the actual power might be close to the original desired power. For example, if the specified power is lowered to .85, the sample sizes are then 4, 8, and 12, and the power is .8690. This is closer to .90 than .9421, and since of course there is a cost associated with sampling, the 4, 8, and 12 combination might actually be preferable.

Consequently, unless PASS is used, when variances are believed to be unequal and infinite populations are involved, an experimenter might take an ad hoc approach and solve for a single $n$ that would be the appropriate sample size for testing a single mean with a postulated value of $\sigma$, which in this case would be the average of the assumed values of $\sigma$ of each population. Then solve for the $n_i$ such that the values of $\sigma_{\bar{x}_i} = \sigma_i/\sqrt{n_i}$ are approximately equal.

### 6.2.4 Analysis of Covariance

Analysis of Covariance, generally represented by the acronym ANCOVA or ANACOV, is an extension of ANOVA and is when there is a need to adjust for the effect of a quantitative variable (i.e., a covariate) that would likely affect the results of comparisons between the levels of a factor in ANOVA. Sample size determination for ANCOVA is not widely available in software, but it is available in PASS, which uses the approach given by Keppel (1991). Another approach was given by Borm, Fransen, and Lemmens (2007), which was restricted to the comparison of two treatments and thus was not a general approach.

The adjustment that is made is a natural one, as the adjustment is the correlation between the covariate and the response variable. To illustrate with a simple example, let's assume that a factor has three levels and the objective is to determine the sample size so as to detect means of 1, 2, and 3 for the three levels, respectively. Let's further assume that the within-level standard deviation is 1, the desired power is .90, and the significance level is .05. If the correlation between the covariate and the response variable is .50, then PASS gives $n = 5$ as the number of observations per level, with the actual power being .9244. What would be the value of $n$ if the correlation were smaller, or even zero, or ignored? If the correlation were .30, then $n = 6$ and if the correlation is either 0 or ignored and ANOVA used instead, then $n = 8$. Again the power is .9244. These results

are intuitive because the use of a covariate helps explain some of the variability in the response variable and thus reduces the amount of unexplained variability in essentially the same way that an additional variable reduces the unexplained variability in a regression analysis. The smaller the unexplained variability, then the smaller the sample size needed in ANCOVA.

### 6.2.5 Randomized Complete Block Designs

A randomized block design is a one-factor design with a single blocking factor. It is used when there is an extraneous factor that is likely to have an effect on the experimental values and accordingly should be separated from the error term. This design is one of the menu items for Lenth's applet. Let's assume the same scenario as in Example 6.1 in terms of $\sigma^2$, $\alpha$, and $\sigma_\tau^2$. Significance of the blocking factor is of interest only as an indicator of whether or not a randomized complete block design was an appropriate design to use. Using Lenth's applet, if 6 blocks are used so that the number of observations for each treatment level is 6 (one per block), the power is .7592—less than the .8053 value obtained without blocking. Why is that? Blocking reduces the degrees of freedom for the error term, which reduces power. Because of this loss of power, blocking should be used only when there is a strong belief that there is an identifiable extraneous factor that will affect experimental results, so that blocking should be done to remove that effect.

Note that even if the number of blocks is such that the treatment and error degrees of freedom for a particular completely randomized design (CRD) and some randomized complete block design (RCBD) are the same, the power values will differ because the number of observations that each level mean is computed from will differ. Remember from Section 6.2 that the noncentrality parameter for a one-factor CRD design with $k$ levels is $\sum_{i=1}^{k}[n_i(\tau_i - \bar{\tau})^2/\sigma^2]$, so it will increase if the $n_i$ increase and there is no other change. For example, assume a CRD with 3 levels and 7 observations per level, relative to a RCBD with 10 blocks. The error *d.f.* is 18 in both cases and the treatment *d.f.* is 2. The RCBD will have greater power because each mean is computed from 10 observations versus 7 observations for the CRD. The RCBD has a power of .9648 (still assuming $\sigma^2 = \sigma_\tau^2 = 1$ and $\alpha = .05$), compared to .877 for the CRD, as noted previously. Of course, the power values can always be computed "semimanually" by computing the value of the noncentrality parameter and entering that into software. The expression for the noncentrality parameter of the RCBD (e.g., see Giesbrecht and Gumpertz, 2004, p. 92) is equivalent to the expression for the noncentrality parameter of the CRD under the assumption of blocks being fixed, or under the (standard) assumption that there is no block $\times$ treatment interaction when blocks are random. Using the Power and Precision software for this example also gives a power of .9648, in agreement with the power obtained using Lenth's applet.

PASS also has (some) capability for a RCBD. Instead of specifying $\sigma^2$ and $\sigma_\tau^2$, however, the user enters $\sigma$ and the hypothesized means for each block, which

of course is more intuitive than $\sigma_\tau^2$. PASS cannot be used to solve for the number of blocks, however; it simply gives the power for the entered number of blocks.

### 6.2.6 Incomplete Block Designs

An incomplete block design is a design for which every level of the factor does not appear in each block. A balanced incomplete block (BIB) design has all factor levels appearing the same number of times across the blocks, and pairs of levels appearing the same number of times within the blocks. Because of these requirements regarding balance, there are restrictions on the number of blocks that can be used, which means that there are restrictions on the choice of sample size.

Most software for sample size determination are of no help with BIB designs, and this applies generally to designs that are not among the simplest and most frequently used. This is apparent from the sample size and power capabilities of SAS/STAT 9.2 and is also apparent from the capabilities of Release 16 of MINITAB. Similarly, PASS 11 can be used to generate BIBs only, not solve for the number of blocks required or give the power for an entered number of blocks.

Lenth's applet also cannot handle BIB designs for sample size determination because whereas the number of blocks determines the error degrees of freedom, the number of blocks is not equal to the number of observations used in computing the mean for each level. Consequently, power would have to be computed manually by computing the value of the noncentrality parameter for the $F$-test and specifying the number of degrees of freedom for the error term and for the factor.

To illustrate, for simplicity we will again assume three levels of the factor, with the $\tau_i$ as given in Example 6.1, and also again assume that $\sigma^2 = 1$ and $\alpha = .05$. A very simple BIB design is obtained by starting with a RCBD in three blocks and then eliminating a different level from each block. Doing so produces the following BIB design with the columns being the blocks.

| | | |
|---|---|---|
| A | C | B |
| B | A | C |

This satisfies the requirement that every treatment (i.e., level of the factor) occurs the same number of times (twice), and pairs of treatments occur together in the same block an equal number of times (once). The $F$-statistic has 2 degrees of freedom for the numerator since there are three levels, but the error term has only 1 degree of freedom since Blocks has 2 degrees of freedom. The expression for the noncentrality parameter is (Giesbrecht and Gumpertz, 2004, p. 219) $(t\lambda/m)\sum_{i=1}^{k}(\tau_i^2/\sigma^2)$, with $t$ = number of treatments, $k$ = number of levels, $m$ = block size, $r$ = number of replications, and $\lambda = [r(k-1)/(t-1)]$. Here $t = 3$, $m = 2$, $k = 3$, and $r = 2$, so $\lambda = 1$ and $(t\lambda/m) = 1.5$. The noncentrality

parameter is then (1.5) $\sum_{i=1}^{3}(\tau_i^2/\sigma^2) = (1.5)[(1)^2 + 0^2 + (-1)^2]/1 = 3.0$. The power is very low at .081 because the denominator *d.f.* is only 1. If this design were replicated so that $r = 4$, the noncentrality parameter is then 6.0 and the denominator *d.f.* increases to 4 and the power increases to .31—still low. This low power relates to what is discussed in Section 6.2.7 on Latin square designs.

Although we can't determine sample size directly (or at least not easily) for a specified power, this is not a major problem because of the restrictions on sample size in order for the design to be balanced. Thus, it would be straightforward just to examine power for the feasible sample size; that is, the feasible number of blocks.

Partially balanced incomplete block (PBIB) designs present similar problems, but as with BIB designs, it is just a matter of computing the value of the noncentrality parameter and determining the appropriate numerator and denominator degrees of freedom for the $F$-statistic.

### 6.2.7 Latin Square Designs

A Latin square design is a design for a single factor of interest with two blocking factors. The layout of the design is a square with the number of cells in the square being the square of the number of levels, with each level appearing exactly once in each column and once in each row. The following is a $3 \times 3$ Latin square in standard order, so called because the first row and first column are in alphabetical order.

| | | |
|---|---|---|
| A | B | C |
| B | C | A |
| C | A | B |

Ryan (2007) stated that a single Latin square should not be used because there is not sufficient power to detect a factor effect of moderate magnitude, such as equal to $\sigma$. Indeed, the average at each level is computed from only three observations, and the error term has only two degrees of freedom. (This is not a true error term because there is no replication, but rather results from the assumption of no interactions.)

For example, if we make the same assumptions regarding the values of $\sigma^2$, $\alpha$, and $\sigma_{\hat{\tau}}^2$ that were made in Example 6.1, the power is only .1823 when a single Latin square design is used, but the power increases sharply to .5402 when two Latin squares are used, with the power being .8318 when three Latin squares are used. These are the numbers obtained by using the value of the noncentrality parameter and computing the power manually.

The noncentrality parameter for a single Latin square design is the same as for a CRD; namely, $\sum_{i=1}^{k}(n_i\tau_i^2/\sigma^2)$, for a Latin square of order $k$, as given at the start of this section, which simplifies to $\lambda = k\sum_{i=1}^{k}(\tau_i^2/\sigma^2)$. Thus, with a $3 \times 3$

Latin square and the assumptions of Exercise 6.1, the value of the noncentrality parameter is 6 and the probability of a noncentral-$F_{2,2}$ variate with $\lambda = 6$ exceeding its .05 critical value is .1823, as indicated in the preceding paragraph.

Because of the low power that will often occur when only a single Latin square design is used, Ryan (2007, p. 77) recommended the use of multiple Latin squares, and Giesbrecht and Gumpertz (2005, p. 126) indicated that this will often be desirable. When multiple Latin squares are used, the error degrees of freedom depends on the relationship between the squares, as illustrated in detail by Giesbrecht and Gumpertz (2005, pp. 127–130). The power is thus affected by this relationship, since it depends in part on the error degrees of freedom. If the squares are unrelated, the error degrees of freedom is $s(k-1)(k-2)$, with $s$ denoting the number of squares, and as before, $k$ denotes the order of each square. The noncentrality parameter is then $ks\sum_{i=1}^{k}(\tau_i^2/\sigma^2)$. This expression for the error degrees of freedom is based on the assumption that the model has square and the treatment × square interaction as model terms, in addition to the usual terms for treatment, row, column, and error. When this model is specified in Lenth's applet and square is treated as a fixed factor, the power for the treatment effect is .5402 and .8318 for two and three squares, respectively, in agreement with the numbers given previously in this section that were computed using the values of the noncentrality parameter for two and three unrelated squares, respectively.

Care should be exercised when using software to compute power for multiple Latin squares, as the results may not agree across different software because of different ways in which the error degrees of freedom is determined. For example, Lenth's applet only breaks out degrees of freedom for row, column, treatment, and error when the default model is used, although the user can specify a different model, if desired. Thus, multiple Latin squares will increase the error degrees of freedom, with the progression being 2, 11, 20, and 29, for a single Latin square, two, three, and four Latin squares, respectively. This is appropriate only if all the squares have common rows and columns and the effects due to squares and all interactions involving squares are not listed in an ANOVA table, with the degrees of freedom for these effects instead used to contribute to the error degrees of freedom.

If the squares are unrelated, the degrees of freedom for row and for column are the degrees of freedom for a single Latin square (the order of the square minus one) multiplied by the number of squares. Although certain terms might be combined to form the error term, the appropriate starting point should be the breakdown of degrees of freedom for each of the ways in which the multiple squares could be constructed (unrelated, common rows and columns, etc.), and then a decision made regarding the pooling of terms in forming the error term.

With experimental designs in general, it is not a good idea to assume that certain effects do not exist, as the existence of certain effects can constitute a violation of the model assumptions. Thus, appropriate checks should be performed.

Consider the following example.

## ■ EXAMPLE 6.2

Jaech (1969) gave a nuclear production reactor example that involved process tubes and positions on each tube. The experiment and data are apparently fictitious, which explains why the factor was not stated as being fixed or random. There was simply the statement: "Say there are five treatments to be evaluated in a given experiment." Presumably the dataset was based on the author's consulting experience. We will assume that the factor is fixed.

There were 10 tubes used and 20 positions on each tube for a total of 200 tube–position combinations. Eight $5 \times 5$ Latin squares were used to represent these combinations. The squares must obviously be related because if rows represented tubes, then 40 tubes would be required if different tubes were used for the Latin squares. Similarly, 40 tube positions would be required but only 20 were used. Thus, there were pairs of squares that used the same five tube positions (1 and 6, 2 and 7, 3 and 8, 4 and 9, and 5 and 10), with the five different tubes used in 1 and 6, but then repeated over 2 and 7, 3 and 8, and so on.

The key question is what is the *d.f.* for the error term, as that will determine the power? The degrees of freedom given by Jaech (1969), 164, does not result from a formula given for any of the ways of constructing the squares given by Giesbrecht and Gumpertz (2005). This is because of the pairing of squares for the experimental material. The pairing relative to tube position results in 4 *d.f.* for each group of 5 positions, and since there were 4 pairs of squares, there were thus 16 *d.f.* for positions. There were 8 *d.f.* for tubes, resulting from 4 *d.f.* each for tubes 1–5 and 6–10. With 7 *d.f.* for squares $(8 - 1)$ and 4 *d.f.* for treatments $(5 - 1)$, 164 *d.f.* for error is then obtained by subtraction.

If we still assume $\sigma_{\hat{\tau}}^2 = 1$ (so that $\sum_{i=1}^{5} \tau_i^2 = 4$), $\sigma^2 = 1$, and $\alpha = .05$, as in the previous examples, Power $= 1$ is the result, which is the same value obtained using the value of the noncentrality parameter in conjunction with the Power and Precision software. The atypical use of multiple Latin squares by Jaech (1969) results in an error degrees of freedom that apparently cannot be accommodated by Lenth's applet, so in this case computation of the power by using the value of the noncentrality parameter seems necessary. ■

Such a high value of power results from the fact that so many Latin squares of a moderate size were used, combined with a sizable noncentrality parameter. For a $5 \times 5$ Latin square with $\sigma_{\hat{\tau}}^2 = \sigma^2 = 1$ and $\alpha = .05$, but now reducing the replications to 2 and assuming that the squares are unrelated results in a power of .9986 with Lenth's applet. This of course is the same value that is obtained by directly calculating the power from the noncentrality parameter and using 24 *d.f.* for the error term (not pooling the Treatment × Square interaction with the error term, although that could be done).

In this example it was assumed that the factor is fixed, whereas the power values will be different when the factor is random. For example, for this last scenario with $\sigma_{\hat{t}}^2 = \sigma^2 = 1$ and $\alpha = .05$ and the model as indicated previously, Lenth's applet gives power = .2092 for the treatment effect. Thus, the designation of the factor of interest as fixed or random makes a huge difference in the power in this example.

Since the Power and Precision software doesn't have Latin square as a menu item, power would have to be computed manually using the value of the noncentrality parameter. Obviously, it is preferable to use software with built-in capability for Latin square designs. PASS will not handle this and in general will handle only repeated measures designs and designs for clinical trials, such as crossover designs, in addition to fixed-effects Analysis of Variance with up to three factors with the software computing power after the user specifies the number of observations per cell. Thus, sample size is not determined directly with the latter.

As mentioned previously, sample size determination for designs with random factors is discussed in general in Section 6.13.

#### *6.2.7.1 Graeco-Latin Square Designs*

A Graeco-Latin square design is used when there are three extraneous factors that must be adjusted for, rather than two as in a Latin square design. The design is constructed by placing one Latin square on top of another Latin square. More formally, a pair of mutually orthogonal Latin squares are used, with Latin letters representing the factor levels in one of the squares and Greek letters used for the other square. Each combination of a Latin letter and a Greek letter occurs only once in the design.

A single Graeco-Latin square is plagued by low power in essentially the same way that a single Latin square is affected. For example, if we assume a $4 \times 4$ Graeco-Latin square design with a fixed factor of interest, $\sigma_{\hat{t}} = 1.5$ and $\sigma = 1$, so that we are trying to detect a 1.5-sigma effect, the power is only .613 for the $F$-test. The power is .834 when $\sigma_{\hat{t}} = 2.0$, but that isn't a very good power for detecting a $2\sigma$ effect. If the number of squares is increased to 2, then the power is essentially 1.0 for the 1.5-sigma effect and of course also virtually 1.0 for the $2\sigma$ effect. Thus, just using one additional Graeco-Latin square has a very profound effect. Note that the power with a single Graeco-Latin square is *less* than occurs with a $4 \times 4$ Latin square design, as the latter has power of .886 for detecting a 1.5-sigma effect. This can be explained by the difference in the degrees of freedom for the error term as the Latin square has more degrees of freedom (6 versus 3).

Thus, it is highly desirable to use multiple Graeco-Latin square designs, just as it is highly desirable to use multiple Latin square designs. Each Graeco-Latin square must be at least of order four because the degrees of freedom for the error term is $(t - 3)(t - 1)$, with $t$ denoting the number of levels. Thus,

a $3 \times 3$ Graeco-Latin square would not have any degrees of freedom for the error term.

As is true of Latin square designs, sample size determination for Graeco-Latin square designs thus amounts to determining how many squares to use.

## 6.3 TWO FACTORS

In this section we consider sample size determination when there are two factors, and we will highlight some input differences between some software for sample size determination. At the outset we will note that the "sample size" that the software computes is the number of replicates to use for each factor-level combination, not the total number of observations to obtain.

### ■ EXAMPLE 6.3

We will consider a factorial experiment with two factors, the first at three levels and the second at two levels. Thus, it is a $3 \times 2$ factorial design that can also be described as an example of two-way ANOVA. When Power and Precision is used, "Factorial analysis of variance (two factors)" is selected from the menu. The default option is to enter the effect sizes for each factor and the interaction. Although small, medium, and large are the options, alternatively a value can be entered for the effect size, which would certainly be preferable for the reasons discussed in Section 2.8. Alternative input options include the between-groups standard deviation for each factor and the mean for each level of the factor. (It should be noted that the divisor that is used for that standard deviation is the number of levels of the factor, not the number of levels minus one.) The other input option is to enter the range of the means (largest and smallest) and enter "the pattern of dispersion for remaining means," with the options being centered, uniform, and extreme. The within-cell standard deviation, $\sigma$, must also be specified.

As with sample size estimation in general, there is a burden on the software user to provide estimates that may not be easy to produce. Of course, this burden exists regardless of the software that is used. If PASS is used, very similar input options are available. First, assuming that the factors are fixed, the user would select `Means > Many Means (ANOVA) > Fixed Effects ANOVA` from the menu in the order indicated. The means for each level of each factor could be entered, or their standard deviation specified. Similarly, for the interaction effects, either the cell means or their standard deviation would be entered. For this example there are six cells (three times two).

Obviously, these input demands place a considerable burden on the user. If the user actually knew everything that has to be inputted, there would be no

need to conduct the experiment since the magnitude of the main effects and interaction effect would be known. Of course, this information is not going to be known but past experimental data might enable a reasonable estimate of $\sigma$ to be produced, and it is the relationship between the factor level means of each factor that determines the power for a given sample size, not the actual mean values. That is, the same power for a fixed sample size will result from specifying three factor-level means as 2, 4, and 6 as when 20, 40, and 60 are used, provided that $\sigma$ is estimated appropriately. That is, $\sigma$ would be larger by a factor of 10 if the true means were 20, 40, and 60 rather than 2, 4, and 6, and each data point that produced the latter were multiplied by 10. If, however, a constant (100) were added to the first set of factor-level means so as to produce 102, 104, and 106 for the second set, the power would be the same for this second set as for the first set with $\sigma$ being the same for both sets. This is because a constant is being added, not multiplied, so there is no scale change.

Note that when PASS is used for a factorial design with two factors, the sample size cannot be solved for by specifying a desired value for the power. Rather, PASS provides an option for specifying the number of observations per cell (i.e., the number of replications), with the default option being the four values of 2, 4, 8, and 16. When this option is in force, the user can see the power values for each effect that is estimated and decide if the power is acceptable for one of these four values. If not, other values could be inputted, with the user perhaps deciding to enter more than four possible sample size values so that a finer grid of power values could be displayed.

MINITAB can be used to determine sample size for general full factorial designs. The user enters the number of levels for each factor, which for this example would be "3 2" or "2 3." Of course, this indirectly enters the number of factors. The user also inputs an estimate of the standard deviation and one or more values for the maximum difference between the main effect means, in addition to one or more target power values and a single significance level. A power curve is then generated for each combination, in addition to the numerical output.

For example, if the input is "3 2" for the levels, $\alpha = .05$, $\sigma = 2$, "2" for the highest-order term in the model, "4 5" for maximum difference between the main effect means, and ".90 .95" for the power values, the graph in Figure 6.3 is produced, in addition to the numerical output. ■

Although such graphs are useful for determining the number of replicates to use, with the curves in Figure 6.3 for 3, 4, and 5 replicates, respectively, and the user having the option for displaying curves for other numbers of replicates, the user should keep in mind that these power figures are contingent on the condition $\sigma = 2$. Since $\sigma$ almost certainly would not equal whatever value was specified, the power values will, technically, be incorrect. It would be helpful to construct

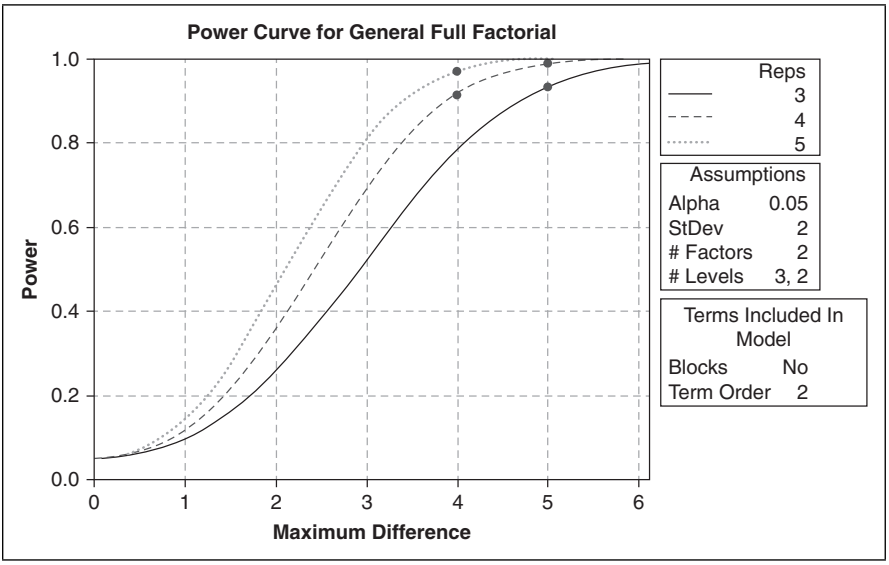


**Figure 6.3** Power curves for determining number of replicates for a $3 \times 2$ factorial design.

an approximate confidence interval on the true power value, along the lines of the discussion in Section 3.3, but this would require a pilot study to estimate $\sigma$, which might not be possible or practical. Nevertheless, the concept is an important one but has not been addressed in the literature on sample size determination for experimental designs.

nQuery could also be used to determine the number of replicates to use for this $3 \times 2$ design, but the required inputs place more exacting demands on the user than is the case with MINITAB. Specifically, whereas MINITAB requires that the maximum difference between main effect means be specified, in addition to a value for $\sigma$, nQuery requires, in addition to $\sigma$, that the variance of the means for each of the two factors be specified, in addition to the variance of the means for the interaction term. Of course, this would require that each of the means be known, and if that were the case, there would be no point in conducting the experiment!

## 6.4 $2^k$ DESIGNS

Two-level designs are the most commonly used factorial designs, with such full factorial designs designated as $2^k$ designs. These designs frequently have only one observation for each of the $2^k$ treatment combinations. When that is the case, the assumption of equal variances for all of the combinations cannot be tested. Consequently, experimenters routinely assume equal variances.

### 6.4.1 $2^2$ Design with Equal and Unequal Variances

We will first consider a replicated $2^2$ design with levels of each factor fixed and the cell variances first considered to be equal and then assumed to be unequal.

For the equal variances case, Lachenbruch (1988) gave the sample size formula for each cell as

$$n = \frac{4\sigma^2 \left(Z_{\alpha/2} + Z_\beta\right)^2}{M^2}$$

with $M$ defined as the contrast for the effect that is being tested. Specifically, with "$R$" denoting the row effect (and thus the effect of the factor that corresponds to the row), $M_R = \mu_{11} + \mu_{12} - \mu_{21} - \mu_{22}$, with $\mu_{ij}$ denoting the mean of the cell in the $i$th row and $j$th column. Similarly, the column effect is defined by the contrast $M_C = \mu_{11} - \mu_{12} + \mu_{21} - \mu_{22}$, and for the interaction effect, $M_I = \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22}$.

For the unequal variances case, Guo and Luh (2010) gave the sample size formula for each cell, $n_{ij}$, as

$$n_{ij} = \frac{S_{ij}\left(S_{11} + S_{12} + S_{21} + S_{22}\right)\left(t_{\alpha/2,df} + t_{\beta,df}\right)^2}{M_R^2} \tag{6.4}$$

with the degrees of freedom, *d.f.*, defined as

$$d.f. = \frac{\left(s_a^2/n_a + s_b^2/n_b\right)^2}{\dfrac{\left(s_a^2/n_a\right)^2}{n_a - 1} + \dfrac{\left(s_b^2/n_b\right)^2}{n_b - 1}}$$

with $n_a = \sum_{j=1}^{2} n_{1j}$, $n_b = \sum_{j=1}^{2} n_{2j}$, $s_a = \sum_{j=1}^{2} s_{1j}$ and $s_b = \sum_{j=1}^{2} s_{2j}$ The $s_{ij}$ are the estimates of the cell standard deviations, $\sigma_{ij}$. Equation (6.4) would of course have to be solved numerically for $n_{ij}$ since there are terms involving $t$-values, which depend upon the sample size through the degrees of freedom.

### 6.4.2 Unreplicated $2^k$ Designs

When a $2^k$ design is unreplicated, the value of $k$ must be determined carefully, in addition to using criteria that are different from the criteria that are normally used in solving for sample size. Indeed, the term "sample size" really doesn't have any meaning when only one observation is being used for each treatment combination.

Instead, the value of $k$ must be determined such that it will be large enough so that the significant effects can be identified using any one of the methods that have

been proposed and examined in the literature for identifying the significant effects with an unreplicated factorial. There are two things that must be considered: the value of $k$ should be large enough so that each effect will be estimated using a reasonable number of observations so that the variance of the estimator will be tolerably small, and $k$ must also be large enough so that the pseudo error term that is formed from high-order interactions will have a reasonable number of degrees of freedom.

Daniel (1976, p. 75) stated that, "as very crude guesses," approximately four real effects is average for a $2^4$ design and seven for a $2^5$ design. Therefore, it would be overkill and wasteful, especially when experimental runs are expensive, to use a $2^5$ design, which would allow the estimation of all 31 effects when only a fraction of those effects will be real effects. Only when the cost of experimentation is extremely low, such as in computer simulation experiments, would it be practical to use a $2^k$ design when $k$ is not small. Computer experiments are discussed in Section 6.15.

When a specific design is planned, Design-Expert can be used to determine the power for detecting effects of specified magnitudes, assuming that all effects will not be estimated, as would be impractical if $k > 3$. (Design-Expert is software solely for design of experiments.) This can be used as a guide for design selection.

## ■ EXAMPLE 6.4

Daniel (1976) gave an example of a $2^3$ design used in a cement experiment for which $\sigma$ was known to be about 12. Because of the latter, it is of interest to see the magnitude of effects than can be detected, relative to the effects that were declared significant using a normal probability plot approach.

If we solve for $k$ in Eq. (6.2), with $n/2 = 4$, we obtain $k = 1.98$. Thus, an effect of magnitude $1.98\sigma$ is the smallest effect that could be detected with the design. With $\sigma = 12$, this is $1.98(12) = 23.78$. A normal probability plot leads to the identification of the $B$ and $C$ main effects and the $BC$ interaction as being the real effects. The estimates of these effects are $-132.75$, $-73.5$, and 47.5, respectively. The next largest effect was the $A$ effect, with an estimate of 15.5. Thus, the normal probability plot identifies the same effects as would be identified doing the hypothesis tests using the known value of $\sigma$.

A $2^3$ design has rather limited practical value, however, because we would prefer to be able to detect effects smaller than $1.98\sigma$. Furthermore, the latter is based on the assumption that $\sigma$ is known. The typical scenario, however, is that $\sigma$ is unknown. Table 3 of Lynch (1993) indicates that, assuming $\alpha = .05$ and $\sigma$ to be estimated from the experimental data, a $2^3$ design must be replicated in order to detect an effect of $2\sigma$ or smaller with power of at least .80. Specifically, two replications are needed to detect an effect of $1.8\sigma$ with power .90. Of course, it is somewhat easier to use software than to use statistical tables, and if we use

MINITAB (Release 16), we observe that the power is .88 for detecting an effect of $1.8\sigma$. This disagrees slightly with the result in Lynch's table, with the difference apparently due to the fact that the Lynch tabular result is based on the assumption that 9 degrees of freedom are available for estimating $\sigma$ (which would be the case if the *ABC* interaction were not estimated), whereas 8 degrees of freedom are available if all effects are estimated. Using MINITAB, we find that 3 replicates are necessary to detect an effect of $1.8\sigma$ if the power is to be at least .90, with the actual power being .985.

One complication relative to sample size determination for designs with two or more factors is that the denominator in the $F$-test for testing for effect significance is not the mean squared error (MSE) when there is a mixture of fixed and random factors. Although MSE will be a component of the denominator, it won't always be the only component. Similarly, the numerator of the statistic won't necessarily be represented by MSE plus a term for what is being tested, as the numerator can contain other terms. For this reason, some Java applets, such as `http://www.math.yorku.ca/SCS/Online/power`, will handle only fixed factors, as then all effects are tested against the error term. ■

### 6.4.3 Software for $2^k$ Designs

Although it is insightful and relatively important to see how power is computed, it is, of course, much easier to use software for this purpose.

In particular, we can use software such as Design-Expert to evaluate the power of $2^k$ designs, as indicated previously. For example, the software immediately warns the user who selects an unreplicated $2^3$ design that "For this design, effects must be at least 2–3 standard deviations in size to be detected. Evaluate power for the expected number of effects. Do you want to continue with this design?"

Of course, the power depends on how many effects we wish to estimate with the $2^3$ design. We cannot estimate all of them without an external estimate of error or a normal probability plot approach because there would be no degrees of freedom for error. Although mentioned in conjunction with Example 6.4, a normal probability plot approach would be somewhat shaky for a $2^3$ design because, if we accept Daniel's rule-of-thumb, the number of real effects may render a normal probability plot useless as the pseudo error term that is created when such plots are used must be constructed using at least a moderate number of effects that are not likely to be significant.

The picture improves if we use two replicates of a $2^3$ design, as then the power is .937 for detecting an effect of magnitude $2\sigma$ for any of the main effects or interactions. (The power is the same for each effect because each effect estimate is based on the difference of two averages that are each computed from 8 observations when a $2^3$ design with two replicates is used.)

Of course, the experimenter must decide the minimum-size effect that is to be computed with high probability.

If a $1\sigma$ effect is deemed important to detect, then two replicates are not sufficient because the power is only .421 for detecting an effect of that magnitude, as obtained using MINITAB. If we use 3 replicates, the power increases to .633, to .775 for 4 replicates, and to .866 for 5 replicates. Thus, we would need 5 replicates in order for the power to be at least .80.

Thus, a $2^3$ design is not a very good design unless replicates are inexpensive or we are interested in detecting large effects and would settle for an unreplicated design.

For a $2^4$ design the picture is somewhat better, as the power of detecting a $2\sigma$ effect when the model contains only main effects and two-factor interactions is .887, but only .368 for a $1\sigma$ effect. If the design is replicated, the numbers improve to .999 and .769, respectively. With a $2^5$ design, the first pair of numbers is .999 and .757, and the second pair of numbers is .999 and .975. (With more than five factors, a full factorial would generally be impractical, exceptions being experiments for which the experimental runs are very inexpensive, as mentioned previously in Section 6.4.2.)

By comparison with these results obtained using MINITAB, the tables of Lynch (1993) are not constructed in such a way that the power is given for $1\sigma$ or $2\sigma$ effects, nor can the tables be used to determine the effect size that can be detected with power .80, as there is a jump from .75 to .85. Power of .90 is included, however.

The picture that emerges when we put all of these numbers together is that a $2^3$ or $2^4$ design should be replicated if there is a need to detect effects smaller than $2\sigma$.

## 6.5 $2^{k-p}$ DESIGNS

Since $2^{k-p}$ designs are generally not replicated, sample size considerations would seem to be irrelevant since standard hypothesis tests cannot be performed. What is not irrelevant, however, is the determination of $k$ and $p$, and the difference, $k - p$, for the size of the experiment is determined by the latter. Furthermore, $k$ should be large enough that the percentage of real effects will be small and the pseudo error term that is used with a normal probability plot will be based both on a reasonable number of observations and real effects. Since every effect that is estimable with a particular $2^{k-p}$ design is computed as the difference of two averages, with each average based on $2^{k-p-1}$ observations, it is imperative that $2^{k-p-1}$ not be so small that each average has a large variance.

As indicated in Section 6.4, Daniel (1976, p. 75) stated that there are typically 7 significant effects when a $2^5$ design is used. It follows that there would also be 7 significant effects when either a $2^{5-2}$ or a $2^{5-1}$ design is used because the real effects and the number of them are of course independent of the fractionization. Therefore, an unreplicated $2^{5-2}$ design should generally not be used because we

won't be able to identify 7 significant effects with only 8 design points because all normal probability plot methods fail if there is a very high percentage of significant effects, as is well known and has been discussed by, for example, Hamada and Balakrishnan (1998) and Ryan (2007, p. 136).

Would using a replicated $2^{5-2}$ design make any sense? Not really, as that would not only offset the savings that are gained by using a highly fractionated design, but with a few replicates there would not be enough observations at each design point to allow for good estimates of $\sigma^2$ at each point, so the estimate of $\sigma^2$ obtained from the average of the estimates over the design points could be an average computed from some large values. Carroll and Cline (1988) showed that at least 8–10 replications of a design point are necessary in order to obtain a good estimate of the variance at that point. Of course, it would be totally impractical to use at least 8 replicates of a $2^{5-2}$ design.

Thus, in selecting a particular $2^{k-p}$ design, the user should be guided by Daniel's rule-of-thumb and any other avilable empirical evidence regarding the number of significant effects that might be expected for a given value of $k$. If an experimenter wants to "push the limit" by using a highly fractionated design (i.e., using a small value of $k - p$), such experimenters should be guided by results given in the study of Hamada and Balakrishnan (1998) regarding the maximum number of real effects that can be handled with normal probability plot methods, and which normal probability plot methods should be used for various expected fractions of real effects.

## 6.6 DETECTING CONDITIONAL EFFECTS

The analysis of conditional effects (also called simple effects in the literature) was emphasized by Ryan (2007), and its use is imperative when there are large interactions. (A conditional effect is an effect computed by splitting the data and using part of it in computing each conditional effect.) Ryan (2007, pp. 113–114) briefly discussed sample size considerations in the estimation of conditional effects. For an unreplicated $2^k$ design, each conditional main effect and each conditional interaction effect is computed from the difference of two averages, with each average computed using $2^{k-2}$ observations, with the data split on one of the factors. Clearly, this can be a problem when $k$ is small. In particular, each average is computed using only two observations when $k = 3$.

If a $2^k$ design is replicated so that an estimate of $\sigma$ can be obtained, the tables of Lynch (1993) could *not* be applied for a given split of the data. To illustrate, assume that a $2^4$ design with 2 replicates has been used. Each average computed to obtain a conditional effect estimate will be computed using 8 observations, which is the same number that would be used in computing each average that is used in obtaining a main effect estimate with an unreplicated $2^4$ design. The error degrees of freedom doesn't match those given in Lynch's table, however.

Another problem that would complicate any inferential approach is that it would be logical to consider the *largest* conditional effect of a factor in trying to determine if the factor has a real effect. For example, with a $2^4$ design and a desire to see if factor $A$ has a real effect, considering all three two-factor interaction terms involving $A$, it would be logical to split the data on the factor that with $A$ comprises the largest two-factor interaction. This selection process would have to be considered in determining an inferential procedure, as a routine approach (such as a two-sample $t$-test) won't work.

If a design is not replicated, a normal probability plot approach will generally be used in determining significant effects. Since conditional effects and regular effects have different variances, it is not possible to construct a normal probability plot that contains a mixture of the two, since variances of all of the plotted effect estimates must be equal. There is also the matter of having only one degree of freedom for regular effect estimates, and we don't generally split degrees of freedom to create fractional degrees of freedom.

One approach that might be used, however, when there are large two-factor interactions, would be to compute all the possible conditional main effects and conditional two-factor interaction effects, then compute the numerical difference between those numbers and the corresponding regular effect estimates. The largest difference for each factor and for two-factor interactions would then be used in a normal probability plot to try to identify significant effects. This is just an ad hoc approach, however, as the differences, while having the same variance, would be computed using different subsets of the data, so no formal inferential procedure could be used. Nevertheless, such a normal probability plot could be very illuminating for certain datasets.

Because of these complications, it would probably be best to view conditional effects analyses as an exploratory data analysis (EDA) tool, but to ensure that there is a sufficient number of observations for estimating both standard effects and conditional effects.

## 6.7 GENERAL FACTORIAL DESIGNS

Although full and fractional factorial two-level designs are the types of factorial designs that are most frequently used in practice, there is often a need to use a design that has more than two levels. Sample size determination for designs that have at least one factor with more than two levels can be performed using Lenth's applet or MINITAB.

To illustrate, with Release 16 of MINITAB the user specifies the number of levels for each factor, the maximum difference between main effect means that is to be detected with the desired power, $\alpha$, a value for $\sigma$, and the highest-order term that is to be included in the model, and the software gives the number of replicates that is required.

■ **EXAMPLE 6.5**

For example, assume that a $3^2$ design is to be used, $\alpha = .05$, the second-order term (the *AB* interaction) is to be included in the model, a term for replicates is *not* to be included in the model (there is an option for that), the maximum difference between treatment means is 4, $\sigma$ is assumed to be 2, and the desired power is .90. MINITAB gives 3 as the number of replicates, with the assumed power value being .946. (This is not close to .90 because the error degrees of freedom jumps by 9 for every additional replicate, so the power numbers will also have large jumps.)

It should be noted that this result assumes fixed factors, and the power is for main effects, not the interaction term. In general, the power for main effects and interactions will differ for designs with more than two levels. When the number of levels differ, MINITAB determines the sample size based on the main effect (i.e., factor) with the largest number of levels, so as to provide conservative results. It is also important to note that the result for this example assumes that the two factors are fixed, as stated, and that *F*-tests are being performed to determine significance. The situation is more complex when at least one factor is random because then both factors are not tested against the error term, as they are when the factors are fixed.

When both factors in a two-factor design are random, the significance of each factor depends on both $\sigma$ and the standard deviation of the interaction term, $\sigma_{AB}$, since each main effect is tested against the interaction term when the *F*-tests for the main effects are performed. In general, for fixed values of $\sigma$ and $\sigma_{AB}$, the power is typically much less when both factors are random than when both factors are fixed. This is because power is determined in part by the denominator degrees of freedom for the *F*-test, and whereas replicates increase the degrees of freedom for the error term, they have no effect on the degrees of freedom for the interaction term.

Lenth's applet can also be used for determining sample size for a general factorial design, although the information that is inputted differs, as $\sigma$ and the standard deviations of the effect estimators must be entered. ■

It will often be impractical to use very many replicates of a general $s \times t$ design unless $s$ and $t$ are small. In the preceding example, the number of design points increased by 9 for every additional replicate. For a $4 \times 4$ design, the number of design points will increase by 16 for every additional replicate, so an impractically large size for the experiment could easily be reached.

## 6.8 REPEATED MEASURES DESIGNS

Repeated measures designs are designs for which multiple measurements are made, over time, on the same subjects. Such designs have both advantages and

disadvantages. Assume that the subjects are people. One advantage is that the people in the study serve as their own control, which permits the use of a smaller sample size than would otherwise be the case. A disadvantage is that there is the potential for carryover effects when two or more treatments are applied over time to the same people. When there is a strong possibility of carryover effects, designs for carryover effects should be used (see Ryan, 2007, p. 432).

Repeated measurements may be made over time without any treatments being applied, however. Vickers (2003) addressed the question of how many measurements to take in medical applications, such as how often patients should be evaluated after thoracic surgery.

Most sample size determination software and freeware will not handle repeated measures designs explicitly; that is, the design is not a design that is selected from a menu. PASS is an exception, however, as it has extensive capability for repeated measures designs. Specifically, if Means is selected from the menu, 67 procedures are listed, with two of them being `Repeated Measures ANOVA` and `Tests for Two Means in a Repeated Measures Design`.

"Before" and "after" measurements on the same subjects with the data analyzed as a test with paired (dependent) samples (Section 3.10) is a type of repeated measures design but a paired-$t$ test is for only two levels of a single factor, so a more general method of analysis must be used for more levels and/or more factors.

We will first consider Example 3.5 that was used in Section 3.10. We will assume the same mean difference, correlation between the two measurements, and standard deviations, but that was a one-tailed test with $\alpha = .05$ whereas an $F$-test, as used with a repeated measures design, is a one-tailed test that sums the two tail areas for a $t$-test. Thus, we will need to convert Example 3.5 to a two-tailed test. Another necessary adjustment is that Example 3.5 used $z$-variates, whereas $t$-variates must be used in order to obtain the same sample size for an $F$-test since $(t_\nu)^2 = F_{1,\nu}$, with $\nu$ denoting the degrees of freedom for the $t$-statistic and 1 and $\nu$ denoting the numerator and denominator degrees of freedom, respectively, for the $F$-statistic. As in Example 3.5, we will let $\mu_d = 1$ and let $\sigma_1 = \sigma_2 = 5$, with the covariance being $\rho_{12}\sigma_1\sigma_2 = .8(5)(5) = 20$.

Using PASS and selecting `Repeated Measures ANOVA` with `n-Regular F Test` to be solved for and `K (Means Multipliers)` set to the default value of 1.0, entries would be made in the input section `Within-Subject Repeated Factors`, indicating, in turn, two levels for "W," $\alpha = .05$, power = .80, and any pair of mean values specified in the last part of that line such that the difference in the means is 1.0. The user would then click on the `Covariance` tab `Covariance Matrix Columns` and select `2) Covariance Matrix Columns` and click the `Input Spreadsheet` icon at the bottom of the page after entering `C1 C2` in that section to specify the columns that contain the covariance matrix. The $2 \times 2$ matrix would then be entered in the spreadsheet using the first two rows and first two columns, with the entries being 25, 20, 20, and

25 going across the two columns for the first row and then doing the same for the second row. Then specifying "Step 0 l" for `Time Metric` results in a sample size of 81, so that two measurements would be made on each of 81 subjects.

It may be instructive to see how this relates to Example 3.5. We will, however, need to make the adjustments mentioned earlier in this section. Specifically, $t_{80,.025} = 1.99006$ and $t_{80,.20} = 0.846137$. When these numbers are used in determining the sample size, as in Example 3.5, the result is $n = [(1.99006 + 0.846137)\sqrt{10}]^2 = 80.44$, so $n = 81$ would be used.

Determining sample size when there is more than one factor is potentially complicated because the user must decide what assumptions are tenable regarding error terms and correlations between factors within subjects and between subjects. The assumptions that can be made will determine how the data should be analyzed, and the method of analysis *should* determine the sample size, but the way to proceed may not be determinable until data have been obtained, which would then be used in testing certain assumptions.

A strategy for approaching the analysis of repeated measures data was given in Section 8.4 of Littell, Stroup, and Freund (2002). This section can be read online at `http://faculty.ucr.edu/~hanneman/linear_models/c8.html#8.4`.

Regardless of how one proceeds, the covariance structure must be specified and estimates of the variances and covariances must be provided, with the structure and estimates being factors that greatly influence the sample size, as the reader will see in working Exercise 6.12. See Littell, Pendergast, and Natarajan (2000) and Milliken (2004) for guidance in determining the covariance structure. The latter is a key step in determining both the sample size and how the data will be analyzed.

Repeated measurements are involved in pharmacokinetic and pharmacodynamic experiments and a moderate amount of research has been performed in determining sample size for such experiments. See, in particular, Ogungbenro and Aarons (2010a,b). Ogungbenro and Aarons (2008, 2009) and Ogungbenro, Aarons, and Graham (2006) may also be of interest in regard to such experiments.

Overall and Doyle (1994) provided sample size formulas for repeated measurement designs and Overall (1996) used simulation to examine the effects of different sample sizes. Other papers that may be of interest include the following. Lipsitz and Fitzmaurice (1994) considered sample size for repeated measures studies with a binary response variable and Jiang and Oleson (2011) considered sample size determination for repeated measures studies with multinomial outcomes. See also Muller and Barton (1989), Muller, LaVange, Ramey, and Ramey (1992), Lui and Cumberland (1992), Kirby, Galai, and Munoz (1994), Zucker and Denne (2002), Yi and Panzarella (2002), Jung and Ang (2003), and Liu and Wu (2005, 2008).

### 6.8.1 Crossover Designs

Crossover designs are a specific form of a repeated measures design. A crossover design, also called a changeover design, is one in which there is a switch, or changeover, in the treatments that the subjects receive, such as half of the subjects receiving, say, treatment *A* followed by treatment *B*, with the other half receiving treatment *B* followed by treatment *A*, with the subjects being randomly assigned to each of the two sequences.

The simplest type of crossover design is a $2 \times 2$ design—a design with two treatment sequences and two treatment periods, with the number of treatment periods equal to the number of treatments. Half of the subjects (i.e., one group) would logically receive one of the treatment sequences and the other half would receive the other treatment sequence.

Consider bioequivalence testing, as a $2 \times 2$ crossover design is often used in conjunction with bioequivalence testing. (Recall that equivalence testing was discussed briefly in Section 2.10.) Under the assumption that the upper bioequivalence limit is $0.2\mu_R$ (and the lower limit is $-0.2\mu_R$), with $\mu_R$ denoting the mean of a reference drug, Chow and Wang (2001) gave the necessary sample size for a two-period, two-sequence crossover design as

$$n = \frac{2\,(CV)^2\left(t_{\alpha,2n-2} + t_{\beta,2n-2}\right)^2}{(0.2 - |\theta|)^2} \tag{6.5}$$

Here $\theta = (\mu_T - \mu_R)/\mu_R$, with $T$ and $R$ denoting two treatments, such as a treatment drug ($T$) and a reference drug ($R$) in a drug study. The quantity $CV$, which represents "coefficient of variation," is defined as $CV = \sigma_e/\mu_R$ with $\sigma_e$ denoting the intrasubject standard deviation. Obviously, an iterative solution for $n$ is necessary since components of the fraction also contain $n$.

■ **EXAMPLE 6.6**

Assume that $\mu_T = 7, \mu_R = 6, \sigma_e = 3, \alpha = .05$, and power $= .90$, with upper bioequivalence limit $0.2\mu_R$ and lower limit $-0.2\mu_R$. Using Eq. (6.5) with the degrees of freedom for $t$ set to infinity for illustration, we obtain

$$\begin{aligned} n &= \frac{2\,(CV)^2\left(t_{\alpha,2n-2} + t_{\beta,2n-2}\right)^2}{(0.2 - |\theta|)^2} \\ &= \frac{2\,(0.5)^2\,(1.64485 + 1.28155)^2}{(0.2 - 1/6)^2} \\ &= 3853.72 \end{aligned}$$

so 3854 treatment subjects per group would be needed, and thus a total of $2(3854) = 7708$ subjects would be used. This is in agreement with the solution given by PASS, which uses the two one-sided tests of Schuirmann (1987).

If we just wanted to do a standard hypothesis test for the equality of two means using a $2 \times 2$ crossover design and still assuming $\mu_T = 7, \mu_R = 6, \sigma_e = 3$, $\alpha = .05$, and power $= .90$, but NOT in the context of equivalence testing, the required sample size will be much lower. Specifically, as can be obtained using PASS, $n = 192$ for a two-sided test with $\alpha = .05$ and power $= .90$ (actually .9014). Thus, the use of equivalence testing increases the required sample size greatly for this example.

There are other types of crossover designs, which we will designate as $k \times m$ designs, with $k$ denoting the number of sequences and $m$ denoting the number of periods.

Ahrens, Teresi, Han, Donnell, Vanden Burgt, and Lux (2001) discussed the determination of sample size when a crossover design is used, compared to what sample size would be needed if a parallel design had been used. (The latter is a design in which no subject is measured more than once. An analogy regarding the comparison of a crossover design and a parallel design would be a paired-$t$ test vis-à-vis an independent sample $t$-test.) Fewer subjects are needed for a crossover design compared to a parallel design, as we would expect.

That advantage is offset somewhat, however, by the fact that crossover designs can be undermined by carryover effects, as stated previously. Specifically, the effect of a treatment in a previous time period can affect a measurement from a different treatment on the same subject in a subsequent time period. There are crossover designs that are balanced for carryover effects, however, so this is not an insurmountable problem. Whether or not there are likely to be carryover effects depends on what the treatments are and what is being measured. Of course, one way to try to prevent carryover effects is to have a substantial amount of time between treatments.

The Ahrens et al. (2001) paper contained a description of a study that involved a two-period crossover design. They gave the formula for sample size determination as $n = \sqrt{\text{mean squared error}}/\text{dose response slope}$, with estimates of the latter obtained using the definition: (change in response)/(change in $\log_{10}$ (dose)). Note that this approach determines the sample size for *each* efficacy variable outcome measure, not for a set of them. Of course, the idea would be to look at efficacy variables that do not require large sample sizes and the authors identified three such variables. The required sample sizes for the three were given by the authors as 23, 25, and 37, respectively, whereas the authors stated that corresponding sample sizes from "otherwise identical parallel studies" would be 657, 1438, and 2261—obviously orders of magnitude larger.

Chow, Shao, and Wang (2008) also discussed the problem more generally and gave the sample size for a crossover design relative to what the sample size would

be for a parallel group design if the latter were used, keeping $\alpha$ and the power constant. That formula is

$$n_{\text{crossover}} = \frac{n_{\text{parallel}}}{2\,(1-\theta)}$$

with $\theta$ defined as

$$\theta = \frac{\gamma}{\left(z_{\alpha/2} + z_{\beta}\right)^2}$$

with $\gamma$ defined as $\gamma = \sigma_p^2/\sigma^2$, with $\sigma_p^2$ denoting the variance of the random period effect and $\sigma^2$ being the variance of the response variable. Thus, the crossover design will require approximately half the sample size of a parallel group design when the period effect is negligible, but there might not be any reduction at all if there is a substantial period (carryover) effect. This shows mathematically one reason why it is desirable to avoid carryover effects when a crossover design is used, which can often be accomplished by allowing sufficient time between treatment periods.

See Chen, Chow, and Li (1997) and Qu and Zheng (2003) for information on sample size determination for higher-ordered crossover designs used in bioequivalence studies, and see also Potvin, DiLiberti, Hauck, Parr, Schuirmann, and Smith (2007), and Montague, Potvin, DiLiberti, Hauck, Parr, and Schuirman (2012), with the latter being a follow-up to the former. ■

#### *6.8.1.1 Software*

Software with capability for determining sample sizes for crossover designs include nQuery, SiZ, and PASS, with the latter having extensive capabilities and the user being able to specify one of the following crossover designs for sample size determination: $2 \times 2$, $2 \times 3$, $2 \times 4$, $4 \times 2$, and $4 \times 4$. By comparison, SiZ and nQuery can be used only for a $2 \times 2$ crossover design. Similarly, the user-written Stata command `xsampsi` can be used to determine sample size for a two-period, two-treatment crossover design. MINITAB does not have capability for a crossover design.

The advantage of using a $4 \times 2$ design (a Balaam design; Balaam, 1968) relative to a $2 \times 2$ design is that the two extra sequences of the former are sequences in which a subject receives the same treatment in the two time periods, which permits an assessment of within-subject variability, assuming no carryover effects. As with repeated measures designs, in general, a crossover design should be selected from the many that are available that perform well under a variety of conditions, including model misspecification.

We will consider a simple example to illustrate the use of PASS for a $2 \times 2$ crossover design. This will illustrate why caution should be used when using certain software.

■ **EXAMPLE 6.7**

When PASS is used for a $2 \times 2$ crossover design, the user specifies the difference between the two treatment means under the null hypothesis and must specify either the within-subject standard deviation (i.e., the standard deviation if repeated measurements were obtained when both the subject and treatment are constant) the standard deviation of paired difference or the standard deviation of the difference in the two treatments for a given subject.

If we specify the difference between means as 0.5, the within-subject standard deviation as 1.0, the power as .90, and $\alpha = .05$ for a two-sided test with the difference of the treatment means equal to zero under the null hypothesis, the required sample size is 88, as is obtained using PASS (Release 11).

For a difference, $d$, and within subject standard deviation $S_w$, the correct expression is $s_d^2 = 2s_w^2$, using the result that the variance of a difference of two independent random variables is the sum of the variances. Thus, for this example, $s_d^2 = 2s_w^2 = 2(1)^2 = 2$, so $s_d = \sqrt{2} = 1.414$. Therefore, if the user of PASS prefers to enter this standard deviation (which PASS labels "SdPaired = Std Dev of Paired Differences"), the value entered should be 1.414, after specifying that it is this type of standard deviation that will be entered. Doing so produces the same sample size, $n = 88$, as obtained previously when the within-subject standard deviation of 1.0 was specified. ■

Other software that can be used for sample size determination includes Cytel Studio, Version 8. A user enters the treatment difference that is to be detected ($\delta$), values for $\sigma$ and $\alpha$, whether the test is one-sided or two-sided, and the desired power, and the sample size is computed. For example, with $\delta = 1.5$, $\sigma = 4$, $\alpha = .05$, power $= .80$, and a two-sided test, the software gives 114 as the sample size. Alternatively, power can be examined for a range of sample sizes, such as 10 to 100 in increments of 1, or the necessary sample size could be displayed for a range of power values.

## 6.9 RESPONSE SURFACE DESIGNS

As stated near the beginning of the chapter, more than one design is generally used when a response surface analysis is performed, although Cheng and Wu (2001) and Bursztyn and Steinberg (2001) have proposed that the factor screening that is generally performed in the first stage be combined with the optimization that is used in the second stage and both be accomplished with a single design.

If a traditional approach were taken using two designs, the first design would be used to identify significant first-order effects; the second design would be used to identify the form of the response surface, which is one way of saying that the objective is to identify the form of a suitable model once the important

factors have been identified; then the third stage would be to perform experiments to seek the optimum combinations of levels of the important factors that have been identified.

## 6.10 MICROARRAY EXPERIMENTS

Microarray experiments have been extremely popular for several years and are being performed in large numbers, so sample size determination is important, especially considering the cost of microarray experiments. These experiments present special experimental design challenges because classical experimental designs are not useful in microarray experiments.

Hwang, Schmitt, Stephanopoulos, and Stephanopoulos (2002) pointed out that there is a tendency to perform only a small number of microarray measurements because such measurements are costly, with the consequence that in many cases the number of measurements is inadequate. They addressed this problem by proposing a method that utilizes some discriminant analysis methodology. Their method does not involve a sample size expression, however, because the method is iterative, with the original sample size increased until the (retrospective) power is $1 - \beta$. Thus, it involves retrospective power, which has been criticized by Lenth (2001), in particular, as stated previously.

Lin, Hsueh, and Chen (2010) stressed the importance of using the correlation and effect size heterogeneity between genes in determining sample size and proposed a permutation method for determining sample size.

Dobbin and Simon (2005) stated that the complexity of microarray experiments and the large amount of data that they produce can make the sample size issue seem daunting and tempt researchers to use rules-of-thumb rather than formal calculations. They sought to overcome that problem by presenting formulas "for determining sample sizes to achieve a variety of experimental goals." Specifically, they presented eight formulas, with individual formulas for determining sample size for (1) the number of single-label microarrays to use, (2) the number of microarrays required when comparing two classes in a reference design with no technical replicates, (3) the number of samples required for dye-swap arrays, (4) the number of arrays required when $m$ technical replicates of each sample are to be performed, (5) the number of arrays required for a balanced block design, (6) the number of arrays required for a balanced paired design with no dye-swap arrays, (7) the number of arrays required for a dye-swap paired design, (8) the number of arrays required for predictive and prognostic markers, and (9) the sample size for a reference design.

They emphasized that the sample size formulas that they presented are not for small samples. The reason for this, as they stated, is their formulas assume known parameter values and poor, unreliable estimates could result when small samples are used.

Their formulas are based on the assumption of a normal distribution but their simulations showed that the usefulness of their formulas is relatively insensitive to that assumption.

Because of the considerable research interest in microarray experimentation, many other papers have been written on sample size determination. See also Lee and Whitmore (2002), Pan, Lin, and Lee (2002), Wang and Chen (2004), Wei, Li, and Bumgarner (2004), Müller, Parmigiani, Robert, and Rousseau (2004), Tsai, Wang, Chen, and Chen (2005), Jung, Bang, and Young (2005), Pounds and Cheng (2005), Li, Bigler, Lampe, Potter, and Peng (2005), Pawitan, Michiels, Koscielny, Gusnanto, and Ploner (2005), Liu and Hwang (2007), Matsui, Zeng, Yamanaka, and Shaughnessy (2008), Matsui and Oura (2009), and Hirakawa, Hamada, and Yoshimura (2011) for additional information on determining sample sizes in microarray experiments.

### 6.10.1 Software

Software that can be used to determine sample size for microarray experiments includes PASS, which has capability for (1) a one-sample or paired-$t$ test for microarray data, and (2) a two-sample $t$-test. A program in R for sample size estimation is described by Orr and Liu (2009).

## 6.11 OTHER DESIGNS

The sample size determination capability for experimental designs using Lenth's applet covers many other types of designs and generally covers any design for which the model can be written. This gives it greater capability than sample size determination software that has only programs for specific designs. We look at several other designs in the following sections, with Lenth's applet having a template for each of these designs.

### 6.11.1 Plackett–Burman Designs

Plackett–Burman (PB) designs are two-level designs with the number of runs not a power of two. More specifically, PB designs exist for 12, 20, 24, and 28 runs, up to 100 runs. MINITAB can be used to solve for the sample size for these designs. For example, assume a 12-point PB design for four factors, with $\sigma = 1.0$, $\alpha = .05$, power $= .80$, no center points, and it is desired to detect an effect of size 1.5. (Since there are multiple factors, "effect size" means the smallest effect size—the average response at the high factor level minus the average response at the low factor level—that one wishes to detect. For this example the effect size can be viewed as $1.5\sigma$.) The solution will be in turns of the number of

replicates and for this example two replicates are required, which gives a power of .936—much higher than requested. Of course, the jumps in power will be substantial for 2, 3, 4 reps, and so on, and are .936, .992, and .999, respectively.

MINITAB computes the power for a PB design as follows. The computation involves the noncentrality parameter when this design is used, which is $\lambda = rc\delta^2/4\hat{\sigma}_e^2$, with $r$ = number of replicates, $c$ = number of corner points in design, $\delta$ = effect, and $\hat{\sigma}_e^2$ is the estimated variance. The power is then computed as Power $= 1 - F(f_{\alpha,1,\nu}, 1, \nu, \lambda)$, with $\nu$ denoting the error degrees of freedom and $f_\alpha$ denoting the value for the central $F$-distribution with 1 degree of freedom for the numerator and $\nu$ for the denominator, and for which the right-tail area is $\alpha$. Therefore, for a 12-point PB design with $\hat{\sigma}_e^2 = 1^2 = 1$, $\delta = 1.5$, $c = 12$, and $r = 1$, $\lambda = rc\delta^2/4\hat{\sigma}_e = (1)(12)(1.5)^2/4(1) = 6.75$. With all main effects assumed to be in the model, $\nu = 7$, so $f_{.05,1,7} = 5.59145$. If an unreplicated 12-point PB for four factors is used, the power is then $1 - F(f_{\alpha,1,\nu}, 1, \nu, \lambda) = 1 - F(5.59145, 1, 7, 6.75) = 1 - .3915 = .6085$. Thus, the unreplicated design has low power for detecting an effect of size $1.5\sigma$. For two replicates, $\lambda = 13.50$, and $\nu = 19$, $f_{.05,\,1,\,19} = 4.38075$, so Power $= 1 - F(4.38075, 1, 19, 13.50) = 1 - .0639 = .9361$.

Thus, for a 12-point PB design for four factors, the power is too low for detecting a $1.5\sigma$ effect if the design is unreplicated, but there is a big jump in power to .936 when two replicates are used. This means that if we were to solve for the number of replicates needed to obtain power of .80, the answer would be close to 1.5. Although such a solution would not be of much practical value, the number of replicates in general would be obtained by iteratively solving for $\lambda$ such that $F(f_{\alpha,1,\nu}, 1, \nu, \lambda) = .20$, then solving for $r$ from the expression $\lambda = rc\delta^2/4\hat{\sigma}_e$. Although a closed-form solution for the number of replicates is not possible, Mathews (2010) gave a large-sample approximation, which is $r \geq (4/c)(z_{\alpha/2} + z_\beta)^2(\hat{\sigma}_e/\delta)^2$. For this example, this becomes $r \geq (4/12)(1.96 + 0.84)^2(10/15)^2 = 1.16$. Thus, the value of $r$ is smaller than expected and certainly a large-sample approximation won't work very well when the necessary number of observations should be under 20, as is the case here.

Roughly, if $r$ is estimated to be about 1.4, we then have $\lambda = 9.45$ and $f_{\alpha,1,\nu} = f_{.05,1,12} = 4.747$, so that $F(f_{\alpha,1,\nu}, 1, \nu, \lambda) = .19456$ and Power $= 1 - .19456 = .80544$ and is thus very close to the target value of .80. Of course, it is not possible to have 1.4 observations at each design point but it should be apparent that when we use a PB design we are seeking economy, so large sample approximations won't work very well. The approximation should work better when the required value of $r$ is at least 2. For example, as stated previously, a 12-point PB design with 2 replicates for 4 factors with $\sigma = 1.0$ and $\alpha = .05$ has Power $= .936$ for detecting an effect of 1.5. Since we know the exact value of the power, we can give the approximation some help by using that value. Doing so gives $r \geq (4/12)(1.96 + 1.52)^2(1/1.5)^2 = 1.794$, so the approximation fares somewhat better this time

but is still not close to the actual solution, although that is somewhat immaterial since only an integer number of replicates can be used.

Indeed, replicating PB designs practically defeats the purpose in using them, as they were designed to be used for studying up to $k$ factors in $N = (k + 1)$ runs, and are thus meant to be economical designs.

Although replication will usually be necessary to achieve the desired power, fortunately that won't always be the case. For example, if a PB design with 20 runs is to be used for studying three factors, power = .8824 for detecting a minimum effect of size $1.5\sigma$ *without* the design being replicated.

This result should not be surprising, however, because the number of runs is more than the number of runs for a $2^3$ design with two replicates. Interestingly, the power drops off slightly to .8796 when the design is used to study four factors and to .8763 and .8724 when the design is used to study five and six factors, respectively. Of course, eventually the power starts to drop off sharply as the number of factors is increased, but that doesn't happen until about 15 factors are used.

These designs can be useful although they have received some criticism (probably unwarranted) in the literature. [See the discussion in Ryan (2007, p. 490).] They have been used extensively in industry and biotechnology, in particular.

### 6.11.2 Split-Plot and Strip-Plot Designs

Lenth's applet has separate templates for split-plot designs and strip-plot designs, so we will consider both separately. The former might be labeled a split-unit design, which is a better descriptor because the term "split-plot" originated in agricultural applications of experimental designs, with a plot of land literally being split for the purpose of an experiment. Today, most applications of split-plot designs do not involve land. The simplest type of split-plot design is a design for an experiment with two fixed factors, each at two levels. One of these factors would be designated as the "whole-plot factor" and the other would be termed the "split-plot factor." It is necessary to make this identification, which determines the structure of the design, because the split-plot effect, also termed the subplot effect, will generally be estimated with greater precision than the whole-plot effect, although exceptions can occur (see Giesbrecht and Gumpertz, 2004, p. 169).

Consider an example with the design as explained in the preceding paragraph, with three replications to be used, which constitute the blocks for the experiment, which can be viewed as a random factor. We will use the default model given in Lenth's applet, which is, using the labels therein: Block + Whole + Block $*$ Whole + Split + Whole $*$ Split, with Block $*$ Whole serving as the error term for testing the whole-plot effect and the residual serving as the error term for testing the split-plot effect.

Under the assumption that both $\sigma$ and $\sigma_{\text{Block}*\text{Whole}} = 1$, Lenth's applet indicates that the power is only .237 for detecting a 1.5-sigma effect of the whole-plot

factor, whereas the power is .783 for detecting a 1.5-sigma effect of the split-plot factor. Thus, there is an enormous difference for this simple example and the power is unacceptable for detecting the whole-plot factor effect. The power can be increased by increasing the number of blocks (replications), but it is necessary to use 8 blocks in order to have power of at least .80 for detecting a 1.5-sigma whole-plot factor effect. Using that many blocks of course drives the power for detecting a 1.5-sigma effect of the split-plot factor to virtually 1 (.9998).

Lenth's applet also has a template for a split-plot design with one whole-plot factor and two subplot factors, and for two whole-plot factors and one subplot factor. The addition of either one whole-plot factor or one subplot factor of course reduces the degrees of freedom for the error term, thus reducing the power. Letting $W_1$ and $W_2$ denote the first and second whole-plot factors, respectively, $W_1$ is tested against Block $*$ $W_1$ and $W_2$ is tested against Block $*$ $W_2$. Under the assumption that $\sigma = \sigma_{\text{Block}*\text{W}_1} = \sigma_{\text{Block}*\text{W}_2} = 1$ and three blocks are used, the power for detecting a $1.5\sigma$ effect is .213 for each of $W_1$ and $W_2$, with the power of .9946 of detecting a $1.5\sigma$ effect of the split-plot factor. Thus, the power for the whole-plot factors is slightly less than when there is a single whole-plot factor, but the power for the split-plot factor is much greater.

For the case of three blocks, one whole-plot factor and two split-plot factors, the whole-plot factor is tested against the Block $*$ Whole interaction, and the two split-plot factors are each tested against the error term, just as the single split-plot factor is tested against the error term when there is only one split-plot factor. Using the continuing assumption that both of these standard deviations are 1.0, the power for detecting a $1.5\sigma$ effect is .8881 for each of the two split-plot factors.

A strip-plot design is used when there are multiple stages involved, such as multiple stages in a production process. As such, it is a special case of a split-plot design, which is a design for two or more stages. Lenth's applet uses this name for the design, but since blocking is involved, a better name for the design would be either a strip-block or split-block design, these being two other names for the design that have been used in the literature. The components of the model for a strip-plot design are: blocks (replications), rows, columns, blocks $\times$ columns, blocks $\times$ rows, and rows $\times$ columns, with row and column each representing a factor of interest. [For more information about the design see Giesbrecht and Gumpertz (2004, Section 7.6), who also give two illustrative examples.]

For example, assume that four blocks are used and there are three levels of each of two factors, corresponding to rows and columns, respectively. The row effect is tested against the block $*$ row interaction and we will assume that the standard deviation of the latter is 1.0. Similarly, the column effect is tested against the block $*$ column interaction and we will assume that the standard deviation of the latter is also 1.0. With these assumptions, the power of detecting a $1.5\sigma$ row

effect is .7129, which of course is the same for a $1.5\sigma$ column effect. That is a low power value, which can be increased by increasing the number of blocks (replications). Increasing the number of blocks from four to five increases each power to a much more acceptable .8615.

### 6.11.3 Nested Designs

A nested design is a design for which not all of the levels of the factors are crossed; that is, the number of treatment combinations is less than the product of all of the factor levels. A *nested factor* design has no factorial structure at all; that is, there is no crossing of levels of any of the factors. An example would be a design for three factors, A, B, and C, with B nested within A and C nested within B. The common notation for nesting is B(A), C((B|A)), and so on, and that notation will be used in this chapter.

A *nested factorial* design is a design for which there is some factorial structure within the design, such as when there are four factors at each of two levels and there is a full factorial structure in C and D at each level of B (so that the factorial structure exists for B, C, and D, but $B_1$ occurs only with $A_1$, and $B_2$ occurs only with $A_2$).

Lenth's applet handles both nested factor designs and nested factorial designs. For example, assume a single-stage nested design with two factors and factor B having four levels and each of two levels nested within A, with three replications of each factor combination. The power of detecting a $1.5\sigma$ effect of factor A is .894, and .774 for factor B(A). Both of these power values may be judged in a particular application setting, although the second number might be deemed too low. If so, the number of replications could be increased to four, which would raise the power for factor B to .924.

There is also a template for a nested design for three factors: A, B(A), and C((B|A)), with the latter denoting the levels of factor C nested within B nested within A. With three replications, there is a high power value for each of the three factors for detecting a $1.5\sigma$ effect, with the numbers being .998, .992, and .964 for A, B, and C, respectively. If two replications were used, however, the values would be .959, .883, and .723, respectively, with the last number likely being deemed too small.

Lenth's applet also has a template for a nested factorial design with one nested factor and two factors in a factorial arrangement, as well as templates for a nested factorial design with two whole-plot factors and one subplot factor and a nested factorial design with one whole-plot factor and two subplot factors.

The discussion in Section 6.5 regarding the need to have enough experimental runs so that conditional effects can be identified and estimated with reasonable precision applies to nested designs only if the design has factorial structure, as the investigation of conditional effects is motivated by large interactions, whereas nested designs without factorial structure do not have interactions.

Moerbeek, Van Breukelen, Berger, and Ausems (2003) gave an example of sample size determination for an experimental design in which individuals are nested within clusters.

### 6.11.4 Ray Designs

A ray design is a type of mixture design such that individual compounds are mixed together in amounts such that the proportion between them is constant. Casey, Gennings, Carter, Moser, and Simmons (2006) addressed sample size determination when such designs are used and the primary objective is to be able to detect interactions with reasonable power.

## 6.12 DESIGNS FOR NONNORMAL RESPONSES

Although the typical assumption is that the response variable is a continuous, normally distributed random variable, often this will not be the case. For example, the variable of interest may be the number of nonconforming units in a particular operation, with a desire to run an experiment to determine the primary causes of what might be an unacceptably high number or proportion of nonconforming units. Similarly, an experiment might be conducted to determine the causes of nonconformities, such as product blemishes.

In the first case, it would be reasonable to assume a binomial distribution for the response since there are two possible values for each unit, however defined, conforming or nonconforming, whereas a Poisson distribution might be appropriate in the second case.

Bisgaard and Fuller (1995) provided a table for determining sample size with a binary response when a $2^{k-p}$ design is to be used. They used the arcsin transformation, which is used in transforming binomial proportions to approximate normality. Other approximations were also involved, so the tabular values should be viewed as approximations, although probably good approximations. Indeed, the authors recommend that the table be used as a rough guide in determining the sample size. Of course, the cost of experimentation should also be considered. Readers interested in the theoretical development that led to the table are referred to Bisgaard and Fuller (1995).

For a given number of experimental runs, such as 16 (a common experiment size when a $2^{k-p}$ design is used), the number of units to be inspected at each of the 16 treatment combinations is determined so as to have a power of .90 of detecting a specified degree of improvement relative to the proportion nonconforming in the null hypothesis, $p_0$. To illustrate, if $p_0 = .05$ and it is desired to detect with probability .90 a 50% reduction to .025, Table Al of Bisgaard and Fuller (1995) shows that 197 units would have to be inspected at each of the 16 combinations, for a total of 3152 units. (Note the very large sample size, which is caused by a small value of $p_0$.)

Of course, software is generally easier to use than tables and a Microsoft Excel spreadsheet that solves for the sample size given inputted values for $p_0$, $\Delta$ (the change from $p_0$), $\alpha$, power, and the number of treatment combinations can be downloaded from `www.statease.com/powercalc.html`. To illustrate, assume that $p_0 = .20$ and a reduction of $\Delta = .06$ to $p = .14$ is desired to be detected with probability .90 when a $2^{k-p}$ design with 16 runs is to be used. Use of the spreadsheet shows that 117 units have to be inspected at each treatment combination. [This differs slightly from 116 units given in Table Al of Bisgaard and Fuller (1995). This difference also occurs with other tabular entries and is caused by the spreadsheet rounding *up* the numbers, whereas the table apparently rounds *off* the numbers.] Although 117 may be an acceptable number, it should be kept in mind that this is the number of units *per run*, so with a 16-run design, the total number of units to be inspected is thus 16(117) = 1872. Obviously, the total number of units to be inspected can be prohibitively large when the number of units to be inspected per run is large.

The spreadsheet is useful for various reasons, including the fact that the Bisgaard and Fuller (1995) table is only for a power of .90. Assume that 117 units in the previous example is more than can be afforded, so the experimenter is willing to settle for a power of .80. With this change 87 units would be inspected. The output from the program is given below.

```
p(bar) = 0.2 current proportion
Δ      = 0.06 minimum change in proportion

alpha (α is typically 0.05 or 0.10)   = 0.05 z-alpha/2 = 1.96
Power (1-b is typically 0.80 or more) = 0.80 z-beta    = 0.84

δ = 0.075190826 change in transformed scale

Design size                          N = 16 user input
                                         (number of runs)
Units per run for power                = 87
Units per run to avoid too many zeros = 25
Recommended Units per run;           n = 87

Total units = 1392
```

There have also been applications in which the response variable is the number of defects (e.g., Bisgaard and Fuller, 1994–1995) and the Poisson distribution will often be a logical choice for the distribution of the response variable. Unfortunately, there is apparently no software available for sample size determination for such experiments. The necessary theory would have to be developed and supported, as was done by Bisgaard and Fuller (1995) for the binomial case.

Once that was done, it would be simple and straightforward to develop an Excel spreadsheet or a MINITAB macro, for example, that would solve for the necessary sample size.

## 6.13 DESIGNS WITH RANDOM FACTORS

Sample size determination for designs with fixed factors is somewhat more straightforward and more intuitive than with random factors because it is easier to think in terms of effect sizes than sizes of variance components. Furthermore, there is very little software available for determining sample sizes with random factors. On the other hand, noncentrality parameters are not involved in testing for the effects of random factors, or, in general, testing for random effects, as interaction effects are considered to be random if at least one of the factors involved in the interaction is random.

As mentioned at the beginning of the chapter, Lenth's applet will also handle the random factor case. Specifically, the user can specify which factors are fixed and which are random for any balanced ANOVA model. The classification of a factor as fixed or random will often have a large effect on the sample size that is determined, so the classification should be done carefully and the experimenter must, of course, understand the definition of each type of factor. (A fixed factor is one for which the levels used in an experiment are the only levels of interest; a random factor is one for which interest centers on a range of possible levels, with the levels used in an experiment ideally randomly selected from the range of possible levels.)

For example, assume that there is a single factor with four levels and the factor is random. The model is then $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$, with $\tau_i \sim N(0, \sigma_\tau^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$. The null hypothesis is $H_0 : \sigma_\tau^2 = 0$. Note that since the factor is random, we are not interested in any particular $\tau_i$, including the effects of the treatment levels used in the experiment. Rather, interest is in whether *any* $\tau_i$ in the population of possible treatment levels is nonzero. If $\sigma_\tau^2 = 0$ then all of the possible $\tau_i$ must be zero.

Assume that $\sigma_\tau^2 = \sigma^2 = 1, \alpha = .05$, and an equal number of observations will be used per level. Lenth's applet shows that eight observations must be used at each level in order to have a power of at least .80 (actually .8055) of having a significant $F$-test result, whereas only five observations are needed when the factor is fixed (power = .8303). Intuitively, this makes sense because stating significance for the population of treatment levels is a stronger statement than stating significance for the particular levels used in an experiment.

It can be shown (see, e.g., Sahai and Ageel, 2000, p. 60) that

$$P\left\{F\text{-statistic} > \frac{F_{a-1,a(n-1);\alpha}}{1 + n\sigma_\tau^2/\sigma^2}\right\} = 1 - \beta$$

where $a$ denotes the number of levels and there are $n$ observations per level. Note that the value of the denominator is 1 if the null hypothesis is true, and the probability is then .05.

We can't use this expression to obtain a closed-form solution for $n$ because $n$ is part of the denominator degrees of freedom for the $F$-variate, but it does allow us to "see" the solution, and of course the expression could be used to solve iteratively for $n$, if desired.

Specifically, with $a = 4$, $\sigma_\tau^2 = \sigma^2 = 1$, $\alpha = .05$, and, as stated, Lenth's applet giving $n = 8$, the fraction is then $2.947/(1 + 8(1)) = 0.3274$. $P(F\text{-statistic} > 0.3274) = 1 - .1945 = .8055$, as can be shown using MINITAB or other software. Thus, the solution given by Lenth's applet can be verified in this manner.

## 6.14 ZERO PATIENT DESIGN

Yazici, Biyikil, van der Linden, and Schouten (2001) proposed the idea of a "zero patient" design to compare the prevalences of rare diseases in different geographical regions, although the true prevalences may remain unknown. Their objective was to determine the size of a sample that can be expected to be free of a certain disease. Their sample size formula utilized the approximation, $\ln(1 - p) \approx -p$, which works well when $p$ is small, such as $p < .02$. Utilizing this approximation, they obtained the required sample size as $n = -\ln(0.05)/p_0 = 3/p_0$, with $p_0$ denoting the hypothesized value of $p$. This is the required sample size for the 95% confidence bound $p < 3/n$. They showed by way of an example how their approach could be used in an hypothesis-testing framework to compare prevalences of a rare disease in different countries.

## 6.15 COMPUTER EXPERIMENTS

Computer experiments have become more popular during the past ten years or so, undoubtedly aided by the publication of books on the design of computer experiments such as Santner, Williams, and Notz (2003) and Fang, Li, and Sidjianto (2006). Guidance on sample size determination for such experiments has been lacking, however, as Loeppky, Sacks, and Welch (2009) stated: "Choosing the sample size of a deterministic computer experiment is important but lacks formal guidance."

Although computer experiments differ greatly from physical experiments since there are no experimental units to which treatments are assigned, there may nevertheless be constraints on the sample size in the former due to budgetary constraints that could limit the number of computer runs that could be made.

Chapman, Welch, Bowman, Sacks, and Walsh (1994) and Jones, Schonlau, and Welch (1998) used the rule-of-thumb that the number of runs being ten times the number of important input factors is a good initial guess. Loeppky et al. (2009) examined this recommendation and concluded that it is quite reasonable.

## 6.16 NONINFERIORITY AND EQUIVALENCE DESIGNS

This topic is covered last because although the terms *noninferiority design* and *equivalence design* are used in the literature, this is practically a misnomer. This is because these are not really types of experimental designs but rather "noninferiority" and "equivalence" refer to forms of hypothesis tests. Greene, Morland, Durkalski, and Frueh (2008) essentially indicated this as they stated: "Noninferiority designs are a one-sided test used to determine if a novel intervention is no worse than a standard intervention. Equivalence designs, a two-sided test, pose a similar question, but also allow for the possibility that the novel intervention is no better than the standard one." Thus, these are designs only if the term is used in a very broad sense and is not intended to refer to statistical designs. Even worse, Greene et al. (2008) stated that many studies in which noninferiority or equivalence was claimed were from studies that were not designed in such a way as to allow such conclusions to be drawn, such as claiming equivalence when a null hypothesis of superiority was not rejected.

## 6.17 PHARMACOKINETIC EXPERIMENTS

Pharmacokinetic experiments are performed to compare the effects of several drugs applied to several individuals. Aarons and Ogungbenro (2010) stated: "Despite the need to include sample size calculation in the design of population pharmacokinetic experiments . . . most population pharmacokinetic experiment designs still do not include a sample size calculation." There are various ways of determining sample size for such experiments and these are reviewed by Aarons and Ogungbenro (2010).

## 6.18 BAYESIAN EXPERIMENTAL DESIGN

Although Bayesian experimental design is a well-established field [see, e.g., the review paper by Chaloner and Verdinelli (1995)], with Kathryn Chaloner being one of the major contributors to the field, relatively little research has been published on sample size determination using Bayesian experimental design methods. One such paper is Goldstein and Wooff (1997), who used a Bayes linear approach for sample size determination.

## 6.19 SOFTWARE

Using Design-Expert to determine power for $2^k$ designs was discussed in Sections 6.4.2 and 6.4.3.

Although Power and Precision is useful for determining sample size and power, it has very limited capability for experimental designs, as its capability is limited to one-factor designs, with and without a covariate, and factorial designs with either two or three factors, with or without a covariate.

This contrasts sharply with the capability of Lenth's applet, which can handle any balanced ANOVA model, and has templates for nested factor and factorial designs, crossover designs, split-plot designs, randomized complete block designs, and Latin square and Graeco-Latin square designs. It is almost certainly the best freeware for determining sample size (meaning replicates, usually) for experimental designs.

In general, there are a variety of software from which to select for determining sample size and computing power for experimental designs. Capability varies greatly among freeware, however, with apparently no freeware coming close to the capability of Lenth's applet. For example, Web Power at `http://www.math.yorku.ca/SCS/Online/power` runs a SAS program for calculating power for factorial designs. There is some lack of flexibility, however, because all factors are assumed to be fixed and power can be calculated only for standardized effects sizes (effect size/$\sigma$) of 0.25, 0.50, 0.75, 1.00, and 1.25, with the program help file indicating that this choice of values (all of which could be specified, then the output would be a table) was influenced by Cohen (1988).

G*Power3 is downloadable freeware that has strengths in certain areas. Its capability for experimental design is somewhat limited, however, because it will handle only fixed effects. Furthermore, its default value for effect size is also obviously influenced by Cohen (1988), and it also restricts the user to the choice of 0.25, 0.50, 0.75, 1.00, or 1.25 for the standardized effect size. One potentially nice feature of the software is that it draws two curves for $\beta$ and $\alpha$, but this is shown as a function of values of the $F$-statistic, which means that someone learning the subject would have to convert the $F$-values to sample sizes.

MINITAB users may be interested in the MINITAB macro for computing power for fixed effects balanced designs given by Paul Mathews and available at `http://users.stargate.net/~pmathews/tot/source/power.mac`. This extends the capability of the MINITAB software, as in Release 16 the capability is limited to one-factor designs, replicated or unreplicated $2^k$ designs, replicated or unreplicated Plackett–Burman (PB) designs, and replicated or unreplicated general full factorial designs. Regarding the last three, there has to be some degrees of freedom for the error term, so if an unreplicated $2^k$, PB, or general full factorial design is to be used, the user must specify the number of terms that are to be omitted from the model, thereby releasing degrees of freedom to be

used for the error term. Of course, an estimate of the standard deviation must be provided in either case.

Thus, an estimate is needed for $\sigma$ to be able to solve for the power. Since the sample size is fixed with an unreplicated $2^k$, PB, or general full factorial design, there must be degrees of freedom for the error term so that hypothesis tests can be performed, with the software indicating the effect size that corresponds to the indicated power or the power for the indicated effect size.

In MINITAB, the effect size used with the PB design is the difference of means, so that the effect size is not standardized, with the user entering a value for $\sigma$. In general, the power depends only on (effect size/$\sigma$), so the specification of $\sigma$ wouldn't be necessary for determining power for the standardized effect size if software permitted the input of a standardized effect size, such as G*Power permits, as mentioned previously. This is not how the sample size determination capability of MINITAB is constructed, however. The software can also be used for determining the minimum effect size that can be detected for a specified power. Given below is the output which illustrates this, as the output includes the effect size when the user inputs everything except the effect size.

```
Power and Sample Size

Plackett-Burman Design

Alpha = 0.05 Assumed standard deviation = 10

Factors:           15    Design: 20
Center pts (total): 0

Center       Total
Points  Reps  Runs   Power   Effect
     0     1    20     0.8  16.8200
```

It is also useful to see a graph of power plotted against effect size for the specified design, which is also part of the MINITAB output and is shown in Figure 6.4.

PASS is the most comprehensive sample size determination software and is also probably the easiest to use, although its capabilities for experimental designs are limited. (The software can be used to construct a variety of designs, but can be used to determine sample size for only a limited number of designs, as was noted in this chapter.)

nQuery is comparable in stature to PASS and has a large number of users, but it isn't quite as user friendly. For example, once a set of inputs has been entered, no component can be changed to see that effect without entering all of the components of the worksheet, or cutting and pasting. Like PASS, it really isn't software for determining sample size for experimental designs.

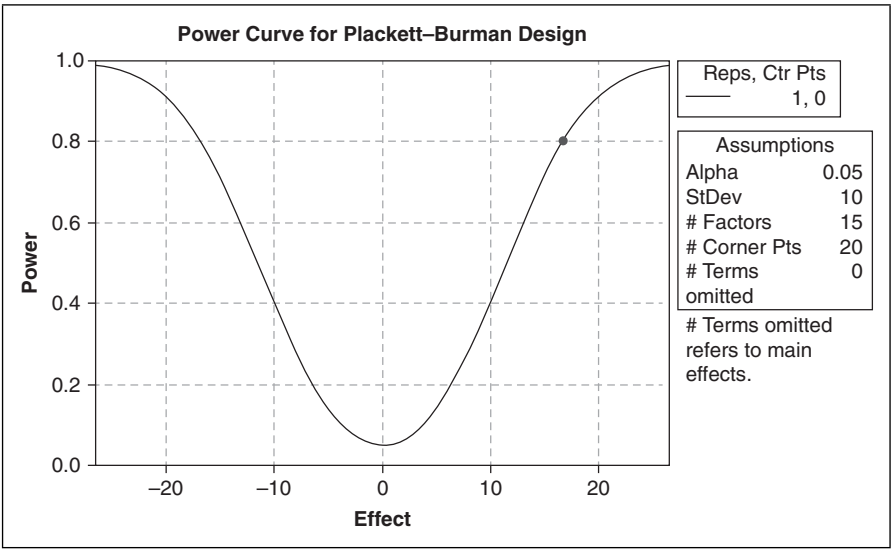


**Figure 6.4** Power curve for 20-point Plackett–Burman design; effect size $= 1.5\sigma$.

This can undoubtedly be explained for PASS and nQuery and perhaps for other software as well by recognizing what the primary markets are for their software (e.g., pharmaceutical firms) and recognizing that these markets are not going to be interested in determining sample size for factorial designs or Taguchi designs, which are used in manufacturing applications. Rather, interest centers on a rather limited collection of designs.

Users of SAS Software can use PROC GLMPOWER for determining sample size for linear models. See `http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_glmpower_a0000000154.htm`. The output for, say, an ANOVA example with two factors has a realistic general form in that a single sample size is not given. Rather, as shown at `http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_clientpss_sect025.htm` for an example involving the two factors `Drug` and `Gender` the output states in part "You need a total sample size between 60 and 108 to yield a power of 0.9 for the `Drug` effect if the standard deviation is between 5 and 7. You need a sample size of half that for the `Gender` effect."

## 6.20 SUMMARY

Sample size determination for designed experiments deserves careful attention for the same reasons that sample size determination in general is important. Books on sample size determination in general have very little on sample size

determination and the need to use multiples of certain designs, such as Latin square designs, is generally not stressed. More attention also needs to be devoted to sample size determination by statistical software companies. Until then, tables such as those given by Bratcher, Moran, and Zimmer (1970), which is for a single factor with and without blocking, can be useful, as can the tables of L. S. Nelson (1985), which are an extension of the Bratcher et al. (1970) tables. See also the extended tables in Diletti, Hauschke, and Steinjans (1992) for testing bioequivalence with a two-period crossover design. The tables provided in Hager and Möller (1986) may also be of interest. Odeh and Fox (1991) gave many charts to aid in sample size determination for experimental designs.

## APPENDIX

### Derivation of Eq. (6.1)

We will first derive the formula for a one-sided test, then arrive at the formula for the two-sided test using deduction. Assume that the alternative hypothesis is "greater than." Then

$$p\,(z < z_\alpha) = \Phi\left(z_\alpha - \frac{\Delta}{\sigma\sqrt{2/m}}\right) = \beta$$

with $m$ denoting the common sample size and $\Delta$ denoting the difference between the two means, so

$$\begin{aligned} z_\alpha - \frac{\Delta}{\sigma\sqrt{2/m}} &= \Phi^{-1}(\beta) \\ &= z_\beta \quad \text{(since the subscript of } z \text{ denotes the right tail area)} \\ &= -z_\beta \end{aligned}$$

so

$$\frac{\Delta}{\sigma\sqrt{2/m}} = -z_\alpha - z_\beta$$

and thus

$$m = \frac{\left(z_\alpha + z_\beta\right)^2 2\sigma^2}{\Delta^2} = \frac{n}{2} \text{ in Eq. (6.1)}$$

The only component of this expression that is affected by whether the test is one-sided or two-sided is $z_\alpha$, which is $z_{\alpha/2}$ if the test is two-sided. Therefore, that substitution gives the formula for $m$ for a two-sided test.

## REFERENCES


Aarons, L. and K. Ogungbenro (2010). Optimal design of pharmacokinetic experiments. *Basic and Clinical Pharmacology and Toxicology*, **106**(3), 250–255.

Ahrens, R. C, M. E. Teresi, S.-H. Han, D. Donnell, J. A. Vanden Burgt, and C. R. Lux (2001). Asthma stability after prednisone: A clinical model for comparing inhaled steroid potency. *American Journal of Respiratory and Critical Care Medicine*, **164**, 1138–1145.

Balaam, L. N. (1968). A two-period design with $t^2$ experimental units. *Biometrics*, **24**(1), 61–73.

Bang, H., S.-H. Jung, and S. George (2005). Sample size calculation for simulation-based multiple-testing procedures. *Journal of Biopharmaceutical Statistics*, **15**(6), 957–967.

Bisgaard, S. and H. T. Fuller (1994–1995). Analysis of factorial experiments with defects or defectives as the response. *Quality Engineering*, **7**(2), 429–443.

Bisgaard, S. and H. T. Fuller (1995). Sample size estimates for $2^{k-p}$ designs with binary responses. *Journal of Quality Technology*, **27**(4), 344–354.

Bishop, T. A. and E. J. Dudewicz (1978). Exact analysis of variance with unequal variances: Test procedures and tables. *Technometrics*, **20**, 419–430.

Borm, G. F., J. Fransen, and W. A. J. G. Lemmens (2007). A simple sample size formula for analysis of covariance in randomized clinical trials. *Journal of Clinical Epidemiology*, **60**, 1234–1238.

Box, G. E. P., W. G. Hunter, and J. S. Hunter (1978). *Statistics for Experimenters.* New York: Wiley. (The current edition is the 2nd edition, 2005.)

Bratcher, T. L., M. A. Moran, and W. J. Zimmer (1970). Tables of sample sizes in the analysis of variance. *Journal of Quality Technology*, **2**(3), 156–164.

Bursztyn, D. and D. M. Steinberg (2001). Rotation designs for experiments in high bias situations. *Journal of Statistical Planning and Inference*, **97**, 399–414.

Carroll, R. J. and D. B. H. Cline (1988). An asymptotic theory for weighted least squares with weights estimated by replication. *Biometrika*, **75**, 35–43.

Casey, M., C. Gennings, W. Carter, V. Moser, and J. Simmons (2006). Power and sample size calculations for linear hypotheses associated with mixtures of many components using fixed-ratio ray designs. *Environmental and Ecological Statistics*, **13**, 11–23.

Chaloner, K. and I. Verdinelli (1995). Bayesian experimental design: A review. *Statistical Science*, **10**(3), 273–304.

Chapman, W. L., W. J. Welch, K. P. Bowman, J. Sacks, and J. E. Walsh (1994). Arctic sea ice variability: Model sensitivities and a multidecadal simulation. *Journal of Geophysical Research*, **99**, 919–935.

Chen, K. W., S. C. Chow, and G. Li (1997). A note on sample size determination for bioequivalence studies with higher-order crossover designs. *Journal of Pharmacokinetics and Biopharmaceutics*, **25**(6), 753–765.

Cheng, S.-W. and C. F. J. Wu (2001). Factor screening and response surface exploration. *Statistica Sinica*, **11**, 553–580. Discussion: **11**, 581–604.

Chow, S.-C. and J. P. Liu (1999). *Design and Analysis of Bioavailability and Bioequivalence Studies*: New York: Marcel Dekker.

Chow, S.-C. and H. Wang (2001). On sample size calculation in bioequivalence trials. *Journal of Pharmacokinetics and Pharmacodynamics*, **28**(2), 155–169.

Chow, S.-C., J. Shao, and H. Wang (2008). *Sample Size Calculations in Clinical Research*, 2nd edition. Boca Raton, FL: Chapman & Hall/CRC.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. New York: Routledge Academic.

Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*. New York: Wiley.

Diletti, E., D. Hauschke, and W. V. Steinjans (1992). Sample size determination: Extended tables for the multiplicative model and bioequivalence ranges of 0.9 to 1.11 and 0.7 to 1.43. *International Journal of Clinical Pharmacology, Therapy, and Toxicology*, **30**, Supplement 1, S59–S62.

Dobbin, K. and R. Simon (2005). Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, **6**(1), 27–38. (Available at `http://biostatistics.oxfordjournals.Org/cgi/reprint/6/1/27.pdf`.)

Fang, K.-T., R. Li, and A. Sudjianto (2006). *Design and Modeling for Computer Experiments.* New York: Chapman and Hall/CRC.

Geng, S. and F. J. Hills (1978). A procedure for determining numbers of experimental and sampling units. *Agronomy Journal*, **70**, 441–444.

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

Goldstein, M. and D. A. Wooff (1997). Choosing sample sizes in balanced experimental designs: A Bayes linear approach. *The Statistician*, **46**(2), 167–183.

Gravetter, F. J. and L. B. Wallnau (2009). *Statistics for the Behavioral Sciences*, 8th edition. Belmont, CA: Wadsworth.

Greene, C. J., L. A. Morland, V. L. Durkalski, and B. C. Frueh (2008). Noninferiority and equivalence designs: Issues and implications for mental health research. *Journal of Trauma Stress*, **21**(5), 433–439.

Guo, J. H. and W.-M. Luh (2010). On sample size calculation for $2 \times 2$ fixed effect ANOVA when variances are unequal. *British Journal of Mathematical and Statistical Psychology*, **62**, 417–425.

Hager, W. and H. Möller (1986). Tables for the determination of power and sample size in univariate and multivariate analyses of variance and regression. *Biometrical Journal*, **28**, 647–663.

Hamada, M. and N. Balakrishnan (1998). Analyzing unreplicated factorial experiments: A review with some new proposals. *Statistica Sinica*, **8**(1), 31–35.

Harris, M., D. G. Horvitz, and A. M. Mood (1948). On the determination of sample sizes in designing experiments. *Journal of the American Statistical Association*, **43**, 391–402.

Hirakawa, A., C. Hamada, and I. Yoshimura (2011). Sample size calculation for a regularized *t*-statistic in microarray experiments. *Statistics and Probability Letters*, **81**, 870–875.

Hwang, D., W. A. Schmitt, George Stephanopolous, and Gregory Stephanopolous (2002). Determination of minimum sample size and discriminatory expression patterns in microarray data. *Bioinformatics*, **18**(9), 1184–1193. (Available at `http://bioinformatics.oxfordjournals.Org/cgi/reprint/18/9/1184.pdf`.)

Jaech, J. L. (1969). The Latin square. *Journal of Quality Technology*, **1**(4), 242–255.

Jiang, D. and J. J. Oleson (2011). Simulation study of power and sample size for repeated measures with multinomial outcomes: An application to sound direction identification experiments. *Statistics in Medicine*, **30**(19), 2451–2466.

Jones, D. R., M. Schonlau, and W. J. Welch (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, **13**, 455–492.

Jung, S.-H. and C. Ang (2003). Sample size estimation for GEE method for comparing slopes in repeated measurements data. *Statistics in Medicine*, **22**(8), 1305–1315.

Jung, S. H., H. Bang, and S. Young (2005). Sample size calculation for multiple testing in microarray data analysis. *Bioinformatics*, **6**(1), 157–169.

Keppel, G. (1991). *Design and Analysis: A Researcher's Handbook*, 3rd edition. Englewood Cliffs, NJ: Prentice Hall.

Kirby, A. J., N. Galai, and A. Munoz (1994). Sample size estimation using repeated measurements on biomarkers as outcomes. *Controlled Clinical Trials*, **15**(3), 165–172.

Lachenbruch, P. A. (1988). A note on sample size computation for testing interactions. *Statistics in Medicine*, **7**, 467–469.

Lee, M.-L. T. and G. A. Whitmore (2002). Power and sample size for DNA microarray studies. *Statistics in Medicine*, **21**(23), 3543–3570.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, **55**(3), 187–193. (Available at `http://www.stat.uiowa.edu/techrep/tr303.pdf`.)

Li, S. S., J. Bigler, J. W. Lampe, J. D. Potter, and Z. Peng (2005). FDR-controlling testing procedures and sample size determination for microarrays. *Statistics in Medicine*, **24**, 2267–2280.

Lin, W.-J., H.-M. Hsueh, and J. J. Chen (2010). Power and sample size estimation in microarray studies. *BMC Bioinformatics*, **11** (Supplement 1), S52.

Lipsitz, S. R. and G. M. Fitzmaurice (1994). Sample size for repeated measures studies with binary responses. *Statistics in Medicine*, **13**, 1233–1239.

Littell, R. C., J. Pendergast, and R. Natarajan (2000). Modeling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, **19**, 1793–1819.

Littell, R. C., W. W. Stroup, and R. J. Freund (2002). *SAS for Linear Models*, 4th edition. Cary, NC: SAS Institute, Inc.

Liu, P. and J. T. G. Hwang (2007). Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics*, **23**(6), 739–746.

Liu, H. and T. Wu (2005). Sample size calculation and power analysis of time-averaged difference. *Journal of Modern Applied Statistical Methods*, **4**(2), 434–445.

Liu, H. and T. Wu (2008). Sample size calculation and power analysis of changes in mean response over time. *Communications in Statistics—Simulation and Computation*, **37**(9), 1785–1798.

Loeppky, J. L., J. Sacks, and W. J. Welch (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics*, **51**(4), 366–376.

Luh, W.-M. and J.-H. Guo (2010). Developing the noncentrality parameter for calculating group sample sizes in heterogeneous analysis of variance. *The Journal of Experimental Education*, **79**(1), 53–63.

Lui, K-.J. and W. G. Cumberland (1992). Sample size requirement for repeated measurements in continuous data. *Statistics in Medicine*, **11**, 633–641.

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

Mathews, P. (2010). *Sample Size Calculations: Practical Methods for Engineers and Scientists.* Harbor, OH: Mathews, Malnar, and Bailey, Inc.

Matsui, S. and T. Oura (2009). Sample sizes for robust ranking of genes in microarray experiments. *Statistics in Medicine*, **28**, 2617–2638.

Matsui, S., S. Zeng, T. Yamanaka, and J. Shaughnessy (2008). Sample size calculations based on ranking and selection in microarray experiments. *Biometrics*, **64**, 217–226.

Milliken, G. A. (2004). Mixed models and repeated measures: Some illustrative industrial examples. In *Handbook of Statistics, Vol. 22: Statistics in Industry*, Chap. 5 (R. Khattree and C. R. Rao, eds.). Amsterdam, The Netherlands: Elsevier Science B.V.

Moerbeek, M., G. J. P. Van Breukelen, M. P. F. Berger, and M. Ausems (2003). Optimal sample sizes in experimental designs with individuals nested within clusters. *Understanding Statistics*, **2**(3), 151–175.

Montague, T. H., D. Potvin, C. E. DiLiberti, W. W. Hauck, A. F. Parr, and D. J. Schuirman (2012). Additional results for "Sequential design approaches for bioequivalence studies with crossover designs." *Pharmaceutical Statistics*, **11**(1), 8–13.

Muller, K. E. and C. N. Barton (1989). Approximate power for repeated measures ANOVA lacking sphericity. *Journal of the American Statistical Association*, **84**(406), 549–555.

Muller, K. E., L. E. LaVange, S. L. Ramey, and C. T. Ramey (1992). Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association*, **87**(420), 1209–1226.

Muller, P., G. Parmigiani, C. Robert, and J. Rousseau (2004). Optimal sample size for multiple testing: The case of gene expression micoarrays. *Journal of the American Statistical Association*, **99**(468), 990–1001.

Nelson, L. S. (1985). Sample size tables for analysis of variance. *Journal of Quality Technology*, **17**, 167–169.

Nelson, P. R. (1983). A comparison of the sample sizes for the Analysis of Means and the Analysis of Variance. *Journal of Quality Technology*, **15**(1), 33–39.

Nelson, P. R. (1985). Power curves for the Analysis of Means. *Technometrics*, **27**, 65–73.

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*. ASA-SIAM Series on Statistics and Applied Probability. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Odeh, R. E. and M. Fox (1991). *Sample Size Choice: Charts for Experiments with Linear Models*, 2nd edition. Boca Raton, FL: CRC Press.

Ogungbenro, K. and L. Aarons (2008). How many subjects are necessary for population pharmacokinetic experiments? Confidence interval approach. *European Journal of Clinical Pharmacology*, **64**(7), 705–713.

Ogungbenro, K. and L. Aarons (2009). Sample-size calculations for multi-group comparison in population pharmacokinetic experiments. *Pharmaceutical Statistics*, **9**(4), 255–268.

Ogungbenro, K. and L. Aarons (2010a). Sample size/power calculations for repeated ordinal measurements in population pharmacodynamic experiments. *Journal of Pharmacokinetics and Pharmacodynamics*, **37**(1), 67–83.

Ogungbenro, K. and L. Aarons (2010b). Sample size/power calculations for population pharmacodynamic experiments involving repeated-count measurements. *Journal of Biopharmaceutical Statistics*, **20**(5), 1026–1042.

Ogungbenro, K., L. Aarons, and G. Graham (2006). Sample size calculations based on generalized estimating equations for population pharmacokinetic experiments. *Journal of Biopharmaceutical Statistics*, **16**(2), 135–150.

Orr, M. and P. Liu (2009). Sample size estimation while controlling false discovery rate for microarray experiments using the ssize.fdr Package. *The R Journal*, **1**(1), 47–53.

Overall, J. E. (1996). How many repeated measurements are useful? *Journal of Clinical Psychology*, **52**(3), 243–252.

Overall, J. E. and S. R. Doyle (1994). Estimating sample sizes for repeated measurement designs. *Controlled Clinical Trials*, **15**(2), 100–123.

Pan, W., J. Lin, and C. T. Lee (2002). How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biology*, **3**, 0022.1–0022.10.

Pan, Z. and L. Kupper (1999). Sample size determination for multiple comparison studies treating confidence interval width as random. *Statistics in Medicine*, **18**(12), 1475–1488.

Pawitan, Y., S. Michiels, S. Koscielny, A. Gusnanto, and A. Ploner (2005). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, **18**(9), 1184–1193.

Pounds, S. and C. Cheng (2005). Sample size determination for the false discovery rate. *Bioinformatics*, **21**(3), 4263–4271. Erratum: **25**(5), 698–699.

Potvin, D., C. E. DiLiberti, W. W. Hauck, A. F. Parr, D. J. Schuirmann, and R. A. Smith (2007). Sequential design approaches for bioequivalence studies with crossover designs. *Pharmaceutical Statistics*, **7**, 245–262.

Qu, R. P. and H. Zheng (2003). Sample size calculation for bioequivalence studies with high-order crossover designs. *Controlled Clinical Trials*, **24**(4), 436–439.

Ryan, T. P. (2007). *Modem Experimental Design*. Hoboken, NJ: Wiley.

Sahai, H. and M. I. Ageel (2000). *The Analysis of Variance: Fixed, Random and Mixed Models*. Boston: Birkhaüser.

Santner, T. J., B. J. Williams, and W. Notz (2003). *The Design and Analysis of Computer Experiments.* New York: Springer.

Schuirmann, D. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, **15**(6), 657–680.

Schwertman, N. C. (1987). An alternative procedure for determining analysis of variance sample size. *Communications in Statistics—Simulation and Computation*, **16**(4), 957–967.

Senn, S. (2002). *Cross-over Trials in Clinical Research.* New York: Wiley.

Simon, R. M., E. L. Korn, L. M. McShane, M. D. Radmacher, G. W. Wright, and Y. Zhao (2004). *Design and Analysis of DNA Microarray Investigations.* New York: Springer.

Tsai, C.-A., S- J. Wang, D.-T. Chen, and J. J. Chen (2005). Sample size for gene expression microarray experiments. *Bioinformatics*, **21**(8), 1502–1508.

Vickers, A. J. (2003). How many repeated measures in repeated measures designs? Statistical issues for comparative trials. *BMC Medical Research Methodology*, **3**, 1–22.

Wang, S. J. and J. J. Chen (2004). Sample size for identifying differentially expressed genes in microarray experiments. *Journal of Computational Biology*, **11**(4), 714–726.

Wei, C., J. Li, and R. E. Bumgarner (2004). Sample size for detecting differentially expressed genes in microarray experiments. *BMC Genomics*, **5**(1), 87. (Electronic resource, paper available at `www.biomedcentral.com/1471-2164/5/87`.)

Witte, J. S., R. C. Elston, and L. R. Cardon (2000). On the relative sample size required for multiple comparisons. *Statistics in Medicine*, **19**, 369–372.

Yazici, H., M. Biyikil, S. van der Linden, and H. J. A. Schouten (2001). The "zero patient" design to compare the prevalences of rare diseases. *Rheumatology*, **40**(2), 121–122.

Yi, Q. and T. Panzarella (2002). Estimating sample sizes for tests on trends across repeated measurements with missing data based on the interaction term in a mixed model. *Controlled Clinical Trials*, **23**(5), 481–496.

Zucker, D. M. and J. Denne (2002). Sample size redetermination for repeated measures studies. *Biometrics*, **58**(3), 548–559. Discussion: **60**, 284–285.

## EXERCISES

**6.1.** Assume that you wish to detect a difference between means of two populations of 2.0 when the standard deviation of each population is assumed to be 6, $\alpha = .05$, and the power is to be .90. How large should the common sample size be? If software is to be used to determine the sample size, is it necessary to use software with experimental design capability? Explain.

**6.2.** Consider a $4 \times 4$ Latin square design with, as in Example 6.1, $\sigma^2 = \sigma_{\hat{\tau}}^2 = 1$ and $\alpha = .05$. Determine the power figures for 1, 2, and 3 Latin squares, with each member of each set of multiple Latin squares being unrelated to the other members of the set.

**6.3.** Assume that six $4 \times 4$ Latin squares are to be used and that $\alpha = .05$ and $\sigma^2 = 1$. How large a value of $\sigma_{\hat{\tau}}^2$ (or equivalently, a value for $\sum_{i=1}^{4} \tau_i^2$) can be detected with a power of .90? If we wanted to detect that value with a power of .80, how many Latin squares should be used?

**6.4.** Assume that an experiment is to be conducted using three levels of a single factor, with these levels being the only ones of interest, so that the factor is fixed. If you have no idea of the effect that each level will have, what complications, if any, does this present? Explain.

**6.5.** Consider designs for nonnormal responses, as discussed in Section 6.12. Determine the sample size that would be needed to detect a reduction from $p_0 = .10$ to $p = .05$ with a two-tailed test with $\alpha = .05$ and power $= .90$ when a $2^4$ design is used.

**6.6.** Assume that a 20-run Plackett–Burman design is to be used for 14 factors. What is the power for detecting a $1\sigma$ effect if the design is not replicated? Would you suggest that such a design be used, or would you prefer to have two replicates and accept the power that is obtained using two replicates? In particular, does using two replicates instead of one help very much? Explain.

**6.7.** Assume that a one-way ANOVA is to be used with three levels of the factor. With $\alpha = .05$, $\sigma^2 = 1$, and $\sigma_\tau^2 = 2$, how many observations should be used with each level in order to have power $= .90$?

**6.8.** Consider the standardized effect sizes used by Cohen (1988) that were listed in Section 6.2. If you wanted to design software to make it as simple as possible for the user and decided to select six standardized effect sizes from which the user would select and thus avoid having to estimate $\sigma$, which six values would you select? Do they differ considerable from those given by Cohen? If so, explain why.

**6.9.** When will the desired power for a particular experimental design not be at least approximately met?

**6.10.** Regarding Exercise 6.1, if one of the standard deviations was believed to be 9 and the other two 6, how would you proceed in determining sample size? In particular, would you use different sample sizes? Explain.

**6.11.** Geng and Hills (1978) considered the determination of sample size for certain experimental designs. In one example that they cited from the literature, a replicated $6 \times 6$ Latin square design was used. This study was used to provide necessary parameter estimates for another study in which the same design is to be used. The experimenter wanted to detect one mean being 10% greater than the overall mean, assumed to be 13.5. They computed the noncentrality parameter to be 2.24. This should apparently be 2.40, although they define the noncentrality parameter as some writers have done, by taking the square root of the expression used in this chapter. Since software for computing power for experimental designs was not readily available in the mid-1970s, they used a Pearson and Hartley chart and concluded that the power would be greater than .90. For either definition of the noncentrality parameter, do you agree that the power is greater than .90? Explain. From a practical standpoint, if five of the six treatment means are equal to the hypothesized average value and only one mean differs from the average, and it differs by only 10%, would we expect an $F$-test, thinking about what the test is designed to detect, to be able to detect this difference with high power without using a very large number of observations?

**6.12.** Consider Example 3.5, which was analyzed as a repeated measures design in Section 6.8. Use PASS or other software and enter the covariance matrix as 25, $a$, $a$, 25, with $a = 19$ and 22, in addition to the $a = 20$ in that example.

**(a)** What effect do these changes in the value of $a$ have on the sample size?

**(b)** Considering the definition, $\rho_{XY} = \text{Cov}(X, Y)/(\sigma_X \sigma_Y)$, with $\rho_{XY}$ denoting the correlation between $X$ and $Y$ and $\text{Cov}(X, Y)$ denoting the covariance between $X$ and $Y$, with $\sigma_X$ and $\sigma_Y$ fixed in this example, what is happening to the correlation between the two factor levels as $a$ is increased? Does the change in the sample size relative to the change in the correlation make sense intuitively? What does this tell you?

**(c)** Although negative correlations generally do not occur with repeated measures designs, now let $a = -19$, $-20$, and $-22$ and determine the sample size for each of these values. Comment relative to what you saw using positive values of $a$.