CHAPTER 4

Completely Randomized Design

4.1 Description of the Design

Chapters 1 to 3 introduced some basic concepts and statistical tools that are used in experimental design. In this and the following chapters, those designs that appear to have the greatest usefulness to researchers in the behavioral sciences, health sciences, and education are examined in detail.

One of the simplest experimental designs from the standpoint of data analysis and assignment of subjects or experimental units to treatment levels is the completely randomized design. The design is denoted by the letters CR-*p,* where CR stands for *completely randomized* and *p* is the number of levels of the treatment. The layout for a completely randomized design with four treatment levels is shown in Figure 4.1-1.

A CR-*p* design is appropriate for experiments that meet, in addition to the general assumptions of analysis of variance summarized in Section 3.5, the following two conditions:

1. One treatment with $p \geq 2$ treatment levels. The levels of the treatment can differ either quantitatively or qualitatively. When the experiment contains only two treatment levels, the design is indistinguishable from the *t* test for independent-samples design that is described in Section 2.2.

2. Random assignment of experimental units to the treatment levels, with each experimental unit designated to receive only one level. The number of experimental units in each treatment level need not be equal, although this is desirable. According to Section 3.5, the *F* statistic is more robust to violation of some assumptions when the sample *n*s are equal.

It is apparent that the completely randomized design is applicable to a broad range of experimental situations. As I discuss in Section 2.2, the design is one of the three building block designs that can be used by itself or in combination to form more complex designs. An understanding of the completely randomized design is fundamental to understanding a number of more complex designs.

		Treat. Level	Dep. Var.
Group_1	Subject_1	a_1	Y_{11}
	Subject_2	a_1	Y_{21}
	⋮	⋮	⋮
	Subject_5	a_1	Y_{51}
			$\overline{Y}_{\cdot 1}$
Group_2	Subject_1	a_2	Y_{12}
	Subject_2	a_2	Y_{22}
	⋮	⋮	⋮
	Subject_5	a_2	Y_{52}
			$\overline{Y}_{\cdot 2}$
Group_3	Subject_1	a_3	Y_{13}
	Subject_2	a_3	Y_{23}
	⋮	⋮	⋮
	Subject_5	a_3	Y_{53}
			$\overline{Y}_{\cdot 3}$
Group_4	Subject_1	a_4	Y_{14}
	Subject_2	a_4	Y_{24}
	⋮	⋮	⋮
	Subject_5	a_4	Y_{54}
			$\overline{Y}_{\cdot 4}$

Figure 4.1-1 ■ Layout for a completely randomized design (CR-4 design) with $p = 4$ treatment levels denoted by a_1, a_2, a_3, and a_4. The subjects are randomly assigned to the treatment levels. The $n = 5$ subjects in Group_1 receive treatment level a_1, those in Group_2 receive treatment level a_2, and so on. The dependent-variable means for the subjects who receive treatment levels a_1, a_2, a_3, and a_4 are denoted by $\overline{Y}_{\cdot 1}$, $\overline{Y}_{\cdot 2}$, $\overline{Y}_{\cdot 3}$, and $\overline{Y}_{\cdot 4}$, respectively.

Experimental Design Model for a CR-*p* Design

I describe the model equation for a completely randomized design in Section 2.2 and the assumptions for the model in Sections 3.3 and 3.5. Here I elaborate on the assumptions for the fixed-effects model.

1. The model equation $Y_{ij} = \mu + \alpha_j + \varepsilon_{i(j)}$ $(i = 1, \ldots, n; j = 1, \ldots, p)$ for a CR-p design contains all of the sources of variation that affect observation Y_{ij} for subject i in treatment level j.

μ is the grand mean, the mean of the p population means, μ_j.

α_j is the treatment effect for population j and is equal to $\mu_j - \mu$, the deviation of the grand mean from the jth population mean.

$\varepsilon_{i(j)}$ is the error effect associated with Y_{ij} and is equal to $Y_{ij} - \mu_j$. The error effect represents effects unique to subject i, effects attributable to chance fluctuations in subject i's behavior, and any other effects that have not been controlled—in other words, all effects not attributable to treatment level a_j.

2. The experiment contains all of the treatment levels, α_js, of interest. As a result, the treatment effects sum to zero, $\sum_{j=1}^{p} \alpha_j = 0$.

3. The error effect, $\varepsilon_{i(j)}$, is normally and independently distributed within each treatment population with mean equal to zero and variance equal to σ_ε^2. This assumption is often abbreviated as $\varepsilon_{i(j)}$ is $NID(0, \sigma_\varepsilon^2)$, where $NID(0, \sigma_\varepsilon^2)$ denotes normally and independently distributed with mean = 0 and variance = σ_ε^2.

The fixed-effects model is the most commonly used model for a CR-p design. The random-effects model in which the p treatment levels are randomly sampled from a population of P levels ($p < P$) is discussed in Section 4.6.

4.2 Exploratory Data Analysis

The emphasis in this book is on **confirmatory data analysis**—using samples to tell us something about the populations from which they came and assessing the precision of our inferences concerning the populations. But every confirmatory data analysis should be preceded by an **exploratory data analysis**—looking at data to see what they seem to say. Eyeballing data is an important first step in any confirmatory data analysis. Such an exploration may uncover, for example, suspected data recording errors, assumptions that appear untenable, and unexpected promising lines of investigation. Several exploratory techniques are described here. For in-depth coverage, the reader should refer to Tukey (1977) and Hoaglin, Mosteller, and Tukey (1991).

Checking the Model Assumptions

Suppose that I am interested in the effects of sleep deprivation, treatment A, on hand-steadiness. The four levels of sleep deprivation of interest are 12, 18, 24, and 30 hours, which are denoted by a_1, a_2, a_3, and a_4, respectively. Suppose that I have conducted an experiment in which 32 subjects were randomly assigned to the four levels of sleep deprivation, with the restriction that 8 subjects were assigned to each level. The dependent variable is the number of times during a 2-minute interval that a stylus makes contact with the side of a half-inch hole. The layout for the design is similar to that shown in Figure 4.1-1. The research hypothesis that led to the experiment is based on the idea that sleep deprivation affects hand-steadiness. A hypothetical set of data for the experiment is shown in Table 4.2-1(i). The

Table 4.2-1 ■ Summary of Hand-Steadiness Data

(i) Data

Treatment Levels

a_1	a_2	a_3	a_4
3	4	4	3
2	4	4	5
2	3	3	6
3	3	2	5
1	1	4	6
3	3	7	6
6	6	5	8
4	4	5	9

(ii) Descriptive statistics

	a_1	a_2	a_3	a_4
	12 Hours	18 Hours	24 Hours	30 Hours
$\bar{Y}_{\cdot j}$	3.00	3.50	4.25	6.00
$\hat{\sigma}_j$	1.51	1.41	1.49	1.85

$$\bar{Y}_{\cdot j} = \sum_{i=1}^{n} Y_{ij} / n \qquad \hat{\sigma}_j = \sqrt{\frac{\sum_{i=1}^{n} Y_{ij}^2 - \frac{\left(\sum_{i=1}^{n} Y_{ij}\right)^2}{n}}{n-1}}$$

descriptive statistics in part (ii) support the research hypothesis: that the sample means for the sleep deprivation conditions differ.

In Section 4.1, I described the assumptions of the error effects: The $\varepsilon_{i(j)}$s should be normally distributed, have equal variances, and be mutually independent. To determine whether the assumptions are tenable, it is helpful to examine a plot of **standardized residuals,** denoted by $z_{i(j)}$. A residual (error effect) is given by $\hat{\varepsilon}_{i(j)} = Y_{ij} - \bar{Y}_{\cdot j}$. Standardization is achieved by dividing the residuals by their standard deviation. For a completely randomized design, the standard deviation of the residuals is $\hat{\sigma}_z = \sqrt{SSWG/(N-1)}$, where $N = n_1 + \cdots + n_p$. The computation of *SSWG* is illustrated in Section 4.3. A standardized residual for subject *i* in treatment level *j* is given by

$$z_{i(j)} = \hat{\varepsilon}_{i(j)} / \hat{\sigma}_z = (Y_{ij} - \bar{Y}_{\cdot j}) / \sqrt{SSWG/(N-1)}$$

Table 4.2-2 ■ Residuals and Standardized Residuals for the Data in Table 4.2-1

Treatment Levels							
a_1		a_2		a_3		a_4	
$\hat{\varepsilon}_{i(1)}$	$z_{i(1)}$	$\hat{\varepsilon}_{i(2)}$	$z_{i(2)}$	$\hat{\varepsilon}_{i(3)}$	$z_{i(3)}$	$\hat{\varepsilon}_{i(4)}$	$z_{i(4)}$
0	0	0.50	0.33	–0.25	–0.17	–3.00	–2.00
–1.00	–0.67	0.50	0.33	–0.25	–0.17	–1.00	–0.67
–1.00	–0.67	–0.50	–0.33	–1.25	–0.83	0	0
0	0	–0.50	–0.33	–2.25	–1.50	–1.00	–0.67
–2.00	–1.34	–2.50	–1.67	–0.25	–0.17	0	0
0	0	–0.50	–0.33	2.75	1.84	0	0
3.00	2.00	2.50	1.67	0.75	0.50	2.00	1.34
1.00	0.67	0.50	0.33	0.75	0.50	3.00	2.00

$$\hat{\varepsilon}_{i(j)} = Y_{ij} - \overline{Y}_{\cdot j} \qquad z_{i(j)} = \hat{\varepsilon}_{i(j)} / \sqrt{SSWG/(N-1)}$$

$$= \hat{\varepsilon}_{i(j)} / \sqrt{69.5000/(32-1)}$$

$$= \hat{\varepsilon}_{i(j)} / 1.4973$$

SSWG is computed in Table 4.3-1.

If the assumptions of the model are tenable, the standardized residuals should be normally and independently distributed with mean equal to 0 and variance equal to 1; $z_{i(j)}$ is $NID(0, 1)$. Hence, to check on the model assumption, one looks for deviations from patterns that would be expected of independent observations from a standard normal distribution.

Residuals and standardized residuals for the data in Table 4.2-1 are shown in Table 4.2-2. In Figure 4.2-1(a), the standardized residuals in Table 4.2-2 are displayed in the form of frequency distributions. If the model assumptions are tenable, approximately 68.3% of the standardized residuals should be between –1 and 1, approximately 95.4% between –2 and 2, and approximately 99.7% between –3 and 3. Based on the residual plots, there is no reason to doubt the tenability of the normality and homogeneity of variance assumptions. Other procedures for testing the hypothesis of homogeneity of the population variances are described in Section 3.5.

Figure 4.2-1(b) displays a different kind of information. Here, the residuals are plotted against the order in which the hand-steadiness measurements were collected. If the independence assumption is tenable, the standardized residuals should be randomly distributed around zero with no discernable pattern. Nonindependence is indicated if the $z_{i(jk)}$s show a consistent downward or upward trend or they have the shape of a megaphone. The independence assumption appears to be satisfied for treatment levels a_1 through a_3. However,

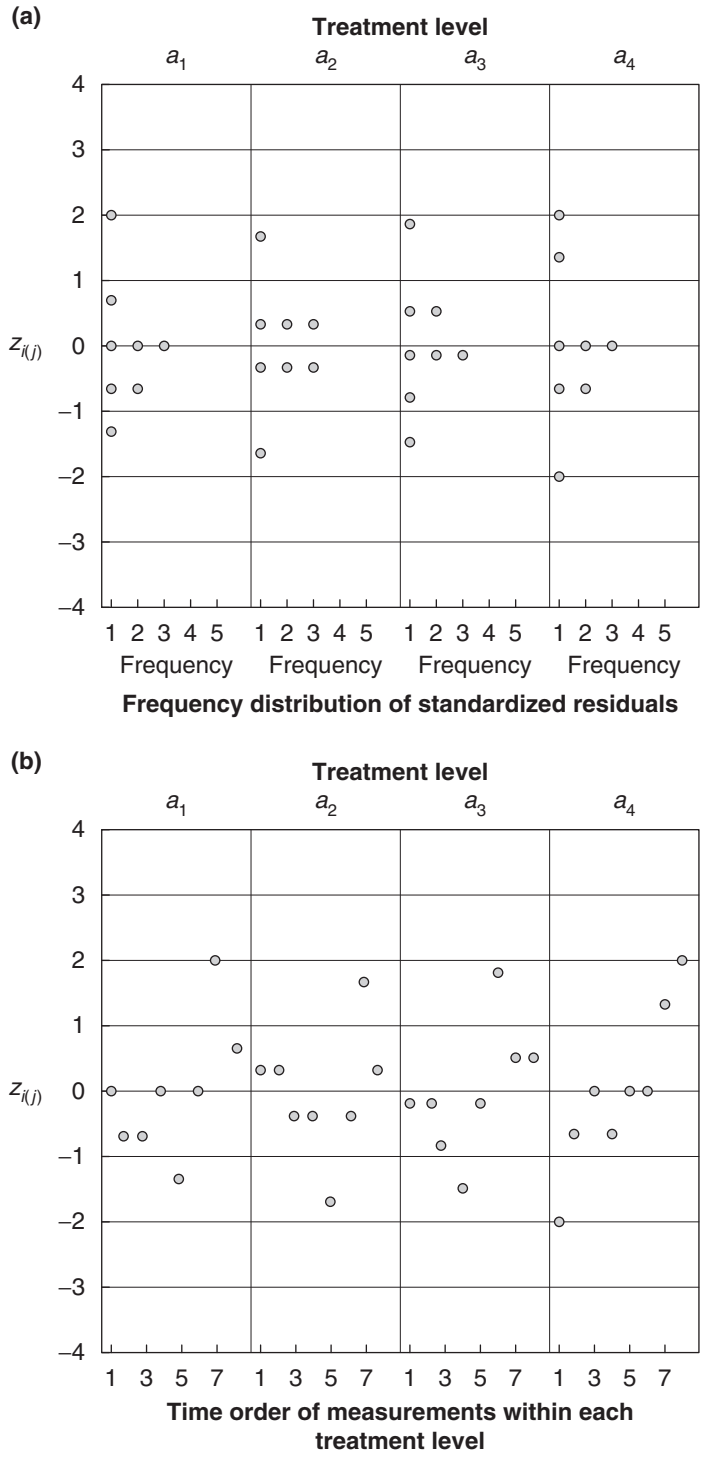

Figure 4.2-1 ■ (a) Frequency distributions of standardized residuals, $z_{i(j)} = (Y_{ij} - \overline{Y}_{\cdot j}) / \sqrt{SSWG / (N-1)}$. (b) Plot of standardized residuals versus the order in which measurements within each treatment level were obtained.

the standardized residuals in treatment level a_4 increase as a function of the order in which the measurements were collected—strong evidence that the independence assumption is violated. A researcher would certainly want to review the data collection procedures for this treatment level.

Outliers

Occasionally one encounters data with one or more observations that deviate markedly from other observations in the sample. Such observations are called **outliers.** In a standardized residual plot, they are observations for which $|z_{i(j)}| > 2.5$. An examination of Figure 4.2-1(a) reveals that there are no outliers. Box plots also are useful for detecting outliers and treatment populations that are not symmetrical. Box plots are discussed in most introductory statistics books.

When outliers occur, they call for detective work. A researcher must decide whether the residuals merely represent extreme manifestations of the random variability inherent in data or are the result of deviations from prescribed experimental procedures, recording errors, equipment malfunctions, and so on. If they reflect the random variability inherent in data, they should be retained and processed in the same manner as the other observations. If some physical explanation for the outlier can be found, a researcher may (1) replace the observation with new data, (2) correct the observation if records permit, or (3) reject the observation and Winsorize. Winsorization is described in Section 3.6. After performing an exploratory data analysis and deciding that the assumptions of the model are tenable, the next step is a confirmatory data analysis.

4.3 Computational Example for CR-4 Design

The statistical hypotheses for the hand-steadiness data in Table 4.2-1 are

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \qquad \text{or} \qquad H_0: \alpha_j = 0 \text{ for all } j$$

$$H_1: \mu_j \neq \mu_{j'} \text{ for some } j \text{ and } j' \qquad H_1: \alpha_j \neq 0 \text{ for some } j$$

The level of significance adopted is $\alpha = .05$. Procedures for computing the sums of squares used in testing the null hypothesis are illustrated in Table 4.3-1. The *AS* Summary Table is so named because variation among the 32 scores reflects the effects of the treatment *A* and the subjects, denoted by *S* for subjects. The computational scheme in parts (ii) and (iii) of the table uses the abbreviated symbols [*AS*], [*A*], and [*Y*] that were introduced in Section 3.2. This abbreviated notation simplifies the presentation of the computational formulas.

An ANOVA table summarizing the results of the analysis is shown in Table 4.3-2. The mean square (*MS*) in each row is obtained by dividing the sum of squares (*SS*) by the degrees of freedom (*df*) in its row. Recall from Section 3.3 that an *MS* is an estimator of a population variance and is given by

$$MS = \hat{\sigma}^2 = \frac{SS}{df}$$

Table 4.3-1 ■ Computational Procedures for a CR-4 Design

(i) Data and notation [Y_{ij} denotes a score for subject i in treatment level j; $i = 1, \ldots, n$ subjects (s_i); $j = 1, \ldots, p$ levels of treatment A (a_j).]

AS Summary Table
Entry is Y_{ij}

a_1	a_2	a_3	a_4
3	4	4	3
2	4	4	5
2	3	3	6
3	3	2	5
1	1	4	6
3	3	7	6
6	6	5	8
4	4	5	9
$\sum_{i=1}^{n} Y_{ij} =$ 24	28	34	48

(ii) Computational symbols

$$\sum_{j=1}^{p}\sum_{i=1}^{n} Y_{ij} = 3 + 2 + \cdots + 9 = 134.000$$

$$\frac{\left(\sum_{j=1}^{p}\sum_{i=1}^{n} Y_{ij}\right)^2}{np} = [Y] = \frac{(134)^2}{(8)(4)} = 561.125$$

$$\sum_{j=1}^{p}\sum_{i=1}^{n} Y_{ij}^2 = [AS] = (3)^2 + (2)^2 + \cdots + (9)^2 = 672.000$$

$$\sum_{j=1}^{p} \frac{\left(\sum_{i=1}^{n} Y_{ij}\right)^2}{n} = [A] = \frac{(24)^2}{8} + \frac{(28)^2}{8} + \cdots + \frac{(48)^2}{8} = 602.500$$

(iii) Computational formulas

$SSTO = [AS] - [Y] = 672.000 - 561.125 = 110.875$

$SSBG = [A] - [Y] = 602.500 - 561.125 = 41.375$

$SSWG = [AS] - [A] = 672.000 - 602.500 = 69.500$

The F statistic is obtained by dividing the mean square in the first row by the mean square in the second row. This is indicated symbolically by $\left[\frac{1}{2}\right]$. According to Appendix Table E.4, the value of F that cuts off the upper .05 region of the sampling distribution for 3 and 28 degrees of freedom is $F_{.05;\,3,\,28} = 2.95$. Because the obtained $F = 5.56$ exceeds the table value, $F > F_{.05;\,3,\,28}$, the null hypothesis is rejected.

Table 4.3-2 ■ ANOVA Table for CR-4 Design

Source	*SS*	*df*	*MS*	*F*	*p*	$\hat{\omega}^2$
1. Between groups (sleep deprivation levels)	41.375	$p - 1 = 3$	13.792	$\left[\frac{1}{2}\right]$ 5.56	.004	0.30
2. Within groups	69.500	$p(n - 1) = 28$	2.482			
3. Total	110.875	$np - 1 = 31$				

It is customary to include in an ANOVA table the p value associated with the F statistic and a measure of effect magnitude. The p value for the F statistic was obtained from Microsoft's Excel FDIST function

FDIST (x,deg_freedom1,deg_freedom2)

To illustrate, I replaced x with 5.56 (the value of the F statistic), deg_freedom1 with 3, and deg_freedom2 with 28 as follows

FDIST (5.56,3,28)

Excel returned the p value of .004. The effect magnitude statistic, $\hat{\omega}^2_{Y|A} = 0.30$, in Table 4.3-2 is discussed in Section 4.4. A decision to reject or not reject the null hypothesis should be based on the researcher's preselected level of significance, .05 in the example. The inclusion of the p value permits readers to, in effect, set their own level of significance.

In reports of the results of an experiment, a descriptive summary of the data—means, standard deviations, and perhaps a graph—should always precede the reporting of significance tests. For the sleep deprivation experiment, the descriptive statistics in Table 4.2-1(ii) provide an adequate summary. The results of the F significance test can be presented either by means of a table like Table 4.3-2 or in the text. For simple designs like the completely randomized design, it is customary to present the results in the text. Using this form, the researcher might say, "We can infer from the analysis of variance that the hand-steadiness population means differ, $F(3, 28) = 5.56$, $p < .001$, $\hat{\omega}^2_{Y|A} = 0.30$." Notice that the degrees of freedom for the F statistic are enclosed in parentheses, followed by the value of the F statistic, its p value, and the measure of effect magnitude. If the

experimental design is complex and requires reporting numerous F statistics, the *Publication Manual of the American Psychological Association* (American Psychological Association, 2010, p. 141) states that a tabular presentation can minimize the need for lengthy textual descriptions.

After the omnibus null hypothesis[1] is rejected, the next step in the analysis is to decide which population means differ. Multiple comparison procedures are used for this purpose and are described in Chapter 5.

4.4 Measures of Strength of Association and Effect Size

The importance of distinguishing between statistical significance and practical significance is discussed in Section 2.5. Statistical significance is concerned with whether an observed treatment effect is due to chance. Practical significance is concerned with whether an observed effect is large enough to be useful in the real world. As discussed in Section 2.5, trivial treatment effects can achieve statistical significance if enough subjects are included in an experiment. Small p values—say, .01 or .001—are widely believed to indicate large treatment effects and, hence, practical significance. This interpretation of p values is incorrect because p values are affected by the size of the treatment effects as well as the size of the sample. A p value of .05 for an experiment with 6 subjects per group may reflect larger treatment effects than a p value of .0001 for an experiment with 70 subjects per group. Unfortunately, there is no measure of the practical significance of research results. However, *measures of effect magnitude* can help a researcher make this kind of assessment (Kirk, 2003). Most measures of effect magnitude fall into one of two categories: (1) measures of effect size (typically, standardized mean differences) and (2) measures of strength of association. I describe measures of strength of association first.

Strength of Association

The most widely used measures of strength of association in analysis of variance are **omega squared,** ω^2, introduced by William Hays (1994, p. 408) for fixed treatment effects and the **intraclass correlation,** ρ_I, for random treatment effects. For a completely randomized design, both measures are defined as

$$\frac{\sigma_\alpha^2}{\sigma_\varepsilon^2 + \sigma_\alpha^2} \tag{4.4-1}$$

where σ_α^2 is the variance of the treatment effects and σ_ε^2 is the variance of the error effects. ω^2 and ρ_I indicate the proportion of the population variance in the dependent variable that is accounted for by specifying the treatment-level classification, and thus they are identical in general meaning. Both ω^2 and ρ_I are measures of strength of association for a qualitative or quantitative independent variable and a quantitative dependent variable.

[1]The omnibus null hypothesis states that all of the population means are equal.

The parameters σ_α^2 and σ_ε^2 in equation (4.4-1) are generally unknown, but they can be estimated from sample data. In Section 3.3, you learned that

$$E(MSBG) = \sigma_\varepsilon^2 + \frac{n\sum_{j=1}^{p}\alpha_j^2}{p-1} \qquad \text{and} \qquad E(MSWG) = \sigma_\varepsilon^2$$

for the fixed-effects model and

$$E(MSBG) = \sigma_\varepsilon^2 + n\sigma_\alpha^2 \qquad \text{and} \qquad E(MSWG) = \sigma_\varepsilon^2$$

for the random-effects model. It follows that unbiased estimators of σ_α^2 and σ_ε^2 are given by

$$\frac{p-1}{np}(MSBG - MSWG) = \frac{\sum_{j=1}^{p}\hat{\alpha}_j^2}{p} = \hat{\sigma}_\alpha^2 \qquad \text{and} \qquad MSWG = \hat{\sigma}_\varepsilon^2$$

for the fixed-effects model and by

$$\frac{1}{n}(MSBG - MSWG) = \hat{\sigma}_\alpha^2 \qquad \text{and} \qquad MSWG = \hat{\sigma}_\varepsilon^2$$

for the random-effects model. If the estimators for σ_α^2 and σ_ε^2 are substituted in equation (4.4-1), the following formulas for $\hat{\omega}^2$ and $\hat{\rho}_I$ can be obtained with the aid of a little algebra:

$$\hat{\omega}^2 = \frac{SSBG - (p-1)MSWG}{SSTO + MSWG} \qquad \hat{\rho}_I = \frac{MSBG - MSWG}{MSBG + (n-1)MSWG}$$

For the hand-steadiness data in Table 4.3-2,

$$\hat{\omega}^2 = \frac{41.375 - (4-1)2.482}{110.875 + 2.482} = 0.30$$

Thus, the four levels of sleep deprivation account for 30% of the variance in the hand-steadiness scores. Not only is the association statistically significant, as is evident from the significant F statistic in Table 4.3-2, but also the association is quite strong.

Based on Cohen's (1988, pp. 284–288) classic work, the following guidelines are suggested for interpreting strength of association:

$\omega^2 = .010$ is a small association.

$\omega^2 = .059$ is a medium association.

$\omega^2 = .138$ or larger is a large association.

When a sample omega squared is negative, the best estimate of the population value is 0. Sedlmeier and Gigerenzer (1989) and Cooper and Findley (1982) reported that the typical strength of association in the journals that they examined was around .06—a medium association.

Omega squared and the intraclass correlation also can be computed from a knowledge of the F statistic, sample size in each treatment level, and number of treatment levels. The alternative formula for $\hat{\omega}^2$ and the value of $\hat{\omega}^2$ for the hand-steadiness data are

$$\hat{\omega}^2 = \frac{(p-1)(F-1)}{(p-1)(F-1)+np} = \frac{(4-1)(5.56-1)}{(4-1)(5.56-1)+(8)(4)} = .30$$

where F, n, and p are obtained from Table 4.3-2. If treatment A represents random effects, the intraclass correlation can be computed from

$$\hat{\rho}_I = \frac{F-1}{(n-1)+F}$$

These formulas for $\hat{\omega}^2$ and $\hat{\rho}_I$ can be used to assess the practical significance of published research where only the F statistic and degrees of freedom are provided.

The formulas for $\hat{\omega}^2$ given earlier assume that the sample ns are equal. If the sample ns are not too different, Vaughan and Corballis (1969) have suggested the following formula for approximating omega squared:

$$\hat{\omega}^2 = \frac{SSBG-(p-1)MSWG}{SSBG+p(\overline{n}-1)MSWG+MSWG}$$

where $\overline{n}$ is the mean of the sample ns.

Omega squared and the intraclass correlation, like the F statistic, are omnibus (overall) statistics. Researchers generally are not as interested in this omnibus statistic as they are in knowing how much of the variance in the dependent variable is accounted for by the difference between selected treatment levels, say, the means for treatment levels a_1 and a_2. One degree-of-freedom omega-squared correlation measures that address this kind of question are discussed in Section 6.5.

In interpreting omega squared, it is important to remember that the treatment levels are selected a priori rather than by random sampling as is the case for the intraclass correlation. The presence of a truncated range or the selection of extreme values of a quantitative independent variable can markedly affect the value of $\hat{\omega}^2$. Omega squared applies to the treatment levels in the experiment; any generalization to levels not included in the experiment is a leap of faith. Note also that $\hat{\omega}^2$ and $\hat{\rho}_I$ are computed from the ratio of unbiased estimators; hence, they are biased estimators of the corresponding population parameters. In general, the ratio of two unbiased estimators is not itself an unbiased estimator. Carroll and Nordholm (1975) have shown that the degree of bias in $\hat{\omega}^2$ is slight.

Other statistics such as R^2, **coefficient of multiple determination** or eta squared ($\hat{\eta}^2$), and $\tilde{R}^2$ also are used to measure the strength of association between the independent and dependent variables. The R^2 statistic is given by

$$R^2 = \frac{SSBG}{SSTO}$$

and indicates the *sample* proportion of variance in the dependent variable that is accounted for by specifying the treatment-level classification. R^2 tends to overestimate the population parameter. For the hand-steadiness data in Table 4.2-1, R^2 = 41.375/110.87 = .37. An adjustment due to Wherry (1931) can be applied to R^2 to obtain a better estimate of the population parameter. The adjusted (shrunken) coefficient is denoted by $\tilde{R}^2$ and is computed from $\tilde{R}^2 = 1 - \frac{N-1}{N-p}(1-R^2)$, where $N = n_1 + n_2 + \cdots + n_p$. For the hand-steadiness data, $\tilde{R}^2$ = .31.

Effect Size

A second approach to assessing the practical significance of research results is based on differences among means. In Section 2.5, I describe a measure popularized by Jacob Cohen (1988) called effect size and denoted by *d.* The effect-size formulas for one- and two-sample experiments are, respectively,

$$d = \frac{|\mu - \mu_0|}{\sigma_\varepsilon} \qquad \text{and} \qquad d = \frac{|\mu_1 - \mu_2|}{\sigma_\varepsilon}$$

In both formulas, a difference among means is expressed in units of the within-groups population standard deviation. This idea with modifications can be extended to the case in which there are three or more means:

$$f = \sqrt{\frac{\sum_{j=1}^{p} (\mu_j - \mu)^2 / p}{\sigma_\varepsilon^2}} \qquad \text{or} \qquad f = \sqrt{\frac{\sum_{j=1}^{p} \alpha_j^2 / p}{\sigma_\varepsilon^2}}$$

A sample estimate of *f* for the hand-steadiness data in Table 4.3-2 is given by

$$\hat{f} = \sqrt{\frac{\sum_{j=1}^{p} \hat{\alpha}_j^2 / p}{\hat{\sigma}_\varepsilon^2}} = \sqrt{\frac{1.060}{2.482}} = 0.65$$

where

$$\frac{\sum_{j=1}^{p} \hat{\alpha}_j^2}{p} = \frac{p-1}{np}(MSBG - MSWG) = \frac{4-1}{(8)(4)}(13.792 - 2.482) = 1.060$$

$$\hat{\sigma}_\varepsilon^2 = MSWG = 2.482$$

Cohen (1988, pp. 284–288) suggested the following guidelines for interpreting the $\hat{f}$ measure of effect size:

f = .10 is a small effect size.

f = .25 is a medium effect size.

f = .40 or larger is a large effect size.

Based on Cohen's guidelines, the treatment effects for the sleep deprivation experiment are classified as large effects. The same conclusion was reached using $\hat{\omega}^2$. In fact, the two indexes are related as follows:

$$\hat{f} = \sqrt{\frac{\hat{\omega}^2}{1 - \hat{\omega}^2}}$$

For a discussion of the merits of measures of strength of association and effect size, the reader is referred to Cumming (2012), Henson (2006), Huberty (2002), and Kline (2004).

In summary, a significant F statistic for treatment effects in a completely randomized design indicates that there is some association between the independent and dependent variables and that at least one treatment effect is not equal to zero. The $\hat{\omega}^2$ and $\hat{\rho}_I$ statistics estimate the population strength of the association between a qualitative or quantitative independent variable and a quantitative dependent variable. Cohen's $\hat{f}$ and similar measures estimate the relative size of treatment effects. Both kinds of measures provide important information that is not contained in a test of significance. When the results of significance tests are reported, researchers should always include a measure of effect magnitude.

4.5 Power and the Determination of Sample Size

Introduction to the Calculation of Power

Power, denoted by $1 - \beta$, is the probability of rejecting a false null hypothesis. Knowledge of power is useful for assessing the sensitivity of a statistical test and for determining the sample size to use. If the null hypothesis is true, then $F = MSBG/MSWG$ is distributed as a **central *F* distribution.** The central F distribution depends on two parameters: ν_1 and ν_2, the degrees of freedom of the F statistic. F values that cut off the upper .25, .10, .05, and .01 portions of the central F distribution are given in Appendix Table E.4. If the null

hypothesis is false, then $F = MSBG/MSWG$ is distributed as a **noncentral *F* distribution.** This latter distribution is used in determining the power of a test. The noncentral F distribution depends on three parameters: ν_1, ν_2, and a **noncentrality parameter** λ (Greek lambda), where

$$\lambda = \frac{\sum_{j=1}^{p} \alpha_j^2}{\sigma_\varepsilon^2 / n}$$

The parameter λ is a measure of the degree to which the null hypothesis is false. The value of λ is determined by the size of the sum of squared treatment effects relative to σ_ε^2 / n. Tang (1938) prepared charts that simplify the calculation of power. Tang's charts, which are reproduced in Appendix Table E.12, are based on a function of the noncentrality parameter. To use the charts, the parameter ϕ (Greek phi),

$$\phi = \sqrt{\frac{\lambda}{p}} = \sqrt{\frac{\sum_{j=1}^{p} \alpha_j^2 / p}{\sigma_\varepsilon^2 / n}} \tag{4.5-1}$$

is entered in the appropriate chart for $\nu_1 = p - 1$ and $\nu_2 = p(n - 1)$ degrees of freedom and a significance level of either .05 or .01.

Calculation of Power Using Tang's Charts

The calculation of power is illustrated for the data summarized in Table 4.3-2. In practice, the parameters $\sum_{j=1}^{p} \alpha_j^2$ and σ_ε^2 in equation (4.5-1) are unknown. However, as you learned in Section 3.3, the parameters can be estimated from sample data as follows:

$$\frac{\sum_{j=1}^{p} \hat{\alpha}_j^2}{p} = \frac{p-1}{np}(MSBG - MSWG) = 1.060 \text{ and } \hat{\sigma}_\varepsilon^2 = MSWG = 2.482$$

An estimate of ϕ is

$$\hat{\phi} = \sqrt{\frac{\sum_{j=1}^{p} \hat{\alpha}_j^2 / p}{\hat{\sigma}_\varepsilon^2 / n}} = \sqrt{\frac{1.060}{2.482 / 8}} = 1.85$$

with $\nu_1 = p - 1 = 3$ and $\nu_2 = p(n - 1) = 4(8 - 1) = 28$. Appendix Table E.12 contains eight power charts: a chart for $\nu_1 = 1, \ldots, 8$. Each chart contains power curves for $\alpha = .05$ and $\alpha = .01$. Use the .05 curves because .05 is the level of significance adopted in the sleep deprivation experiment. The value of $\hat{\phi} = 1.85$ is located along the $\alpha = .05$ baseline in the

$\nu_1 = 3$ chart. Extend an imaginary vertical line above $\hat{\phi} = 1.85$ until it intersects a point just to the right of the $\nu_2 = 30$ curve; the chart does not contain a $\nu_2 = 28$ curve. If you read across to the vertical axis, the power of the ANOVA F test is found to be approximately .83, which just exceeds the minimum acceptable power of .80.

Cohen (1988, pp. 289–354) provides more extensive tables for determining power than those in Appendix E.12. His tables contain values for $\nu_1 = 1$ through 6, 8, 10, 12, 15, and 24 and $\alpha = .10$, .05, and .01. To use his tables, a researcher computes Cohen's $\hat{f}$ effect size. This effect size can be computed from the noncentrality parameter, $\hat{\lambda}$, or Tang's $\hat{\phi}$ as follows: $\hat{f} = \sqrt{\hat{\lambda} / np} = \hat{\phi} / \sqrt{n}$. Cohen's tables and those in Appendix E.12 are appropriate for fixed effects. Montgomery (2009, pp. 625–628) gives tables for calculating power for random effects.

A plethora of free easy-to-use power and sample size calculators can be found on the Internet. One of my favorites is G*Power 3.

Estimating Sample Size From a Pilot Study

Choosing a sample size is a bewildering task for many researchers. Researchers want to use enough subjects to detect meaningful effects, but they don't want to use too many subjects and squander research resources. Three approaches to estimating sample size are illustrated. The procedures differ in terms of the information that a researcher must provide and in their simplicity. The first approach requires the most information. A researcher must specify the (1) level of significance, α; (2) power, $1 - \beta$; (3) size of the population variance, σ_ε^2; and (4) the sum of the squared population treatment effects, $\sum_{j=1}^{p} \alpha_j^2$. In practice, σ_ε^2 and $\sum_{j=1}^{p} \alpha_j^2$ are unknown. However, there are ways to circumvent this problem. One way is to estimate σ_ε^2 and $\sum_{j=1}^{p} \alpha_j^2$ from a pilot study. Alternatively, estimates of σ_ε^2 and $\sum_{j=1}^{p} \alpha_j^2$ may be obtained from research that is similar to that under consideration.

For the purpose of illustration, suppose that the hand-steadiness data in Table 4.2-1 were obtained in a pilot study to estimate sample size; let $\alpha = .05$ and $1 - \beta = .80$. This choice of values for α and $1 - \beta$ is based on the widely accepted conventions that the probably of making a Type I error should be less than or equal to .05 and the minimum acceptable power should be greater than or equal to .80. With these conventions and the pilot-study information from Table 4.3-2, a researcher can use trial and error to estimate the required sample size. The process consists of inserting trial sample-size values, denoted by n', in

$$\hat{\phi} = \sqrt{n'} \sqrt{\frac{\sum_{j=1}^{p} \hat{\alpha}_j^2 / p}{\hat{\sigma}_\varepsilon^2}}$$

and determining from Tang's charts whether a power of .80 has been achieved. I begin the trial-and-error process with $n' = 7$.

$$\hat{\phi} = \sqrt{7} \sqrt{\frac{1.060}{2.482}} = (2.646)(0.654) = 1.73$$

with $\nu_1 = p - 1 = 3$ and $\nu_2 = p(n' - 1) = 4(7 - 1) = 24$. According to Tang's chart in Appendix Table E.12, $\hat{\phi} = 1.73$ corresponds to a power of .76, which is less than the desired power. Substituting $n' = 8$ in the formula

$$\hat{\phi} = \sqrt{8}\sqrt{\frac{1.060}{2.482}} = (2.828)(0.654) = 1.85$$

with $\nu_1 = 3$ and $\nu_2 = 4(8 - 1) = 28$ gives a power of .83. Thus, if a researcher uses $np = (8)(4) = 32$ subjects, the power is approximately .83.

Estimating Sample Size Using *d*

If accurate estimates of $\sum_{j=1}^{p} \alpha_j^2$ and σ_ε^2 are not available from a pilot study or previous research, the procedure just described for calculating n cannot be used. However, there is an alternative approach that does not require this information. The approach does require a general idea about the size of the difference between the largest and smallest population means that would be useful to detect relative to the size of σ_ε. To use this approach, the difference between the largest and smallest population means that a researcher wants to detect is specified as some multiple, denoted by *d,* of the population standard deviation; that is, $\mu_{max} - \mu_{min} = d\sigma_\varepsilon$. An examination of Figure 4.5-1 should help to clarify the meaning of *d.* For example, the difference between μ_{max} and μ_{min} that a researcher wants to detect might be one and a half times larger than σ_ε, $d = 1.5$, or the difference might be three

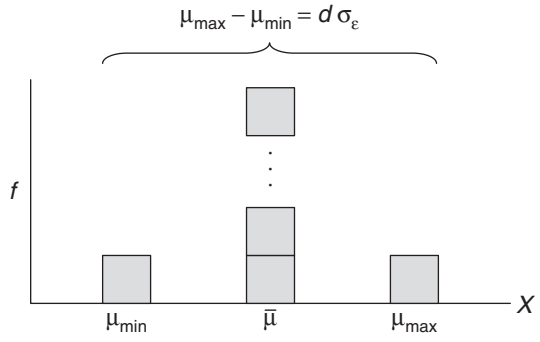

Figure 4.5-1 ■ Each treatment mean is represented by a square. The mean of the p means, the grand mean, is denoted by $\bar{\mu}$. Two of the treatment effects, $\alpha_{min} = \mu_{min} - \bar{\mu} = -d\sigma_\varepsilon / 2$ and $\alpha_{max} = \mu_{max} - \bar{\mu} = d\sigma_\varepsilon / 2$, are not equal to zero. The remaining $p - 2$ treatment effects, $\alpha_j = \mu_j - \bar{\mu} = 0$, are equal to zero. It should be apparent that $\sum_{j=1}^{p} \alpha_j^2$ is minimal when μ_{min} and μ_{max} are not equal to the grand mean and all of the remaining means are equal to the grand mean.

fourths as large as σ_ε, $d = 0.75$. This approach to estimating sample size requires the specification of d but not $\sum_{j=1}^{p}\alpha_j^2$, $\mu_{max} - \mu_{min}$, and σ_ε. Obviously, to specify d, it is necessary to have some idea about the size of $\mu_{max} - \mu_{min}$ that would be worth detecting and to be able to express this difference as a multiple of σ_ε.

When there are more than two means in an experiment, many configurations of means will produce the same value of $\mu_{max} - \mu_{min} = d\sigma_\varepsilon$. It can be shown that the sum of the squared treatment effects, $\sum_{j=1}^{p}\alpha_j^2$, is minimal when two of the means, μ_{min} and μ_{max}, are not equal and the remaining $p - 2$ means are equal to the grand mean. This configuration of means is illustrated in Figure 4.5-1. It should be apparent from the figure that the treatment effect for μ_{min} is equal to $\alpha_{min} = \mu_{min} - \bar{\mu} = -d\sigma_\varepsilon / 2$. Similarly, the treatment effect for μ_{max} is equal to $\alpha_{max} = \mu_{max} - \bar{\mu} = d\sigma_\varepsilon / 2$. Substituting α_{min} and α_{max} for two of the α_js in $\sum_{j=1}^{p}\alpha_j^2$ and zero for the remaining α_js gives

$$\sum_{j=1}^{p}\alpha_j^2 = \left(-\frac{d\sigma_\varepsilon}{2}\right)^2 + (0)^2 + \cdots + (0)^2 + \left(\frac{d\sigma_\varepsilon}{2}\right)^2 = \frac{2d^2\sigma_\varepsilon^2}{4} = \frac{d^2\sigma_\varepsilon^2}{2}$$

Because power increases with an increase in $\sum_{j=1}^{p}\alpha_j^2$, it follows that a choice of values for the α_js other than these will always lead to greater power. Hence, if the sample size necessary to achieve a given power is computed for these treatment effects, a researcher can be certain that any other configuration for which the maximum difference between means is equal to $d\sigma_\varepsilon$ will yield a power greater than that specified. The ϕ formula for estimating sample size is obtained by replacing $\sum_{j=1}^{p}\alpha_j^2$ with $d^2\sigma_\varepsilon^2 / 2$ as follows:

$$\phi = \sqrt{n'}\sqrt{\frac{\sum_{j=1}^{p}\alpha_j^2 / p}{\sigma_\varepsilon^2}} = \sqrt{n'}\sqrt{\frac{\left(d^2\sigma_\varepsilon^2 / 2\right) / p}{\sigma_\varepsilon^2}} = \sqrt{n'}\sqrt{\frac{d^2}{2p}}$$

Assume that an experiment contains four treatment levels and I am interested in detecting differences among means such that $\mu_{max} - \mu_{min}$ is equal to $1.5\sigma_\varepsilon$. In this example, $d = 1.5$, $\alpha = .05$, $1 - \beta = .80$, and $\nu_1 = p - 1 = 3$. Various trial sample-size values, n', can be tried in the formula for ϕ until the desired power is obtained. I begin the trial-and-error process with $n' = 8$.

$$\phi = \sqrt{n'}\sqrt{\frac{d^2}{2p}} = \sqrt{8}\sqrt{\frac{(1.5)^2}{(2)(4)}} = \sqrt{8}(0.530) = 1.50$$

where $\nu_1 = p - 1 = 3$ and $\nu_2 = p(n' - 1) = 4(8 - 1) = 28$. According to Appendix Table E.12, $\phi = 1.50$ corresponds to a power of .64. Obviously, a larger sample n' is required. Substituting $n' = 11$ in the formula gives

$$\phi = \sqrt{n'}\sqrt{\frac{d^2}{2p}} = \sqrt{11}\sqrt{\frac{(1.5)^2}{(2)(4)}} = \sqrt{11}(0.530) = 1.76$$

where $\nu_1 = p - 1 = 3$ and $\nu_2 = p(n' - 1) = 4(11 - 1) = 40$. I get a power of .81. Thus, to detect a difference between the largest and smallest means that is 1.5 times as large as σ_ε, I should use $np = (11)(4) = 44$ subjects. The advantage of this approach to estimating sample size is that it is not necessary to know or estimate $\sum_{j=1}^{p} \alpha_j^2$ and σ_ε. However, it is necessary to specify *d,* which is a kind of effect-size measure.

Estimating Sample Size Using ω^2 and *f*

The third approach to estimating sample size can be used when a researcher knows nothing about $\sum_{j=1}^{p} \alpha_j^2$ and σ_ε and is unable to express $\mu_{max} - \mu_{min}$ as a multiple of σ_ε. This approach requires a researcher to specify the (1) level of significance, α; (2) power, $1 - \beta$; and (3) either the strength of association, ω^2, or the effect size, *f,* that is of interest. The use of ω^2 is described first.

In Section 4.4, Cohen's guidelines for interpreting ω^2 are described. Recall that

$\omega^2 = .010$ is a small association.

$\omega^2 = .059$ is a medium association.

$\omega^2 = .138$ or larger is a large association.

Suppose that a researcher is interested in determining the sample size necessary to detect a large association, $\omega^2 = .138$, for a completely randomized design with $p = 4$ treatment levels. Assume that the researcher has followed the convention of setting $\alpha = .05$ and $1 - \beta = .80$. The sample size can be determined from Appendix Table E.13 for $\nu_1 = 4 - 1 = 3$ and $\nu_2^{CR} = 4(n - 1)$, where ν_1 and ν_2^{CR} denote the degrees of freedom for a completely randomized design.[2] The value of *n* is obtained from the column headed by $\omega^2 = .138$ and the row labeled $1 - \beta = .80$. According to Table E.13, the sample *n* is 18. The experiment requires $np = (18)(4) = 72$ subjects.

The effect-size index, *f,* developed by Cohen (1988) also can be used to determine the required sample size. Cohen suggested the following guidelines for interpreting *f:*

$f = .10$ is a small effect size.

$f = .25$ is a medium effect size.

$f = .40$ or larger is a large effect size.

Suppose that a researcher is interested in determining the sample size necessary to detect a large effect size, $f = .40$, for a completely randomized design with $p = 4$ treatment levels. Assume that $\alpha = .05$ and $1 - \beta = .80$. The required sample size can be determined from Appendix Table E.13 for $\nu_1 = 4 - 1 = 3$ and $\nu_2^{CR} = 4(n - 1)$, where ν_1 and ν_2^{CR} denote the degrees of freedom for a completely randomized design. The value of *n* is obtained from the column headed by $f^* = f = .400$ and the row labeled $1 - \beta = .80$. According to Table E.13, the sample *n* is 18. The experiment requires $np = (18)(4) = 72$ subjects.

[2]I am indebted to Barbara Mobley Foster, who developed the sample-size tables from which Table E.13 was taken.

Appendix Table E.13 can be used to estimate the sample size if $\alpha = .05$, $1 - \beta = .70$, .80, or .90, and the design contains two to four treatment levels. If these conditions are not satisfied, Tang's charts in Appendix Table E.12 can be used to estimate n. The charts are entered with

$$\phi = \sqrt{n'}\sqrt{\frac{\omega^2}{1-\omega^2}} \qquad \text{or} \qquad \phi = \sqrt{n'}f^* \quad (f^* = f)$$

depending on whether one wants to use a strength of association measure or an effect-size measure.

Suppose that a researcher plans to use a completely randomized design and wants to detect a large strength of association, $\omega^2 = .138$, for an experiment with $p = 5$ treatment levels. Assume that $\alpha = .05$ and $1 - \beta = .80$. Various n's can be tried in the formula for ϕ until the desired power is obtained. I begin with $n' = 13$.

$$\phi = \sqrt{n'}\sqrt{\frac{\omega^2}{1-\omega^2}} = \sqrt{13}\sqrt{\frac{.138}{1-.138}} = 3.6056(0.4001) = 1.44$$

with $\nu_1 = 5 - 1 = 4$ and $\nu_2 = 5(13 - 1) = 60$. According to Appendix Table E.12, a power of approximately .70 is obtained if $n' = 13$. Obviously, a larger n' is required. If $n' = 16$, a power of approximately .80 is obtained.

$$\phi = \sqrt{16}\sqrt{\frac{.138}{1-.138}} = 4.0000(0.4001) = 1.60$$

with $\nu_1 = 5 - 1 = 4$ and $\nu_2 = 5(16 - 1) = 75$. The experiment requires $np = (16)(5) = 80$ subjects.

There is a tendency among researchers to underestimate the sample size required to obtain practical significance. In the last example, $np = (16)(5) = 80$ subjects are required to detect a large association. Medium and small associations require, respectively, (39)(5) = 195 subjects and (240)(5) = 1200 subjects.

Three approaches to estimating sample size have been described. The use of ω^2 or f combined with Cohen's guidelines for interpreting values of ω^2 and f requires the least amount of information and is the simplest. Cohen's guidelines are offered as a useful starting point. Researchers should use their subject-matter knowledge to specify appropriate values of ω^2 and f. What constitutes small, medium, and large associations, for example, can vary from one research area to another. Easy-to-use programs for estimating sample size are available on the Internet. Most of the programs require the researcher to specify the type of ANOVA design, an effect magnitude measure, α, $1 - \beta$, and the number of treatment levels.

An estimate of the sample size necessary to detect effects that are practically significant should always be made before an experiment is performed. A researcher may find, for example, that the contemplated sample size is wastefully large, in which case the sample

size can be reduced. On the other hand, a researcher may find that the contemplated sample size is too small and gives less than a 60% chance of detecting treatment effects considered of practical significance. In this case, a researcher may (1) attempt to secure enough subjects to obtain a power of .80, (2) decide not to conduct the experiment, or (3) attempt to modify the experiment so as to reduce the required number of subjects. The modification could involve selecting a less stringent level of significance, settling for lower power, increasing the size of treatment effects that are of interest, or redesigning the experiment to obtain a more precise estimate of treatment effects and a smaller error term.

4.6 Random-Effects Model

The experimental design model equation for a completely randomized design is given in Section 4.1 as

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{i(j)} \qquad (i = 1, \ldots, n;\ j = 1, \ldots, p)$$

There I assumed that the treatment effects are fixed effects, μ is a constant, and $\varepsilon_{i(j)}$ is $NID(0, \sigma^2_\varepsilon)$. This model is called a fixed-effects model or **model I.** Alternatively, the p treatment levels in the experiment may represent a random sample from a population of P levels, where P is large relative to p. For this case, the treatment effects are random effects and the α_js are assumed to be $NID(0, \sigma^2_\alpha)$. As before, μ is a constant and $\varepsilon_{i(j)}$ is $NID(0, \sigma^2_\varepsilon)$. This model is called a random-effects model or **model II.**

A comparison of the expected values of the mean squares for the two models is given in Table 4.6-1. The derivation of $E(MS)$ is given in Section 3.8. For both models, a test of the null hypothesis $\alpha_j = 0$ for all j (model I) or $\sigma^2_\alpha = 0$ (model II) is given by

$$F = \frac{MSBG}{MSWG} = \frac{f(\text{error effects}) + f(\text{treatment effects})}{f(\text{error effects})}$$

where $f(\ )$ denotes a function of the effects in parentheses. If any treatment effects exist, the numerator of the F statistic should be larger than the denominator. This F statistic adheres to a basic principle that is shared by all ANOVA F statistics: The expected value of the numerator should always contain one more term than the expected value of the denominator. For the random-effects model, $E(MSBG) = \sigma^2_\varepsilon + n\sigma^2_\alpha$ and $E(MSWG) = \sigma^2_\varepsilon$. The F test can be regarded as a procedure for deciding, on the basis of sample data, which of the following model equations

$$Y_{ij} = \mu + \varepsilon_{i(j)}$$

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{i(j)}$$

underlies observations in the population.[3] If the null hypothesis is rejected, the second equation is adopted; if not, the first equation remains tenable.

[3]This view is explored in detail in Chapter 7.

Table 4.6-1 ■ Comparison of $E(MS)$ for Models I and II

Source	Model I $E(MS)$	Model II $E(MS)$
MSBG	$\sigma_\varepsilon^2 + n\sum_{j=1}^{p}\alpha_j^2/(p-1)$	$\sigma_\varepsilon^2 + n\sigma_\alpha^2$
MSWG	σ_ε^2	σ_ε^2

As you have seen, the fixed- and random-effects models are identical except for the assumptions about the nature of the treatment effects. This difference is important because it determines the nature of the conclusions that can be drawn from an experiment. For the fixed-effects model, conclusions are restricted to the p treatment levels in the experiment. For the random-effects model, conclusions apply to the P treatment populations from which the p treatment levels were randomly sampled.

4.7 Advantages and Disadvantages of CR-*p* Design

The major advantages of the completely randomized design are as follows:

1. The layout of the design is simple.
2. Statistical analysis and interpretation of results are relatively straightforward.
3. The design does not require equal sample sizes for each treatment level.
4. It allows for the maximum number of degrees of freedom for the error sum of squares.
5. The design does not require a subject to participate under more than one treatment level or the use of subjects who have been matched on an appropriate variable.

The major disadvantages of the design are as follows:

1. The effects of differences among subjects are controlled by random assignment of the subjects to treatment levels. For this to be effective, subjects should be relatively homogeneous or a large number of subjects should be used.
2. When many treatment levels are included in the experiment, the required sample size may be prohibitive.

4.8 Review Exercises

1. Terms to remember:
   a. confirmatory data analysis (4.2)
   b. exploratory data analysis (4.2)
   c. standardized residual (4.2)
   d. outlier (4.2)
   e. omega squared (4.4)
   f. intraclass correlation (4.4)

g. coefficient of multiple determination (4.4)
h. central F distribution (4.5)
i. noncentral F distribution (4.5)
j. noncentrality parameter (4.5)
k. model I (4.6)
l. model II (4.6)

*2. Two approaches to learning problem solving strategies—more specifically, generating alternative solutions—were investigated. Thirty sixth-graders were randomly assigned to one of the two approaches and a control condition. Treatment level a_1, referred to as the training condition, involved participating in five sessions per week during 3 consecutive weeks. Students assigned to this condition observed a videotape introduction for 10 minutes, practiced the skill for 15 minutes, observed peer models via videotape for 15 minutes, and watched a videotaped review for 10 minutes. Treatment level a_2, a film and discussion condition, was conducted concurrently with the training condition and for the same amount of time. Films related to generating alternative solutions were shown followed by group discussions. The students in the control condition, treatment level a_3, did not receive any form of training. At the conclusion of the experiment, five problem situations were presented and the students were instructed to write down as many solutions to each one as they could. The dependent variable was the number of solutions proposed, summed across the five problems. The following data were obtained. (Experiment suggested by Poitras-Martin, D., & Steve, G. L. Psychological education: A skills-oriented approach. *Journal of Counseling Psychology.*)

a_1	a_2	a_3
11	11	7
12	14	18
19	10	16
13	9	11
17	12	9
15	13	10
17	10	13
14	8	14
13	14	12
16	11	12

*a. [4.2] Perform an exploratory data analysis on these data (see Table 4.2-1 and Figure 4.2-1). Assume that the observations within each treatment level are listed in the order in which the observations were obtained. Interpret the analysis.

*b. [4.3] Test the null hypothesis $\mu_1 = \mu_2 = \mu_3$; let $\alpha = .05$. Construct an ANOVA table and make a decision about the null hypothesis.

*c. [4.4] Compute and interpret $\hat{\omega}^2$ and $\hat{f}$ for these data.

*d. [4.5] Calculate the power of the test in part (b).

*e. [4.5] Use the results of part (b) as a pilot study and determine the number of subjects required to achieve a power of approximately .80.

*f. [4.5] Determine the number of subjects required to achieve a power of .80, where the largest difference among means is $1.10\sigma_\epsilon$.

*g. [4.5] Determine the number of subjects required to detect a medium association with power equal to .80.

h. Prepare a "results and discussion section" appropriate for the *Journal of Counseling Psychology.*

*3. The effects of instructions-to-learn on performance on a delayed-recall test were investigated. Twenty men and women college undergraduate volunteers were randomly assigned to two instructional conditions. The subjects assigned to treatment level a_1 were informed of a subsequent recall test prior to the presentation of a word list and were told to use any kind of rehearsal that they felt would aid their recall. The subjects in treatment level a_2 were not informed of a subsequent recall test. Thirty concrete nouns were shown to the subjects. Each noun was presented for 1 second with a 9-second interstimulus interval. As each noun was shown, the subjects were required to write it down. Twenty-four hours later, the subjects were given a 10-minute written recall test. The dependent variable was the number of nouns recalled. The following data were obtained. (Experiment suggested by McDaniel, Mark A., & Masson, M. E. Long-term retention: When incidental semantic processing fails. *Journal of Experimental Psychology: Human Learning and Memory.*)

a_1	a_2
10	15
6	8
12	10
9	7
8	5
17	4
15	9
11	11
14	9
11	12

*a. [4.2] Perform an exploratory data analysis on these data (see Table 4.2-1 and Figure 4.2-1). Assume that the observations within each treatment level are listed in the order in which the observations were obtained. Interpret the analysis.

*b. [4.3] Use ANOVA to test the hypothesis $\mu_1 = \mu_2$, let $\alpha = .05$. Construct an ANOVA table and make a decision about the null hypothesis.

*c. [4.4] Compute and interpret $\hat{\omega}^2$ and $\hat{f}$ for these data.

*d. [4.5] Calculate the power of the test in part (b).

*e. [4.5] Use the results of part (b) as a pilot study and determine the number of subjects required to achieve a power of approximately .80.

*f. [4.5] Determine the number of subjects required to detect a large association; let $1 - \beta = .80$.

*g. [4.5] Determine the number of subjects required to achieve a power of .80, where the largest difference among means is $1.15\sigma_\varepsilon$.

h. Prepare a "results and discussion section" appropriate for the *Journal of Experimental Psychology: Human Learning and Memory.*

4. The effects of written instructions designed to maximize subject attention to hypnotic facilitative information were investigated. The subjects were 36 hypnotically naive male and female college students who scored in the low and moderate ranges on the Harvard Group Scale of Hypnotic Susceptibility. The subjects were randomly assigned to one of four groups with nine subjects in each group. Subjects in the programmed active information group, treatment level a_1, read a booklet about hypnosis. Interspersed throughout the booklet were incomplete sentences designed to test the subject's knowledge of the material. Answers were provided on the following page of the booklet. Subjects in the active information group, treatment level a_2, read a booklet that covered the same information but did not contain the self-testing feature. Subjects in the passive information group, treatment level a_3, read a booklet about the historical development of hypnosis but with no information about how to experience hypnosis. Subjects in the control group, treatment level a_4, were given several magazines and told to browse through them in a relaxed manner. Following this phase of the experiment, subjects took the Stanford Hypnotic Susceptibility Scale, Form C. The dependent variable was the subject's score on this scale. The following data were obtained. (Experiment suggested by Diamond, Michael Jay, Steadman, Clarence, Harada, D., & Rosenthal, J. The use of direct instructions to modify hypnotic performance: The effects of programmed learning procedures. *Journal of Abnormal Psychology.*)

a_1	a_2	a_3	a_4
4	10	4	4
7	6	6	2
5	3	5	5
6	4	2	7
10	7	10	5
11	8	9	1
9	5	7	3
7	9	6	6
8	7	7	4

a. [4.2] Perform an exploratory data analysis on these data (see Table 4.2-1 and Figure 4.2-1). Assume that the observations within each treatment level are listed in the order in which the observations were obtained. Interpret the analysis.

b. [4.3] Test the hypothesis $\mu_1 = \mu_2 = \mu_3 = \mu_4$; let $\alpha = .05$. Construct an ANOVA table and make a decision about the null hypothesis.

c. [4.4] Compute and interpret $\hat{\omega}^2$ and $\hat{f}$ for these data.

d. [4.5] Calculate the power of the test in part (b).

e. [4.5] Use the results of part (b) as a pilot study and determine the number of subjects required to achieve a power of approximately .80.

f. [4.5] Determine the number of subjects required to detect a medium association; let $1 - \beta = .80$.

g. [4.5] Determine the number of subjects required to achieve a power of .80, where the largest difference among means is $0.95\sigma_\varepsilon$.

h. Prepare a "results and discussion section" for the *Journal of Abnormal Psychology.*

5. An experiment was designed to evaluate the effects of different levels of training on children's ability to acquire the concept of an equilateral triangle. Fifty 3-year-old children were recruited from daycare facilities and randomly assigned to one of five groups, with 10 children in each group. Each group contained an equal number of boys and girls. Children in treatment level a_1 (visual condition) were shown 36 blocks, one at a time, and instructed to look at them but not to touch them. Children in treatment level a_2 (visual plus motor condition) looked at the blocks and were permitted to play with them. They also were asked to perform specific tactile-kinesthetic exercises, such as tracing the perimeter of the blocks with their index finger. Children in treatment level a_3 (visual plus verbal condition) looked at the blocks and were told to notice differences in their shape, color, size, and thickness. Children in treatment level a_4 (visual plus motor plus verbal condition) used a combination of visual, motor, and verbal means of stimulus predifferentiation. Children in treatment level a_5 (control condition) engaged in unrelated play activity. All training was done individually. The day after training, the children were shown a "target" block for 5 seconds and then asked to identify the block in a group of seven blocks. This task was repeated six times using different target blocks. The dependent variable was the number of target blocks correctly identified. The following data were obtained. (Experiment suggested by Nelson, G. K. Concomitant effects of visual, motor, and verbal experiences in young children's concept development. *Journal of Educational Psychology.*)

a_1	a_2	a_3	a_4	a_5
0	2	2	2	1
1	3	3	4	0
3	4	4	5	2
1	2	4	3	1
1	1	2	2	1
2	1	1	1	2
2	2	2	3	1
1	2	3	3	0
1	3	2	2	1
2	4	2	4	3

a. [4.2] Perform an exploratory data analysis on these data (see Table 4.2-1 and Figure 4.2-1). Assume that the observations within each treatment level are listed in the order in which the observations were obtained. Interpret the analysis.

b. [4.3] Test the null hypothesis $\mu_1 = \mu_2 = \cdots = \mu_5$; let $\alpha = .05$. Construct an ANOVA table and make a decision about the null hypothesis.

c. [4.4] Compute and interpret $\hat{\omega}^2$ and $\hat{f}$ for these data.

d. [4.5] Calculate the power of the test in part (b).

e. [4.5] Use the results of part (b) as a pilot study and determine the number of subjects required to achieve a power of approximately .80.

f. [4.5] Use Appendix Table E.12 to determine the number of subjects required to detect a medium association; let $1 - \beta = .80$.

g. [4.5] Determine the number of subjects required to achieve a power of .80, where the largest difference among means is $0.95\sigma_\varepsilon$.

h. Prepare a "results and discussion section" for the *Journal of Educational Psychology.*

*6. [4.4] With $\hat{\omega}^2 = \hat{\sigma}_\alpha^2 / (\hat{\sigma}_\varepsilon^2 + \hat{\sigma}_\alpha^2)$ as the starting point where $\hat{\sigma}_\varepsilon^2 = MSWG$ and $\hat{\sigma}_\alpha^2 = [(p-1)/np](MSBG - MSWG)$, derive the computational formula for omega squared:

$$\hat{\omega}^2 = \frac{SSBG - (p-1)MSWG}{SSTO + MSWG}$$

*7. [4.4] With $\hat{\rho}_I = \hat{\sigma}_\alpha^2 / (\hat{\sigma}_\varepsilon^2 + \hat{\sigma}_\alpha^2)$ as the starting point where $\hat{\sigma}_\varepsilon^2 = MSWG$ and $\hat{\sigma}_\alpha^2 = (1/n)(MSBG - MSWG)$, derive the computational formula for the intraclass correlation:

$$\hat{\rho}_I = \frac{MSBG - MSWG}{MSBG + (n-1)MSWG}$$

*8. [4.4] For the following designs, estimate the number of subjects required to achieve a power of .80, where the largest difference among means is equal to $d\sigma_\varepsilon$.

*a. CR-3 design; let $\alpha = .05$ and $d = 0.8$

*b. CR-4 design; let $\alpha = .01$ and $d = 1.2$

*c. CR-2 design; let $\alpha = .05$ and $d = 1.0$

d. CR-3 design; let $\alpha = .01$ and $d = 1.2$

e. CR-4 design; let $\alpha = .05$ and $d = 1.0$

f. CR-5 design; let $\alpha = .01$ and $d = 1.4$

*9. Section 4.2 described an experiment concerning the effects of sleep deprivation on hand-steadiness. Assume that a second sleep deprivation experiment was performed in which the dependent variable was simple reaction time to the onset of a light. The following data (in hundredths of a second) were obtained.

a_1	a_2	a_3	a_4
12 hours	18 hours	24 hours	30 hours
20	21	25	25
20	20	23	22
17	21	22	22
19	22	23	20
20	20	21	22
19	20	22	26
21	23	22	23
19	19	23	23

*a. [4.2] Perform an exploratory data analysis on these data (see Table 4.2-1 and Figure 4.2-1). Assume that the observations within each treatment level are listed in the order in which the observations were obtained. Interpret the analysis.

*b. [4.3] Test the null hypothesis $\mu_1 = \mu_2 = \mu_3 = \mu_4$; let $\alpha = .05$. Construct an ANOVA table and make a decision about the null hypothesis.

*c. [4.4] Compute and interpret $\hat{\omega}^2$ and $\hat{f}$ for these data.

*d. [4.5] Calculate the power of the test in part (b).

*e. [4.5] Use the results of part (b) as a pilot study and determine the number of subjects required to achieve a power of approximately .80.

*f. [4.5] Use Appendix Table E.12 to determine the number of subjects required to detect a medium association; let $1 - \beta = .80$.

*g. [4.5] Determine the number of subjects required to achieve a power of .80, where the largest difference among means is $1.5\sigma_\varepsilon$.

h. Prepare a "results and discussion section" for the *Review of General Psychology.*

10. The effects of viewing "mug shots" on accuracy of eyewitness identification were investigated. Twenty-four subjects observed a videotape of six men who they were later asked to identify in a recognition test. The subjects were randomly assigned to one of four groups. Subjects in group a_4 searched through a sequence of 75 mug shots to identify the suspects, those in group a_3 searched through 50 mug shots, and those in group a_2 searched through 25 mug shots. Subjects in a_1 spent an equivalent amount of time looking for articles about crime in *Time* magazine. Following this, the subjects were shown pictures that included the suspects and asked to identify them. The dependent variable is the number of suspects identified. The following data were obtained.

a_1	a_2	a_3	a_4
5	4	3	0
6	3	0	1
3	6	1	0
4	3	2	2
5	5	2	1
4	4	1	2

a. [4.2] Perform an exploratory data analysis on these data (see Table 4.2-1 and Figure 4.2-1). Assume that the observations within each treatment level are listed in the order in which the observations were obtained. Interpret the analysis.

b. [4.3] Test the null hypothesis $\mu_1 = \mu_2 = \cdots = \mu_4$; let $\alpha = .05$. Construct an ANOVA table and make a decision about the null hypothesis.

c. [4.4] Compute and interpret $\hat{\omega}^2$ and $\hat{f}$ for these data.

d. [4.5] Calculate the power of the test in part (b).

e. [4.5] Use the results of part (b) as a pilot study and determine the number of subjects required to achieve a power of approximately .80.

f. [4.5] Use Appendix Table E.12 to determine the number of subjects required to detect a large association; let $1 - \beta = .80$.

g. [4.5] Determine the number of subjects required to achieve a power of .80, where the largest difference among means is $2.0\sigma_\varepsilon$.

h. Prepare a "results and discussion section" for the *Review of General Psychology.*

*11. [4.6] How do model I and model II differ for a CR-*p* design?