CHAPTER 1

Experimental Design

ROGER E. KIRK

SOME BASIC EXPERIMENTAL DESIGN CONCEPTS

Experimental design is concerned with the skillful interrogation of nature. Unfortunately, nature is reluctant to reveal her secrets. Joan Fisher Box (1978) observed in her biography of her famous father, Ronald A. Fisher, "Far from behaving consistently, however, Nature appears vacillating, coy, and ambiguous in her answers" (p. 140). Nature's most effective tool for confusing researchers is variability—in particular, variability among subjects or experimental units. Although nature can be duplicitous, Ronald A. Fisher showed that by comparing the variability among subjects treated differently to the variability among subjects treated alike, researchers can make informed choices between competing hypotheses in science and technology.

We must never underestimate nature—she is a formidable foe. Carefully designed and executed experiments are required to learn her secrets. An *experimental design* is a plan for assigning subjects to experimental conditions and the statistical analysis associated with the plan (Kirk, 2012, p. 1). The design of an experiment involves five interrelated activities:

1. Formulation of statistical hypotheses that are germane to the scientific hypothesis. A statistical hypothesis is a statement about (a) one or more parameters of a population or (b) the functional form of a population. Statistical hypotheses are rarely identical to scientific hypotheses—they are testable formulations of scientific hypotheses.
2. Determination of the experimental conditions (independent variable) to be manipulated, the measurement (dependent variable) to be recorded, and the extraneous conditions (nuisance variables) that must be controlled.
3. Specification of the number of subjects required and the population from which they will be sampled.
4. Specification of the procedure for assigning the subjects to the experimental conditions.
5. Determination of the statistical analysis that will be performed.

In short, an experimental design identifies the independent, dependent, and nuisance variables and indicates the way in which the randomization and statistical aspects of an experiment are to be carried out.

Analysis of Variance

Analysis of variance (ANOVA) is a procedure for decomposing the total variation displayed by a set of observations into two or more identifiable sources of variation. The procedure enables researchers to interpret the variability in designed experiments. The seminal ideas for both ANOVA and experimental design can be traced to Ronald A. Fisher, a statistician, eugenicist, evolutionary biologist, and geneticist who worked at the Rothamsted Experimental Station that is 25 miles northwest of London. According to Box (1978, p. 100), Fisher developed the basic ideas of ANOVA between 1919 and 1925. The first hint of what was to come appeared in a 1918 paper in which Fisher partitioned the total variance of a human

attribute into portions attributed to heredity, environment, and other factors. The analysis of variance table for a two-treatment factorial design first appeared in a 1923 paper published with M. A. Mackenzie (Fisher & Mackenzie, 1923). Fisher referred to the table as a convenient way of arranging the arithmetic. In 1924, Fisher (1925) introduced the Latin square design in connection with a forest nursery experiment. The publication in 1925 of his classic textbook *Statistical Methods for Research Workers* and a short paper the following year (Fisher, 1926) presented all the essential ideas of analysis of variance. The textbook (Fisher, 1925, pp. 244–249) included a table of the critical values of the ANOVA test statistic in terms of a function called z, where $z = 1/2(\ln\hat{\sigma}^2_{\text{Treatment}} - \ln\hat{\sigma}^2_{\text{Error}})$. The statistics $\hat{\sigma}^2_{\text{Treatment}}$ and $\hat{\sigma}^2_{\text{Error}}$ denote, respectively, treatment and error variance. A more convenient form of Fisher's z table that did not require looking up log values was developed by George Snedecor (1934). His critical values are expressed in terms of the function $F = \hat{\sigma}^2_{\text{Treatment}}/\hat{\sigma}^2_{\text{Error}}$ that is obtained directly from the ANOVA calculations. He named it F in honor of Fisher. Fisher's field of experimentation—agriculture—was a fortunate choice because results had immediate application with assessable economic value, because simplifying assumptions such as normality and independence of errors were usually tenable, and because the cost of conducting experiments was modest.

Three Principles of Good Experimental Design

The publication of Fisher's *Statistical Methods for Research Workers* and his 1935 *The Design of Experiments* gradually led to the acceptance of what today is considered to be the cornerstone of good experimental design: randomization. It is hard to imagine the hostility that greeted the suggestion that subjects or experimental units should be randomly assigned to treatment levels. Before Fisher's work, most researchers used systematic schemes, not subject to the laws of chance, to assign subjects. According to Fisher, random assignment has several purposes. It helps to distribute the idiosyncratic characteristics of subjects over the treatment levels so that they do not selectively bias the outcome of the experiment. Also, random assignment permits the computation of an unbiased estimate of error effects—those effects not attributable to the manipulation of the independent variable—and it helps to ensure that the error effects are statistically independent.

Fisher popularized two other principles of good experimentation: replication and local control or blocking. *Replication* is the observation of two or more subjects under identical experimental conditions. Fisher observed that replication enables a researcher to estimate error effects and to obtain a more precise estimate of treatment effects. *Blocking*, on the other hand, is an experimental procedure for isolating variation attributable to a nuisance variable. As the name suggests, *nuisance variables* are undesired sources of variation that can affect the dependent variable. There are many sources of nuisance variation. Differences among subjects comprise one source. Other sources include variation in the presentation of instructions to subjects, changes in environmental conditions, and the effects of maturation, fatigue, and learning when subjects are observed several times. Three experimental approaches are used to deal with nuisance variables:

1. Holding the variable constant.
2. Assigning subjects randomly to the treatment levels so that known and unsuspected sources of variation among the subjects are distributed over the entire experiment and do not affect the subjects in just one or a limited number of treatment levels.
3. Including the nuisance variable as one of the factors in the experiment.

The third experimental approach uses local control or blocking to isolate variation attributable to the nuisance variable so that it does not appear in estimates of treatment and error effects. A statistical approach also can be used to deal with nuisance variables. The approach is called *analysis of covariance* and is described in the last section of this chapter. The three principles that Fisher vigorously championed—randomization, replication, and local control—remain the cornerstones of good experimental design.

THREE BUILDING BLOCK DESIGNS

In this section I describe three simple analysis of variance designs that can be combined to form more complex designs. They are the completely randomized design, the randomized block design, and the Latin square design. I call these designs *building block designs*.

Completely Randomized Design

One of the simplest experimental designs is the randomization and analysis plan that is used with a t statistic for independent samples. Consider an experiment to compare the effectiveness of two diets for obese teenagers. The

independent variable is the two kinds of diets; the dependent variable is the amount of weight loss two months after going on a diet. For notational convenience, the two diets are called *Treatment A*. The levels of Treatment *A* corresponding to the specific diets are denoted by the lowercase letter a and a subscript: a_1 denotes one diet and a_2 denotes the other diet. A particular but unspecified level of Treatment *A* is denoted by a_j, where j ranges over the values 1 and 2. The amount of weight loss in pounds 2 months after subject i went on diet j is denoted by Y_{ij}.

The null and alternative hypotheses for the weight-loss experiment are, respectively,

$$H_0\colon \mu_1 - \mu_2 = 0$$
$$H_1\colon \mu_1 - \mu_2 \neq 0$$

where μ_1 and μ_2 denote the mean weight loss of the respective populations. Assume that 30 girls who want to lose weight are available to participate in the experiment. The researcher assigns $n = 15$ girls to each of the $p = 2$ diets so that each of the $(np)!/(n!)^p = 155{,}117{,}520$ possible assignments has the same probability. This is accomplished by numbering the girls from 1 to 30 and drawing numbers from a random numbers table. The first 15 numbers drawn between 1 and 30 are assigned to treatment level a_1; the remaining 15 numbers are assigned to a_2. The layout for this experiment is shown in Figure 1.1. The girls who are assigned to treatment level a_1 are called Group_1; those assigned to treatment level a_2 are called Group_2. The mean weight losses of the two groups of girls are denoted by $\overline{Y}_{\cdot 1}$ and $\overline{Y}_{\cdot 2}$.

		[a]Treat. Level	[b]Dep. Var.
Group_1	Subject_1	a_1	Y_{11}
	Subject_2	a_1	Y_{21}
	⋮	⋮	⋮
	Subject_{15}	a_1	$Y_{15,1}$
			$\overline{Y}_{\cdot 1}$
Group_2	Subject_1	a_2	Y_{12}
	Subject_2	a_2	Y_{22}
	⋮	⋮	⋮
	Subject_{15}	a_2	$Y_{15,2}$
			$\overline{Y}_{\cdot 2}$

[a]Treatment

[b]Dependent Variable

Figure 1.1 Layout for a t independent-samples design. Thirty girls are randomly assigned to two levels of Treatment *A* with the restriction that 15 girls are assigned to each level. The mean weight loss in pounds for the girls in treatment levels a_1 and a_2 is denoted by $\overline{Y}_{\cdot 1}$ and $\overline{Y}_{\cdot 2}$, respectively.

The t independent-samples design involves randomly assigning subjects to two levels of a treatment. A completely randomized design, which is described next, extends this design strategy to two or more treatment levels. The *completely randomized design* is denoted by the letters CR-p, where CR stands for *completely randomized* and p is the number of levels of the treatment.

Again, consider the weight-loss experiment and suppose that the researcher wants to evaluate the effectiveness of three diets. The null and alternative hypotheses for the experiment are, respectively,

$$H_0\colon \mu_1 = \mu_2 = \mu_3$$
$$H_1\colon \mu_j \neq \mu_{j'} \text{ for some } j \text{ and } j'$$

Assume that 45 girls who want to lose weight are available to participate in the experiment. The girls are randomly assigned to the three diets with the restriction that 15 girls are assigned to each diet. The layout for the experiment is shown in Figure 1.2. A comparison of the layout in this figure with that in Figure 1.1 for a t independent-samples design reveals that they are the same except that the completely randomized design has an additional treatment level. The t independent-samples design can be thought of as a special case of a completely randomized design. When p is equal to two, the layouts and randomization plans for the designs are identical.

Thus far I have identified the null hypothesis that the researcher wants to test, $\mu_1 = \mu_2 = \mu_3$, and described the

		Treat. Level	Dep. Var.
Group_1	Subject_1	a_1	Y_{11}
	Subject_2	a_1	Y_{21}
	⋮	⋮	⋮
	Subject_{15}	a_1	$Y_{15,1}$
			$\overline{Y}_{\cdot 1}$
Group_2	Subject_1	a_2	Y_{12}
	Subject_2	a_2	Y_{22}
	⋮	⋮	⋮
	Subject_{15}	a_2	$Y_{15,2}$
			$\overline{Y}_{\cdot 2}$
Group_3	Subject_1	a_3	Y_{13}
	Subject_2	a_3	Y_{23}
	⋮	⋮	⋮
	Subject_{15}	a_3	$Y_{15,3}$
			$\overline{Y}_{\cdot 3}$

Figure 1.2 Layout for a completely randomized design (CR-3 design). Forty-five girls are randomly assigned to three levels of Treatment *A* with the restriction that 15 girls are assigned to each level. The mean weight loss in pounds for the girls in treatment levels a_1, a_2, and a_3 is denoted by $\overline{Y}_{\cdot 1}$, $\overline{Y}_{\cdot 2}$, and $\overline{Y}_{\cdot 3}$, respectively.

manner in which the subjects are assigned to the three treatment levels. In the following paragraphs, I discuss the composite nature of an observation, describe the classical model equation for a CR-p design, and examine the meaning of the terms *treatment effect* and *error effect*.

An observation, which is a measure of the dependent variable, can be thought of as a composite that reflects the effects of the (a) independent variable, (b) individual characteristics of the subject or experimental unit, (c) chance fluctuations in the subject's performance, (d) measurement and recording errors that occur during data collection, and (e) any other nuisance variables such as environmental conditions that have not been controlled. Consider the weight loss of the fifth subject in treatment level a_2. Suppose that two months after beginning the diet this girl has lost 13 pounds ($Y_{52} = 13$). What factors have affected the value of Y_{52}? One factor is the effectiveness of the diet. Other factors are her weight prior to starting the diet, the degree to which she stayed on the diet, and the amount she exercised during the 2-month trial, to mention only a few. In summary, Y_{52} is a composite that reflects (a) the effects of treatment level a_2, (b) effects unique to the subject, (c) effects attributable to chance fluctuations in the subject's behavior, (d) errors in measuring and recording the subject's weight loss, and (e) any other effects that have not been controlled. My conjectures about Y_{52} or any of the other 44 observations can be expressed more formally by a model equation. The classical model equation for the completely randomized design is

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

where

- Y_{ij} is the weight loss for subject i in treatment level a_j.
- μ is the population grand mean of the three weight-loss population means.
- α_j is the treatment effect for population j and is equal to $\mu_j - \mu$. α_j reflects the effects of diet a_j.
- $\varepsilon_{i(j)}$ is the within-groups error effect associated with Y_{ij} and is equal to $Y_{ij} - \mu - \alpha_j$.

$\varepsilon_{i(j)}$ reflects all effects not attributable to treatment level a_j. The notation $i(j)$ indicates that the ith subject appears only in treatment level j. Subject i is said to be nested within the jth treatment level. Nesting is discussed in the section titled "Hierarchical Designs."

According to the equation for the completely randomized design, each observation is the sum of three parameters μ, α_j, and $\varepsilon_{i(j)}$. The values of the parameters are unknown, but they can be estimated from sample data as follows:

Population parameters	μ	$+ \quad \alpha_j \quad +$		$\varepsilon_{i(j)}$
Sample estimators	$\overline{Y}_{..}$	$+ \ (\overline{Y}_{.j} - \overline{Y}_{..}) \ +$		$(Y_{ij} - \overline{Y}_{.j})$

where $\overline{Y}_{..}$ is the mean of the np observations and $\overline{Y}_{.j}$ is the mean of the n observations in the jth treatment level.

The meaning of the terms *grand mean*, μ, and *treatment effect*, α_j, in the model equation seems fairly clear; the meaning of the error effect, $\varepsilon_{i(j)}$, requires a bit more explanation. Why do observations, Y_{ij}s, in the same treatment level vary from one girl to the next? This variation must be due to differences among the girls and to other uncontrolled variables because the parameters μ and α_j in the model equation are constants for all of the girls in the same treatment level. To put it another way, observations in the same treatment level are different because the error effects, $\varepsilon_{i(j)}$s, for the observations are different. Recall that error effects reflect idiosyncratic characteristics of the subjects—those characteristics that differ from one subject to another—and any other variables that have not been controlled. Researchers attempt to minimize the size of error effects by holding sources of variation that might contribute to the error effects constant and by the judicial choice of an experimental design. Designs that are described next permit a researcher to isolate and remove some sources of variation that would ordinarily be included in the error effects.

Randomized Block Design

The two designs just described use independent samples. Two samples are independent if, for example, a researcher randomly samples from two populations or randomly assigns subjects to p groups. Dependent samples, however, can be obtained by any of the following four procedures:

1. Observe each subject under each treatment level in the experiment—that is, obtain repeated measures on the subjects.
2. Form sets of subjects who are similar with respect to a variable that is correlated with the dependent variable. This procedure is called *subject matching*.
3. Obtain sets of identical twins or littermates in which case the subjects have similar genetic characteristics.
4. Obtain subjects who are matched by mutual selection, for example, husband and wife pairs or business partners.

In the behavioral and social sciences, the characteristics of subjects such as aptitudes, interests, values, and experiences can differ markedly. Individual differences are inevitable, but it is often possible to isolate or partition out a portion of these effects so that they do not appear in estimates of the error effects. One design for accomplishing this is the design used with a t statistic for dependent samples. As the name suggests, the design uses dependent samples. A t dependent-samples design also uses a more complex randomization and analysis plan than a t independent-samples design. However, the added complexity is often accompanied by greater power—a point that I will develop later in connection with a randomized block design.

Let us reconsider the weight-loss experiment. It is reasonable to assume that the ease of losing weight is related to the amount by which a girl is overweight. The design of the experiment can be improved by isolating this nuisance variable. Suppose that instead of randomly assigning 30 girls to the treatment levels, the researcher formed pairs of girls who are similar in obesity prior to beginning a diet. The girls in each pair constitute a block or set of matched subjects. A simple way to form blocks of matched subjects is to rank them from least to most overweight. The girls ranked 1 and 2 are assigned to block one, those ranked 3 and 4 are assigned to block two, and so on. In this example, 15 blocks of dependent samples can be formed from the 30 girls. After all of the blocks have been formed, the two girls in each block are randomly assigned to the two diets. The layout for this experiment is shown in Figure 1.3. If the researcher's hunch is correct—that ease in losing weight is related to the amount by which a girl is overweight—this design should result in a more powerful test of the null hypothesis, $\mu_1 - \mu_2 = 0$, than a t test for independent samples. As I show later, the increased power results from isolating the nuisance variable (the amount by which the girls are overweight) so that it does not appear in the estimate of the error effects.

	Treat. Level	Dep. Var.	Treat. Level	Dep. Var.
Block_1	a_1	Y_{11}	a_2	Y_{12}
Block_2	a_1	Y_{21}	a_2	Y_{22}
Block_3	a_1	Y_{31}	a_2	Y_{32}
⋮	⋮	⋮	⋮	⋮
Block_{15}	a_1	$Y_{15,1}$	a_2	$Y_{15,2}$
		$\overline{Y}_{\cdot 1}$		$\overline{Y}_{\cdot 2}$

Figure 1.3 Layout for a t dependent-samples design. Each block contains two girls who are overweight by about the same amount. The two girls in a block are randomly assigned to the treatment levels. The mean weight loss in pounds for the girls in treatment levels a_1 and a_2 is denoted by $\overline{Y}_{\cdot 1}$ and $\overline{Y}_{\cdot 2}$, respectively.

Earlier I showed that the layout and randomization procedures for a t independent-samples design and a completely randomized design are the same except that a completely randomized design can have more than two treatment levels. The same comparison can be drawn between a t dependent-samples design and a randomized block design. A *randomized block design* is denoted by the letters RB-p, where RB stands for *randomized block* and p is the number of levels of the treatment. The four procedures for obtaining dependent samples that were described earlier can be used to form the blocks in a randomized block design.

The procedure that is used to form the blocks does not affect the computation of significance tests, but the procedure does affect the interpretation of the results. The results of an experiment with repeated measures generalize to a population of subjects who have been exposed to all of the treatment levels. However, the results of an experiment with matched subjects generalize to a population of subjects who have been exposed to only one treatment level. Some writers reserve the designation *randomized block design* for this latter case. They refer to a design with repeated measurements in which the order of administration of the treatment levels is randomized independently for each subject as a *subjects-by-treatments design*. A design with repeated measurements in which the order of the administration of the treatment levels is the same for all subjects is referred to as a *subject-by-trials design*. I use the designation *randomized block design* for all three cases.

Of the four ways of obtaining dependent samples, the use of repeated measures on the subjects typically results in the greatest homogeneity within the blocks. However, if repeated measures are used, the effects of one treatment level should dissipate before the subject is observed under another treatment level. Otherwise the subsequent observations will reflect the cumulative effects of the preceding treatment levels. There is no such restriction, of course, if carryover effects such as learning or fatigue are the researcher's principal interest. If the blocks are composed of identical twins or littermates, it is assumed that the performance of subjects having identical or similar heredities will be more homogeneous than the performance of subjects having dissimilar heredities. If the blocks are composed of subjects who are matched by mutual selection (e.g., husband and wife pairs or business partners), a researcher should ascertain that the subjects in a block

are in fact more homogeneous with respect to the dependent variable than are unmatched subjects. A husband and wife often have similar political attitudes; the couple is less likely to have similar mechanical aptitudes.

Suppose that in the weight-loss experiment, the researcher wants to evaluate the effectiveness of three diets, denoted by a_1, a_2, and a_3. The researcher suspects that the ease of losing weight is related to the amount by which a girl is overweight. If a sample of 45 girls is available, the blocking procedure described in connection with a t dependent-samples design can be used to form 15 blocks of subjects. The three girls in a block are matched with respect to the nuisance variable, amount by which a girl is overweight. The layout for the experiment is shown in Figure 1.4. A comparison of the layout in this figure with that in Figure 1.3 for a t dependent-samples design reveals that they are the same except that the randomized block design has $p = 3$ treatment levels. When $p = 2$, the layouts and randomization plans for the designs are identical. In this and later examples, I assume that all of the treatment levels and blocks of interest are represented in the experiment. In other words, the treatment levels and blocks represent fixed effects. A discussion of the case in which either the treatment levels or blocks or both are randomly sampled from a population of levels, the mixed and random effects cases, is beyond the scope of this chapter. The reader is referred to Kirk (2012, pp. 285–286, 296–299).

A randomized block design enables a researcher to test two null hypotheses:

H_0: $\mu_{\cdot 1} = \mu_{\cdot 2} = \mu_{\cdot 3}$ (Treatment population means are equal.)

	Treat. Level	Dep. Var.	Treat. Level	Dep. Var.	Treat. Level	Dep. Var.	
Block_1	a_1	Y_{11}	a_2	Y_{12}	a_3	Y_{13}	$\overline{Y}_{1\cdot}$
Block_2	a_1	Y_{21}	a_2	Y_{22}	a_3	Y_{23}	$\overline{Y}_{2\cdot}$
Block_3	a_1	Y_{31}	a_2	Y_{32}	a_3	Y_{33}	$\overline{Y}_{3\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Block_{15}	a_1	$Y_{15,1}$	a_2	$Y_{15,2}$	a_3	$Y_{15,3}$	$\overline{Y}_{15,\cdot}$
		$\overline{Y}_{\cdot 1}$		$\overline{Y}_{\cdot 2}$		$\overline{Y}_{\cdot 3}$	

Figure 1.4 Layout for a randomized block design (RB-3 design). Each block contains three girls who are overweight by about the same amount. The three girls in a block are randomly assigned to the treatment levels. The mean weight loss in pounds for the girls in treatment levels a_1, a_2, and a_3 is denoted by $\overline{Y}_{\cdot 1}$, $\overline{Y}_{\cdot 2}$, and $\overline{Y}_{\cdot 3}$, respectively. The mean weight loss for the girls in Block_1, $\text{Block}_2, \ldots,$ Block_{15} is denoted by $\overline{Y}_{1\cdot}$, $\overline{Y}_{2\cdot}, \ldots,$ $\overline{Y}_{15\cdot}$, respectively.

H_0: $\mu_{1\cdot} = \mu_{2\cdot} = \cdots = \mu_{15\cdot}$ (Block population means are equal.)

The second hypothesis, which is usually of little interest, states that the population weight-loss means for the 15 levels of the nuisance variable are equal. The researcher expects a test of this null hypothesis to be significant. If the nuisance variable represented by the blocks does not account for an appreciable proportion of the total variation in the experiment, little has been gained by isolating the effects of the variable. Before exploring this point, I describe the model equation for an RB-p design.

The classical model equation for the randomized block design is

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

where

- Y_{ij} is the weight loss for the subject in Block_i and treatment level a_j.
- μ is the grand mean of the three weight-loss population means.
- α_j is the treatment effect for population j and is equal to $\mu_{\cdot j} - \mu$. α_j reflects the effect of diet a_j.
- π_i is the block effect for population i and is equal to $\mu_{i\cdot} - \mu$. π_i reflects the effect of the nuisance variable in Block_i.
- ε_{ij} is the residual error effect associated with Y_{ij} and is equal to $Y_{ij} - \mu - \alpha_j - \pi_i$. ε_{ij} reflects all effects not attributable to treatment level a_j and Block_i.

According to the model equation for the randomized block design, each observation is the sum of four parameters: μ, α_j, π_i, and ε_{ij}. A residual error effect is that portion of an observation that remains after the grand mean, treatment effect, and block effect have been subtracted from it; that is, $\varepsilon_{ij} = Y_{ij} - \mu - \alpha_j - \pi_i$. The sum of the squared error effects for the randomized block design,

$$\sum\sum \varepsilon_{ij}^2 = \sum\sum (Y_{ij} - \mu - \alpha_j - \pi_i)^2$$

will be smaller than the sum for the completely randomized design,

$$\sum\sum \varepsilon_{i(j)}^2 = \sum\sum (Y_{ij} - \mu - \alpha_j)^2$$

if π_i^2 is not equal to zero for one or more blocks. This idea is illustrated in Figure 1.5 where the total sum of squares

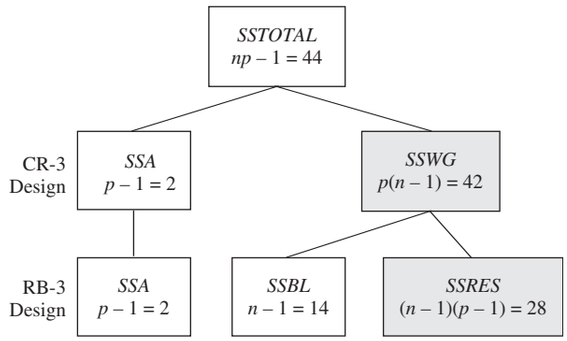

Figure 1.5 Partition of the total sum of squares (*SSTOTAL*) and degrees of freedom ($np - 1 = 44$) for CR-3 and RB-3 designs. The treatment and within-groups sums of squares are denoted by, respectively, *SSA* and *SSWG*. The block and residual sums of squares are denoted by, respectively, *SSBL* and *SSRES*. The shaded rectangles indicate the sums of squares that are used to compute the error variance for each design: $MSWG = SSWG/p(n-1)$ and $MSRES = SSRES/(n-1)(p-1)$. If the nuisance variable (*SSBL*) in the randomized block design accounts for an appreciable portion of the total sum of squares, the design will have a smaller error variance and, hence, greater power than the completely randomized design.

and degrees of freedom for the two designs are partitioned. The F statistic that is used to test the null hypothesis can be thought of as a ratio of error and treatment effects,

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

where $f(\,)$ denotes a function of the effects in parentheses. It is apparent from an examination of this ratio that the smaller the sum of the squared error effects, the larger the F statistic and, hence, the greater the probability of rejecting a false null hypothesis. Thus, by isolating a nuisance variable that accounts for an appreciable portion of the total variation in a randomized block design, a researcher is rewarded with a more powerful test of a false null hypothesis.

As I have shown, blocking with respect to the nuisance variable (the amount by which the girls are overweight) enables the researcher to isolate this variable and remove it from the error effects. But what if the nuisance variable doesn't account for any of the variation in the experiment? In other words, what if all of the block effects in the experiment are equal to zero? In this unlikely case, the sum of the squared error effects for the randomized block and completely randomized designs will be equal. In this case, the randomized block design will be less powerful than the completely randomized design because its error variance, the denominator of the F statistic, has $n - 1$ fewer degrees of freedom than the error variance for the completely randomized design. It should be obvious that the nuisance variable must be selected with care. The larger the correlation between the nuisance variable and the dependent variable, the more likely it is that the block effects will account for an appreciable proportion of the total variation in the experiment.

Latin Square Design

The Latin square design described in this section derives its name from an ancient puzzle that was concerned with the number of different ways that Latin letters can be arranged in a square matrix so that each letter appears once in each row and once in each column. An example of a 3×3 Latin square is shown in Figure 1.6. In this figure I have used the letter a with subscripts in place of Latin letters. The Latin square design is denoted by the letters LS-p, where LS stands for *Latin square* and p is the number of levels of the treatment. The Latin square design enables a researcher to isolate the effects of not one but two nuisance variables. The levels of one nuisance variable are assigned to the rows of the square; the levels of the other nuisance variable are assigned to the columns. The levels of the treatment are assigned to the cells of the square.

Let us return to the weight-loss experiment. With a Latin square design, the researcher can isolate the effects of the amount by which girls are overweight and the effects of a second nuisance variable, for example, genetic predisposition to be overweight. A rough measure of the second nuisance variable can be obtained by asking a girl's parents whether they were overweight as teenagers: c_1 denotes neither parent overweight, c_2 denotes one parent overweight, c_3 denotes both parents overweight. This nuisance variable can be assigned to the columns of the Latin square. Three levels of the amount by which girls are overweight can be assigned to the rows of the Latin square: b_1 is less than 15 pounds, b_2 is 15 to 25 pounds, and b_3 is more than 25 pounds. The advantage of being able to isolate two

	c_1	c_2	c_3
b_1	a_1	a_2	a_3
b_2	a_2	a_3	a_1
b_3	a_3	a_1	a_2

Figure 1.6 Three-by-three Latin square, where a_j denotes one of the $j = 1, \ldots, p$ levels of Treatment A, b_k denotes one of the $k = 1, \ldots, p$ levels of nuisance variable B, and c_l denotes one of the $l = 1, \ldots, p$ levels of nuisance variable C. Each level of Treatment A appears once in each row and once in each column as required for a Latin square.

nuisance variables comes at a price. The randomization procedures for a Latin square design are more complex than those for a randomized block design. Also, the number of rows and columns of a Latin square must each equal the number of treatment levels, which is three in the example. This requirement is very restrictive. For example, it was necessary to restrict the continuous variable of the amount by which girls are overweight to only three levels. The layout of the LS-3 design is shown in Figure 1.7.

The design in Figure 1.7 enables the researcher to test three null hypotheses:

H_0: $\mu_{1..} = \mu_{2..} = \mu_{3..}$ (Treatment population means are equal.)

H_0: $\mu_{.1.} = \mu_{.2.} = \mu_{.3.}$ (Row population means are equal.)

		[a]Treat. Comb.	[b]Dep. Var.
Group_1	Subject_1	$a_1b_1c_1$	Y_{1111}
	⋮	⋮	⋮
	Subject_{15}	$a_1b_1c_1$	$Y_{15,111}$
			$\overline{Y}_{\cdot 111}$
Group_2	Subject_1	$a_1b_2c_3$	Y_{1123}
	⋮	⋮	⋮
	Subject_{15}	$a_1b_2c_3$	$Y_{15,123}$
			$\overline{Y}_{\cdot 123}$
Group_3	Subject_1	$a_1b_3c_2$	Y_{1132}
	⋮	⋮	⋮
	Subject_{15}	$a_1b_3c_2$	$Y_{15,132}$
			$\overline{Y}_{\cdot 132}$
Group_4	Subject_1	$a_2b_1c_2$	Y_{1212}
	⋮	⋮	⋮
	Subject_{15}	$a_2b_1c_2$	$Y_{15,212}$
			$\overline{Y}_{\cdot 212}$
⋮	⋮	⋮	⋮
Group_9	Subject_1	$a_3b_3c_1$	Y_{1331}
	⋮	⋮	⋮
	Subject_{15}	$a_3b_3c_1$	$Y_{15,331}$
			$\overline{Y}_{\cdot 331}$

[a]Treatment Combination

[b]Dependent Variable

Figure 1.7 Layout for a Latin square design (LS-3 design) that is based on the Latin square in Figure 1.6. Treatment *A* represents three kinds of diets, nuisance variable *B* represents amount by which the girls are overweight, and nuisance variable *C* represents genetic predisposition to be overweight. The girls in Group_1, for example, received diet a_1, were less than 15 pounds overweight (b_1), and neither parent had been overweight as a teenager (c_1). The mean weight loss in pounds for the girls in the nine groups is denoted by $\overline{Y}_{\cdot 111}$, $\overline{Y}_{\cdot 123}, \ldots, \overline{Y}_{\cdot 331}$.

H_0: $\mu_{..1} = \mu_{..2} = \mu_{..3}$ (Column population means are equal.)

The first hypothesis states that the population means for the three diets are equal. The second and third hypotheses make similar assertions about the population means for the two nuisance variables. Tests of the nuisance variables are expected to be significant. As discussed earlier, if the nuisance variables do not account for an appreciable proportion of the total variation in the experiment, little has been gained by isolating the effects of the variables.

The classical Latin square model equation for this version of the weight-loss experiment is

$$Y_{ijkl} = \mu + \alpha_j + \beta_k + \gamma_l + \varepsilon_{jkl} + \varepsilon_{i(jkl)}$$
$$(i = 1, \ldots, n;\ j = 1, \ldots, p;\ k = 1, \ldots, p;\ l = 1, \ldots, p)$$

where

Y_{ijkl} is the weight loss for the ith subject in treatment level a_j, row b_k, and column c_l.

μ is the grand mean of the $p^2 = 9$ population means.

α_j is the treatment effect for population j and is equal to $\mu_{j..} - \mu$. α_j reflects the effect of diet a_j.

β_k is the row effect for population k and is equal to $\mu_{.k.} - \mu$. β_k reflects the effect of nuisance variable b_k.

γ_l is the column effect for population l and is equal to $\mu_{..l} - \mu$. γ_l reflects the effects of nuisance variable c_l.

ε_{jkl} is the residual effect that is equal to $\mu_{jkl} - \mu_{j..} - \mu_{.k.} - \mu_{..l} + 2\mu$.

$\varepsilon_{i(jkl)}$ is the within-cell error effect associated with Y_{ijkl} and is equal to $Y_{ijkl} - \mu - \alpha_j - \beta_k - \gamma_l - \varepsilon_{jkl}$.

According to the model equation for the Latin square design, each observation is the sum of six parameters: μ, α_j, β_k, γ_l, ε_{jkl}, and $\varepsilon_{i(jkl)}$. The sum of the squared within-cell error effects for the Latin square design,

$$\sum\sum \varepsilon^2_{i(jkl)} = \sum\sum (Y_{ijkl} - \mu - \alpha_j - \beta_k - \gamma_l - \varepsilon_{jkl})^2$$

will be smaller than the sum for the randomized block design,

$$\sum\sum \varepsilon^2_{ij} = \sum\sum (Y_{ij} - \mu - \alpha_j - \pi_i)^2$$

if the combined effects of $\sum \beta_k^2$, $\sum \gamma_l^2$, and $\sum \varepsilon_{jkl}^2$ are greater than $\sum \pi_i^2$. The benefit of isolating two nuisance variables is a smaller error variance and increased power.

Thus far I have described three of the simplest experimental designs: the completely randomized design, randomized block design, and Latin square design. I call the three designs *building block designs* because complex experimental designs can be constructed by combining two or more of these simple designs (Kirk, 2012, p. 42). Furthermore, the randomization procedures, data analysis, and model assumptions for complex designs represent extensions of those for the three building block designs. The three designs provide the organizational structure for the design nomenclature and classification scheme that is described next.

CLASSIFICATION OF EXPERIMENTAL DESIGNS

A classification scheme for experimental designs is given in Table 1.1. The designs in the category *systematic designs* do not use random assignment of subjects or experimental units and are of historical interest only. According to Leonard and Clark (1939), agricultural field research employing systematic designs on a practical scale dates back to 1834. Over the past 90 years systematic designs have fallen into disuse because designs employing random assignment are more likely to provide valid estimates of treatment and error effects and can be analyzed using the powerful tools of statistical inference such as the analysis of variance. Experimental designs that use random assignment are called *randomized designs*. The randomized designs in Table 1.1 are subdivided into categories based on (a) the number of treatments, (b) whether subjects are assigned to relatively homogeneous blocks prior to random assignment, (c) presence or absence of confounding, (d) use of crossed or nested treatments, and (e) use of a covariate.

The letters p and q in the abbreviated designations denote the number of levels of Treatments A and B, respectively. If a design includes a third and fourth treatment, say Treatments C and D, the number of their levels is denoted by r and t, respectively. In general, the designation for designs with two or more treatments includes the letters CR, RB, or LS to indicate the building block design. The letter F or H is added to the designation to indicate that the design is, respectively, a factorial design or a hierarchical design. For example, the F in the designation CRF-pq indicates that it is a factorial design; the CR and pq indicate that

TABLE 1.1 Classification of Experimental Designs

Experimental Design	Abbreviated Designation[a]
I. Systematic designs (selected examples)	
1. Beavan's chessboard design	
2. Beavan's half-drill strip design	
3. Diagonal square design	
4. Knut Vik square design	
II. Randomized designs with one treatment	
A. Experimental units randomly assigned to treatment levels	
1. Completely randomized design	CR-p
B. Experimental units assigned to relatively homogeneous blocks or groups prior to random assignment	
1. Balanced incomplete block design	BIB-p
2. Crossover design	CO-p
3. Generalized randomized block design	GRB-p
4. Graeco-Latin square design	GLS-p
5. Hyper-Graeco-Latin square design	HGLS-p
6. Latin square design	LS-p
7. Lattice balanced incomplete block design	LBIB-p
8. Lattice partially balanced incomplete block design	LPBIB-p
9. Lattice unbalanced incomplete block design	LUBIB-p
10. Partially balanced incomplete block design	PBIB-p
11. Randomized block design	RB-p
12. Youden square design	YBIB-p
III. Randomized designs with two or more treatments	
A. Factorial designs: Designs in which all treatments are crossed	
1. Designs without confounding	
a. Completely randomized factorial design	CRF-pq
b. Generalized randomized block factorial design	GRBF-pq
c. Randomized block factorial design	RBF-pq
2. Design with group-treatment confounding	
a. Split-plot factorial design	SPF-$p.q$
3. Designs with group-interaction confounding	
a. Latin square confounded factorial design	LSCF-p^k
b. Randomized block completely confounded factorial design	RBCF-p^k
c. Randomized block partially confounded factorial design	RBPF-p^k
4. Designs with treatment–interaction confounding	
a. Completely randomized fractional factorial design	CRFF-p^{k-i}
b. Graeco-Latin square fractional factorial design	GLSFF-p^k
c. Latin square fractional factorial design	LSFF-p^k
d. Randomized block fractional factorial design	RBFF-p^{k-i}

(*continued*)

TABLE 1.1 (*Continued*)

Experimental Design	Abbreviated Designation[a]
B. Hierarchical designs: Designs in which one or more treatments are nested	
1. Designs with complete nesting	
a. Completely randomized hierarchical design	CRH-$pq(A)$
b. Randomized block hierarchical design	RBH-$pq(A)$
2. Designs with partial nesting	
a. Completely randomized partial hierarchical design	CRPH-$pq(A)r$
b. Randomized block partial hierarchical design	RBPH-$pq(A)r$
c. Split-plot partial hierarchical design	SPH-$p.qr(B)$
IV. Randomized designs with one or more covariates	
A. Designs that include a covariate have the letters AC added to the abbreviated designation as in the following examples.	
1. Completely randomized analysis of covariance design	CRAC-p
2. Completely randomized factorial analysis of covariance design	CRFAC-pq
3. Latin square analysis of covariance design	LSAC-p
4. Randomized block analysis of covariance design	RBAC-p
5. Split-plot factorial analysis of covariance design	SPFAC-$p.q$
V. Miscellaneous designs	
1. Solomon four-group design	
2. Interrupted time-series design	

[a]*Note*. Abbreviated designations are discussed in the text.

the design was constructed by combining two completely randomized designs with p and q treatment levels. The letters CF, PF, FF, and AC are added to the designation if the design is, respectively, a confounded factorial design, partially confounded factorial design, fractional factorial design, or analysis of covariance design. Three of these designs are described later. Because of space limitations, I cannot describe all of the designs in Table 1.1. I focus on those designs that are potentially the most useful in the behavioral and social sciences.

It is apparent from Table 1.1 that a wide array of designs is available to researchers. Unfortunately, there is no universally accepted designation for the various designs—some designs have as many as five different names. For example, the completely randomized design has been called a one-way classification design, single-factor design, randomized group design, simple randomized design, and single variable experiment. Also, a variety of design classification schemes have been proposed. The classification scheme in Table 1.1 owes much to Cochran and Cox (1957, Chapters 4–13) and Federer (1955, pp. 11–12).

A quick perusal of Table 1.1 reveals why researchers sometimes have difficulty selecting an appropriate experimental design—there are many designs from which to choose. Because of the wide variety of designs that are available, it is important to identify them clearly in research reports. One often sees statements such as "a two-treatment factorial design was used." It should be evident that a more precise description is required. This description could refer to 10 of the 11 factorial designs in Table 1.1.

Thus far, I have limited my discussion to designs with one treatment and one or two nuisance variables. In the following sections, I describe designs with two or more treatments that are constructed by combining several building block designs.

FACTORIAL DESIGNS

Completely Randomized Factorial Design

Factorial designs differ from those described previously in that two or more treatments are evaluated simultaneously in an experiment. The simplest factorial design from the standpoint of randomization, data analysis, and model assumptions is based on a completely randomized design and, hence, is called a *completely randomized factorial design*. A two-treatment, completely randomized factorial design is denoted by the letters CRF-pq, where p and q denote the number of levels, respectively, of Treatments A and B.

In the weight-loss experiment, a researcher might be interested in knowing also whether walking on a treadmill for 30 minutes a day would contribute to losing weight, as well as whether the difference between the effects of walking or not walking on the treadmill would be the same for each of the three diets. To answer these additional questions, a researcher can use a two-treatment completely randomized factorial design. Let Treatment A consist of the three diets (a_1, a_2, and a_3) and Treatment B consist of no exercise on the treadmill (b_1) and exercise for 30 minutes a day on the treadmill (b_2). This design is a CRF-32 design, where 3 is the number of levels of Treatment A and 2 is the number of levels of Treatment B. The layout of the design is obtained by combining the treatment levels of a CR-3 design with those of a CR-2 design so that each treatment level of the CR-3 design appears once with each level of the CR-2 design and vice versa. The resulting design has $3 \times 2 = 6$ treatment combinations as follows: a_1b_1, a_1b_2, a_2b_1, a_2b_2, a_3b_1, a_3b_2. When treatment levels are combined in this way, the treatments are said to be

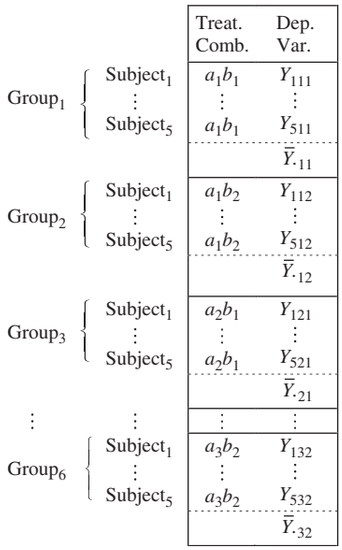

Figure 1.8 Layout for a two-treatment completely randomized factorial design (CRF-32 design). Thirty girls are randomly assigned to six combinations of Treatments A and B with the restriction that five girls are assigned to each combination. The mean weight loss in pounds for the girls in the six groups is denoted by $\overline{Y}_{\cdot 11}, \overline{Y}_{\cdot 12}, \ldots, \overline{Y}_{\cdot 32}$.

crossed. The use of crossed treatments is a characteristic of all factorial designs. The layout of the design with 30 girls randomly assigned to the six treatment combinations is shown in Figure 1.8.

The classical model equation for the two-treatment completely randomized factorial design is

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{i(jk)} \quad (i = 1, \ldots, n;\ j = 1, \ldots, p;\ k = 1, \ldots, q),$$

where

- Y_{ijk} is the weight loss for subject i in treatment combination $a_j b_k$.
- μ is the grand mean of the six weight-loss population means.
- α_j is the treatment effect for population a_j and is equal to $\mu_{j\cdot} - \mu$. α_j reflects the effect of diet a_j.
- β_k is the treatment effect for population b_k and is equal to $\mu_{\cdot k} - \mu$. β_k reflects the effects of exercise condition b_k.
- $(\alpha\beta)_{jk}$ is the interaction effect for populations a_j and b_k and is equal to $\mu_{jk} - \mu_{j\cdot} - \mu_{\cdot k} + \mu$. Interaction effects are discussed later.
- $\varepsilon_{i(jk)}$ is the within-cell error effect associated with Y_{ijk} and is equal to $Y_{ijk} - \mu - \alpha_j - \beta_k - (\alpha\beta)_{jk}$. $\varepsilon_{i(jk)}$ reflects all effects not attributable to treatment level a_j, treatment level b_k, and the interaction of a_j and b_k.

The CRF-32 design enables a researcher to test three null hypotheses:

H_0: $\mu_{1\cdot} = \mu_{2\cdot} = \mu_{3\cdot}$ (Treatment A population means are equal.)

H_0: $\mu_{\cdot 1} = \mu_{\cdot 2}$ (Treatment B population means are equal.)

H_0: $\mu_{jk} - \mu_{jk'} - \mu_{j'k} + \mu_{j'k'} = 0$ for all j and k (All $A \times B$ interaction effects equal zero.)

The last hypothesis is unique to factorial designs. If any difference in the dependent variable for one treatment is different at two or more levels of the other treatment, the treatments are said to interact.

Suppose that 30 girls are available to participate in the weight-loss experiment and have been randomly assigned to the six treatment combinations with the restriction that five girls are assigned to each combination. The data, weight loss for each girl, are given in Table 1.2. A descriptive summary of the data—sample means and standard deviations—is given in Table 1.3.

An examination of Table 1.3 suggests that diet a_3 resulted in more weight loss than did the other diets and that 30 minutes a day on the treadmill was beneficial. The analysis of variance for the weight-loss data is summarized in Table 1.4, which shows that the null hypotheses

TABLE 1.2 Weight-Loss Data for the Diet (a_j) and Exercise Conditions (b_k) (CRF-32 Design)

a_1b_1	a_1b_2	a_2b_1	a_2b_2	a_3b_1	a_3b_2
7	7	9	10	15	13
13	14	4	5	10	16
9	11	7	7	12	20
5	4	14	15	5	19
1	9	11	13	8	12

TABLE 1.3 Descriptive Summary of the Weight-Loss Data: Means ($\overline{Y}$) and Standard Deviations (S)

	Diet a_1	Diet a_2	Diet a_3	Mean Standard Deviation
No treadmill exercise (b_1)	$\overline{Y}_{\cdot 11} = 7.0$ $S_{\cdot 11} = 4.0$	$\overline{Y}_{\cdot 21} = 9.0$ $S_{\cdot 21} = 3.4$	$\overline{Y}_{\cdot 31} = 10.0$ $S_{\cdot 31} = 3.4$	$\overline{Y}_{\cdot\cdot 1} = 8.7$ $S_{\cdot\cdot 1} = 3.8$
Treadmill exercise (b_2)	$\overline{Y}_{\cdot 12} = 9.0$ $S_{\cdot 12} = 3.4$	$\overline{Y}_{\cdot 22} = 10.0$ $S_{\cdot 22} = 3.7$	$\overline{Y}_{\cdot 32} = 16.0$ $S_{\cdot 32} = 3.2$	$\overline{Y}_{\cdot\cdot 2} = 11.7$ $S_{\cdot\cdot 2} = 4.6$
Mean Standard deviation	$\overline{Y}_{\cdot 1\cdot} = 8.0$ $S_{\cdot 1\cdot} = 3.8$	$\overline{Y}_{\cdot 2\cdot} = 9.5$ $S_{\cdot 2\cdot} = 3.6$	$\overline{Y}_{\cdot 3\cdot} = 13.0$ $S_{\cdot 3\cdot} = 4.4$	

TABLE 1.4 Analysis of Variance Table for the Weight-Loss Data

Source	*SS*	*df*	*MS*	*F*	*p*
Treatment *A* (Diet)	131.6667	2	65.8334	4.25	.026
Treatment *B* (Exercise)	67.5000	1	67.5000	4.35	.048
$A \times B$	35.0000	2	17.5000	1.13	.340
Within cell	372.0000	24	15.5000		
Total	606.1667	29			

for Treatments *A* and *B* can be rejected. We know that at least one contrast or difference among the diet population means is not equal to zero.

Also, from Tables 1.3 and 1.4, we know that 30 minutes a day on the treadmill resulted in greater weight loss than did the no-exercise condition. The $A \times B$ interaction test is not significant. When two treatments interact, a graph in which treatment-combination population means are connected by straight lines will always reveal at least two nonparallel lines for one or more segments of the lines. The nonsignificant interaction test in Table 1.4 tells us that there is no reason for believing that the population difference in weight loss between the treadmill and no treadmill conditions is different for the three diets. If the interaction had been significant, our interest would have shifted from interpreting the tests of Treatments *A* and *B* to understanding the nature of the interaction. Procedures for interpreting interactions are described by Kirk (2012, pp. 373–386).

Statistical Significance Versus Practical Significance

The rejection of the null hypotheses for the diet and exercise treatments is not informative. I know in advance that the null hypotheses are false. John Tukey (1991, p. 100) observed that, "All we know about the world teaches us that the effects of A and B are always different—in some decimal place. Thus asking 'Are the effects different?' is foolish." Furthermore, rejection of a null hypothesis tells us nothing about the size of the treatment effects or whether the effects are important or large enough to be useful—that is, their practical significance. The logic and usefulness of null hypothesis significance testing has been debated for over 80 years. For a summary of the issues and major criticisms, see Kirk (2005, 2007).

In spite of the criticisms of null hypothesis significance testing, researchers continue to focus on null hypotheses and *p* values. I think that the focus should be on the data and what the data tell the researcher about the scientific hypothesis (Kirk, 1996, 2001). This is not a new idea. It was originally touched on by Karl Pearson in 1901 and more explicitly by Fisher in 1925. Fisher proposed that researchers supplement null hypothesis significance tests with measures of strength of association. Since then more than 70 supplementary measures of effect magnitude have been proposed (Kirk, 2007). The majority of the measures fall into one of two categories: measures of effect size (typically, standardized mean differences) and measures of strength of association. I describe Cohen's measure of effect size next.

In 1969 Cohen introduced the first effect-size measure that was explicitly labeled as such. His measure, denoted by δ, expresses the size of the absolute difference $\mu - \mu_0$ in units of the population standard deviation,

$$d = \frac{|\mu - \mu_0|}{\sigma}$$

where μ is the population mean and μ_0 is the hypothesized population mean. Cohen recognized that the size of the difference $\mu - \mu_0$ is influenced by the scale of measurement of the means. Cohen divided the difference between the means by the population standard deviation, σ, to rescale the difference in units of the amount of variability in the observations. What made Cohen's contribution unique is that he provided guidelines for interpreting the magnitude of δ.

$\delta = 0.2$ is a small effect
$\delta = 0.5$ is a medium effect
$\delta = 0.8$ is a large effect

According to Cohen (1992), a medium effect of 0.5 is visible to the naked eye of a careful observer. A small effect of 0.2 is noticeably smaller than medium but not so small as to be trivial. A large effect of 0.8 is the same distance above medium as small is below it. These operational definitions turned his measure of effect size into a much more useful statistic. A sample estimator of Cohen's *d* is obtained by replacing μ with $\overline{Y}$ and σ with $\hat{\sigma}$.

$$d = \frac{|\overline{Y} - \mu_0|}{\hat{\sigma}}$$

For experiments with two sample means, Larry Hedges (1981) proposed a modification of Cohen's *d* as follows:

$$g = \frac{|\overline{Y}_1 - \overline{Y}_2|}{\hat{\sigma}_{\text{Pooled}}}$$

where

$$\hat{\sigma}^2_{\text{Pooled}} = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{(n_1 - 1) + (n_2 - 1)} \quad d = \frac{|\overline{Y} - \mu_0|}{\hat{\sigma}}$$

is a pooled estimator of the unknown population standard deviation. Hedges's g is interpreted the same as Cohen's d.

Hays (1963) introduced a measure of strength of association that can assist a researcher in assessing the importance or usefulness of treatments in ANOVA: omega squared, $\hat{\omega}^2$. Omega squared estimates the proportion of the population variance in the dependent variable accounted for by a treatment. For experiments with several treatments, as in the weight-loss experiment, partial omega squared is computed. For example, the proportion of variance in the dependent variable, Y, accounted for by Treatment A eliminating Treatment B and the $A \times B$ interaction is denoted by $\hat{\omega}^2_{Y|A\cdot B,\,AB}$. Similarly, $\hat{\omega}^2_{Y|B\cdot A,\,AB}$ denotes the proportion of the variance accounted for by Treatment B eliminating Treatment A and the $A \times B$ interaction. For the weight-loss experiment, the partial omega squareds for Treatments A and B are, respectively,

$$\hat{\omega}^2_{Y|A\cdot B,\,AB} = \frac{(p-1)(F_A-1)}{(p-1)(F_A-1)+npq}$$

$$= \frac{(3-1)(4.247-1)}{(3-1)(4.247-1)+(5)(3)(2)} = 0.18$$

$$\hat{\omega}^2_{Y|B\cdot A,\,AB} = \frac{(q-1)(F_B-1)}{(q-1)(F_B-1)+npq}$$

$$= \frac{(2-1)(4.376-1)}{(2-1)(4.376-1)+(5)(3)(2)} = 0.10$$

Cohen (1988, pp. 284–288) proposed the following guidelines for interpreting measures of strength of association like omega squared:

.010 is a small association
.059 is a medium association
.138 is a large association

I conclude that the diets accounted for a large proportion of the population variance in weight loss. This is consistent with our perception of the differences between the weight-loss means for the three diets: Girls on diet a_3 lost five more pounds than did those on a_1. Certainly, any girl who is anxious to lose weight would want to be on diet a_3. Likewise, the medium association between the exercise conditions and weight loss is practically significant: Walking on the treadmill resulted in a mean weight loss of 3 pounds. Based on Tukey's HSD statistic, 95% confidence intervals for the three pairwise contrasts among the diet means are

$$-5.9 < \mu_{1\cdot} - \mu_{2\cdot} < 2.9$$

$$-9.4 < \mu_{1\cdot} - \mu_{3\cdot} < -0.6$$

$$-7.9 < \mu_{2\cdot} - \mu_{3\cdot} < 0.9$$

Because the confidence interval for $\mu_{1\cdot} - \mu_{3\cdot}$ does not contain 0, we can be confident that diet a_3 is superior to diet a_1.

Hedges's (1981) effect size for the difference between diets a_1 and a_3 is

$$g = \frac{|\overline{Y}_{\cdot 1\cdot} - \overline{Y}_{\cdot 3\cdot}|}{\hat{\sigma}_{\text{Pooled}}} = \frac{|8.0 - 13.0|}{3.937} = 1.27,$$

a large effect.

The determination of whether results are important or useful must be made by the researcher. Unfortunately, there is no statistic that measures practical significance. However, Cumming (2012) and Kirk (2007) illustrate how confidence intervals and measures of effect magnitude can help the researcher make this decision.

Alternative Models

Thus far, I have described the classical model equation for several experimental designs. This model and associated procedures for computing sums of squares assume that all cell ns in multitreatment experiments are equal. If the cell ns are not equal, some researchers use one of the following procedures to obtain approximate tests of null hypotheses: (a) estimate the missing observations under the assumption that the treatments do not interact, (b) randomly set aside data to reduce all cell ns to the same size, and (c) use an unweighted-means analysis. The latter approach consists of performing an ANOVA on the cell means and then multiplying the sums of squares by the harmonic mean of the cell ns. None of these procedures is entirely satisfactory. Fortunately, exact solutions to the unequal cell n problem exist. Two solutions that are described next are based on a regression model and a cell means model. Unlike the classical model approach, the regression and cell means model approaches require a computer and software for manipulating matrices.

Suppose that halfway through the weight-loss experiment the third subject in treatment combination a_2b_2 ($Y_{322} = 7$) moved to another area of the country and dropped out of the experiment. The loss of this subject resulted in unequal cell ns. Cell a_2b_2 has four subjects; the other cells have five subjects. The regression model for the weight-loss data with a missing observation is described next.

Regression Model

A qualitative regression model equation with $h - 1 = (p - 1) + (q - 1) + (p - 1)(q - 1) = 5$ independent variables $(X_{i1}, X_{i2}, \ldots, X_{i2}X_{i3})$ and $h = 6$ parameters $(\beta_0, \beta_1, \ldots, \beta_5)$,

$$Y_i = \beta_0 + \overbrace{\beta_1 X_{i1} + \beta_2 X_{i2}}^{A \text{ effects}} + \overbrace{\beta_3 X_{i3}}^{B \text{ effects}} + \overbrace{\beta_4 X_{i1} X_{i3} + \beta_5 X_{i2} X_{i3}}^{A \times B \text{ effects}} + e_i$$

can be formulated so that tests of selected parameters of the regression model provide tests of the null hypotheses for A, B, and $A \times B$ in the weight-loss experiment. Tests of the following null hypotheses for the regression model are of particular interest:

H_0: $\beta_1 = \beta_2 = 0$ (Corresponds to the Treatment A null hypothesis.)

H_0: $\beta_3 = 0$ (Corresponds to the Treatment B null hypothesis.)

H_0: $\beta_4 = \beta_5 = 0$ (Corresponds to the $A \times B$ interaction null hypothesis.)

For tests of these null hypotheses to provide tests of ANOVA null hypotheses, it is necessary to establish a correspondence between the five independent variables of the regression model equation and the $(p - 1) + (q - 1) + (p - 1)(q - 1) = 5$ treatment and interaction effects of the CRF-32 design. One way to establish this correspondence is to code the independent variables of the regression model as follows:

$$X_{i1} = \begin{cases} 1, & \text{if an observation is in } a_1 \\ -1, & \text{if an observation is in } a_3 \\ 0, & \text{otherwise} \end{cases}$$

$$X_{i2} = \begin{cases} 1, & \text{if an observation is in } a_2 \\ -1, & \text{if an observation is in } a_3 \\ 0, & \text{otherwise} \end{cases}$$

$$X_{i3} = \begin{cases} 1, & \text{if an observation is in } b_1 \\ -1, & \text{if an observation is in } b_2 \end{cases}$$

$$X_{i1}X_{i3} = \begin{cases} \text{product of coded values} \\ \text{associated with } a_1\ (\mathbf{x}_1) \text{ and } b_1\ (\mathbf{x}_3) \end{cases}$$

$$X_{i2}X_{i3} = \begin{cases} \text{product of coded values} \\ \text{associated with } a_2\ (\mathbf{x}_2) \text{ and } b_1\ (\mathbf{x}_3) \end{cases}$$

This coding scheme, which is called *effect coding*, produced the **X** matrix in Table 1.5. The **y** vector in Table 1.5 contains the 29 weight-loss observations. The first column vector, $\mathbf{x}_0$, in the **X** matrix contains ones; the second through the sixth column vectors contain coded values for X_{i1}, $X_{i2}, \ldots, X_{i2}X_{i3}$. To save space, only a portion of the 29 rows of **X** and **y** are shown. As mentioned earlier, observation Y_{322} is missing. Hence, each of the treatment combinations contains five observations except for a_2b_2, which contains four.

F statistics for testing hypotheses for selected regression parameters are obtained by dividing a regression mean square, MSR, by an error mean square, MSE, where $MSR = SSR/df_{\text{reg}}$ and $MSE = SSE/df_{\text{error}}$. The regression sum of squares, SSR, which reflects the contribution of

TABLE 1.5 Data Vector, y, and X Matrix for the Regression Model (Observation Y_{322} Is Missing)

	y 29 × 1	X 29 × 6 A A B $A \times B$ $A \times B$					
		$\mathbf{x}_0$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_1\mathbf{x}_3$	$\mathbf{x}_2\mathbf{x}_3$
a_1b_1	7	1	1	0	1	1	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	1	1	1	0	1	1	0
a_1b_2	7	1	1	0	−1	−1	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	9	1	1	0	−1	−1	0
a_2b_1	9	1	0	1	1	0	1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	1	1	0	1	1	0	1
	1						
a_2b_2	1						
	0	1	0	1	−1	0	−1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	1	1	0	1	−1	0	−1
	3						
a_3b_1	1						
	5	1	−1	−1	1	−1	−1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	8	1	−1	−1	1	−1	−1
a_3b_2	1						
	3	1	−1	−1	−1	1	1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	1	1	−1	−1	−1	1	1
	2						

independent variables X_1 and X_2 over and above the contribution of X_3, X_1X_3, and X_2X_3, is given by the difference between two error sums of squares, SSE, as follows:

$$SSR(\overbrace{X_1\ X_2}^{A}|\overbrace{X_3}^{B}\ \overbrace{X_1X_3\ X_2X_3}^{A\times B}) = SSE(\overbrace{X_3}^{B}\ \overbrace{X_1X_3\ X_2X_3}^{A\times B}) - SSE(\overbrace{X_1\ X_2}^{A}\ \overbrace{X_3}^{B}\ \overbrace{X_1X_3\ X_2X_3}^{A\times B})$$

For convenience, the designation $SSR(X_1\ X_2|X_3\ X_1X_3\ X_2X_3)$ can be shortened to $SSR(A|B, A \times B)$. An error sum of squares is given by

$$SSE(\ ) = \mathbf{y}'\mathbf{y}\ - [(\mathbf{X}_i'\mathbf{X}_i)^{-1}(\mathbf{X}_i'\mathbf{y})]'(\mathbf{X}_i'\mathbf{y})$$

where the $\mathbf{X}_i$ matrix contains the first column, $\mathbf{x}_0$, of $\mathbf{X}$ and the columns corresponding to the independent variables contained in $SSE(\ )$. For example, the $\mathbf{X}$ matrix that is used in computing $SSE(X_3X_1X_3X_2X_3)$ contains four columns: $\mathbf{x}_0$, $\mathbf{x}_3$, $\mathbf{x}_1\mathbf{x}_3$, and $\mathbf{x}_2\mathbf{x}_3$. For the data in Table 1.2 with $Y_{322} = 7$ missing, the regression sum of squares that corresponds to SSA in ANOVA is

$$SSR(\overbrace{X_1\ X_2}^{A}|\overbrace{X_3}^{B}\ \overbrace{X_1X_3\ X_2X_3}^{A\times B}) = SSE(\overbrace{X_3}^{B}\ \overbrace{X_1X_3\ X_2X_3}^{A\times B}) - SSE(\overbrace{X_1\ X_2}^{A}\ \overbrace{X_3}^{B}\ \overbrace{X_1X_3\ X_2X_3}^{A\times B})$$
$$= 488.1538 - 360.7500 = 127.4038$$

with $p - 1 = 2$ degrees of freedom. This sum of squares is used in testing the regression null hypothesis H_0: $\beta_1 = \beta_2 = 0$. Because of the correspondence between the regression and ANOVA parameters, a test of this regression null hypothesis is equivalent to testing the ANOVA null hypothesis for Treatment A.

The regression sum of squares that corresponds to SSB in ANOVA is

$$SSR(\overbrace{X_3}^{B}|\overbrace{X_1\ X_2}^{A}\ \overbrace{X_1X_3\ X_2X_3}^{A\times B}) = SSE(\overbrace{X_1\ X_2}^{A}\ \overbrace{X_1X_3\ X_2X_3}^{A\times B}) - SSE(\overbrace{X_1\ X_2}^{A}\ \overbrace{X_3}^{B}\ \overbrace{X_1X_3\ X_2X_3}^{A\times B})$$
$$= 436.8000 - 360.7500 = 76.0500$$

with $q - 1 = 1$ degree of freedom. A shorter designation for the regression sum of squares is $SSR(B|A, A \times B)$.

The regression sum of squares that corresponds to $SSA \times B$ in ANOVA is

$$SSR(\overbrace{X_1X_3\ X_2X_3}^{A\times B}|\overbrace{X_1\ X_2}^{A}\ \overbrace{X_3}^{B}) = SSE(\overbrace{X_1\ X_2}^{A}\ \overbrace{X_3}^{B}) - SSE(\overbrace{X_1\ X_2}^{A}\ \overbrace{X_3}^{B}\ \overbrace{X_1X_3\ X_2X_3}^{A\times B})$$
$$= 388.5385 - 360.7500 = 27.7885$$

with $(p - 1)(q - 1) = 2$ degrees of freedom. A shorter designation for the regression sum of squares is $SSR(A \times B|A, A \times B)$.

The regression error sum of squares that corresponds to $SSWCELL$ in ANOVA is

$$SSE(\overbrace{X_1\ X_2}^{A}\ \overbrace{X_3}^{B}\ \overbrace{X_1X_3\ X_2X_3}^{A\times B}) = 360.7500$$

with $N - h = 29 - 6 = 23$ degrees of freedom.

The total sum of squares is

$$SSTO = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{J}\mathbf{y}N^{-1} = 595.7931$$

where $\mathbf{J}$ is a 29×29 matrix of ones and $N = 29$, the number of weight-loss observations. The total sum of squares has $N - 1 = 28$ degrees of freedom. The analysis of the weight-loss data is summarized in Table 1.6. The null hypotheses $\beta_1 = \beta_2 = 0$ and $\beta_3 = 0$ can be rejected. Hence, independent variables X_1 or X_2 as well as X_3 contribute to predicting the dependent variable. As I show in the next section, the F statistics in Table 1.6 are identical to the ANOVA F statistics for the cell means model.

Cell Means Model

The classical model equation for a CRF-pq design,

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{i(jk)}$$
$$(i = 1, \ldots, n;\ j = 1, \ldots, p;\ k = 1, \ldots, q)$$

TABLE 1.6 Analysis of Variance Table for the Weight-Loss Data (Observation Y_{322} Is Missing)

Source	SS	df	MS	F	p
$SSR(A\|B, A \times B)$	127.4038	$p - 1 = 2$	63.7019	4.06	.031
$SSR(B\|A, A \times B)$	76.0500	$q - 1 = 1$	76.0500	4.85	.038
$SSR(A \times B\|A, B)$	27.7885	$(p - 1)(q - 1) = 2$	13.8943	0.89	.426
SSE	360.7500	$N - h = 23$	15.6848		
Total	595.7931	$N - 1 = 28$			

focuses on the grand mean, treatment effects, and interaction effects. The cell means model equation for the CRF-*pq* design,

$$Y_{ijk} = \mu_{jk} + \varepsilon_{i(jk)} \quad (i = 1, \ldots, n;\ j = 1, \ldots, p;\ k = 1, \ldots, q),$$

focuses on cell means, where μ_{jk} denotes the population mean in cell a_j and b_k. Although I described the classical model first, this is not the order in which the models evolved historically. According to Urquhart, Weeks, and Henderson (1973), Fisher's early development of ANOVA was conceptualized by his colleagues in terms of cell means. It was not until later that a cell mean was given a linear structure in terms of the grand mean and model effects, that is, $\mu_{jk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}$. The classical model equation for a CRF-*pq* design uses four parameters, $\mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}$, to represent one parameter, μ_{jk}. Because of this structure, the classical model is overparameterized. For example, the expectation of the classical model equation for the weight-loss experiment contains 12 parameters: μ, α_1, α_2, α_3, β_1, β_2, $(\alpha\beta)_{11}$, $(\alpha\beta)_{12}$, $(\alpha\beta)_{21}$, $(\alpha\beta)_{22}$, $(\alpha\beta)_{31}$, $(\alpha\beta)_{32}$. However, there are only 6 cell means with which to estimate the 12 parameters. When there are missing cells in multitreatment designs, a researcher is faced with the question of which parameters or parametric functions are estimable. For a discussion of this and other problems with the classical model, see Hocking (1985), Hocking and Speed (1975), Kirk and Bekele (2009), and Searle (1987).

The cell means model avoids the problems associated with overparameterization. A population mean can be estimated for each cell that contains one or more observations. Thus, the model is fully parameterized. Unlike the classical model, the cell means model does not impose a structure on the analysis of data. Consequently, the model can be used to test hypotheses about any linear combination of population cell means. It is up to the researcher to decide which tests are meaningful or useful based on the original research hypotheses, the way the experiment was conducted, and the data that are available.

I use the weight-loss data in Table 1.2 to illustrate the computational procedures for the cell means model approach. Again, I assume that observation Y_{322} is missing. The null hypothesis for Treatment *A* is

$$H_0\colon \mu_{1.} = \mu_{2.} = \mu_{3..}$$

An equivalent null hypothesis that is used with the cell means model is

$$H_0\colon \begin{aligned} \mu_{1.} - \mu_{2.} &= 0 \\ \mu_{2.} - \mu_{3.} &= 0. \end{aligned} \tag{1}$$

This hypothesis can be expressed in terms of cell means as

$$H_0\colon \begin{aligned} \frac{\mu_{11} + \mu_{12}}{2} - \frac{\mu_{21} + \mu_{22}}{2} &= 0 \\ \frac{\mu_{21} + \mu_{22}}{2} - \frac{\mu_{31} + \mu_{32}}{2} &= 0 \end{aligned} \tag{2}$$

where $\mu_{1.} = (\mu_{11} + \mu_{12})/2$, $\mu_{2.} = (\mu_{21} + \mu_{22})/2$, and so on. In matrix notation, the null hypothesis is

$$\begin{matrix} & \mathbf{C}_A' & & \boldsymbol{\mu} & \mathbf{0} \\ & (p-1) \times h & & h \times 1 & (p-1) \times 1 \end{matrix}$$

$$H_0\colon \frac{1}{2}\begin{bmatrix} 1 & 1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 & -1 \end{bmatrix}\begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \\ \mu_{31} \\ \mu_{32} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

where p is the number of levels of Treatment *A* and h is the number of cell means. For the null hypothesis $\mathbf{C}_A'\boldsymbol{\mu} = \mathbf{0}$ to be testable, the $\mathbf{C}_A'$ matrix must be of full row rank. This means that each row of $\mathbf{C}_A'$ must be linearly independent of every other row. The maximum number of such rows is $p - 1$, which is why it is necessary to express the null hypothesis as Equation 1 or 2. An estimator of the null hypothesis, $\mathbf{C}_A'\boldsymbol{\mu} - \mathbf{0}$, is incorporated in the formula for computing a sum of squares. For example, the estimator appears as $\mathbf{C}_A'\hat{\boldsymbol{\mu}} - \mathbf{0}$ in the formula for the Treatment *A* sum of squares

$$SSA = (\mathbf{C}_A'\hat{\boldsymbol{\mu}} - \mathbf{0})'[\mathbf{C}_A'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_A]^{-1}(\mathbf{C}_A'\hat{\boldsymbol{\mu}} - \mathbf{0}), \tag{3}$$

where $\hat{\boldsymbol{\mu}}$ is a vector of sample cell means. Equation 3 simplifies to

$$SSA = (\mathbf{C}_A'\hat{\boldsymbol{\mu}})'[\mathbf{C}_A'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_A]^{-1}(\mathbf{C}_A'\hat{\boldsymbol{\mu}})$$

because $\mathbf{0}$ is a vector of zeros. In the formula, $\mathbf{C}_A'$ is a coefficient matrix that defines the null hypothesis, $\hat{\boldsymbol{\mu}} = [(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})] = \left[\overline{Y}_{.11}\ \ \overline{Y}_{.12}\ \ \cdots\ \ \overline{Y}_{.23}\right]'$, and $\mathbf{X}$ is a structural matrix. The structural matrix for the weight-loss experiment is given in Table 1.7. The vectors, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_6$, of $\mathbf{X}$ are coded as follows:

$$\mathbf{x}_1 = \begin{cases} 1, & \text{if an observation is in } a_1b_1 \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbf{x}_2 = \begin{cases} 1, & \text{if an observation is in } a_1b_2 \\ 0, & \text{otherwise} \end{cases}$$

TABLE 1.7 Data Vector, y, and X Matrix for the Cell Means Model (Observation Y_{322} Is Missing)

	y 29 × 1	X 29 × 6					
		x_1	x_2	x_3	x_4	x_5	x_6
a_1b_1	7	1	0	0	0	0	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	1	1	0	0	0	0	0
a_1b_2	7	0	1	0	0	0	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	9	0	1	0	0	0	0
a_2b_1	9	0	0	1	0	0	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	1	0	0	1	0	0	0
	1						
a_2b_2	1						
	0	0	0	0	1	0	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	1	0	0	0	1	0	0
	3						
a_3b_1	1						
	5	0	0	0	0	1	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	8	0	0	0	0	1	0
a_3b_2	1						
	3	0	0	0	0	0	1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	1	0	0	0	0	0	1
	2						

$$\mathbf{x}_3 = \begin{cases} 1, & \text{if an observation is in } a_2b_1 \\ 0, & \text{otherwise} \end{cases}$$

$$\vdots$$

$$\mathbf{x}_6 = \begin{cases} 1, & \text{if an observation is in } a_3b_2 \\ 0, & \text{otherwise} \end{cases}$$

For the weight-loss data, the sum of squares for Treatment A is

$$SSA = (\mathbf{C}_A'\hat{\mu})'[\mathbf{C}_A'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_A]^{-1}(\mathbf{C}_A'\hat{\mu}) = 127.4038$$

with $p - 1 = 2$ degrees of freedom.

The null hypothesis for Treatment B is

$$H_0\colon \mu_{\cdot 1} = \mu_{\cdot 2}$$

In terms of cell means, this hypothesis is expressed as

$$H_0\colon \frac{\mu_{11} + \mu_{21} + \mu_{31}}{3} - \frac{\mu_{12} + \mu_{22} + \mu_{32}}{3} = 0$$

In matrix notation, the null hypothesis is

$$\underset{(q-1)\times h}{\mathbf{C}_B'} \quad \underset{h\times 1}{\mu} \quad = \quad \underset{(q-1)\times 1}{\mathbf{0}}$$

$$H_0 : \frac{1}{3}\begin{bmatrix} 1 & -1 & 1 & -1 & 1 & -1 \end{bmatrix}\begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \\ \mu_{31} \\ \mu_{32} \end{bmatrix} = [0]$$

where q is the number of levels of Treatment B and h is the number of cell means. The sum of squares for Treatment B is

$$SSB = (\mathbf{C}_B'\hat{\mu})'[\mathbf{C}_B'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_B]^{-1}(\mathbf{C}_B'\hat{\mu}) = 76.0500$$

with $q - 1 = 1$ degree of freedom.

The null hypothesis for the $A \times B$ interaction is

$$H_0\colon \mu_{jk} - \mu_{jk'} - \mu_{j'k} + \mu_{j'k'} = 0 \quad \text{for all } j \text{ and } k$$

This hypothesis can be expressed as

$$H_0\colon \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} = 0$$
$$\mu_{21} - \mu_{22} - \mu_{31} + \mu_{32} = 0$$

The two rows of the null hypothesis correspond to the two sets of means connected by crossed lines in Figure 1.9.

In matrix notation, the null hypothesis is

$$\underset{(p-1)(q-1)\times h}{\mathbf{C}_{A\times B}'} \quad \underset{h\times 1}{\mu} \quad = \quad \underset{(p-1)(q-1)\times 1}{\mathbf{0}}$$

$$H_0 : \begin{bmatrix} 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 1 \end{bmatrix}\begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \\ \mu_{31} \\ \mu_{32} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

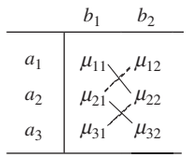

Figure 1.9 Two interaction terms of the form $\mu_{jk} - \mu_{jk'} - \mu_{j'k} + \mu_{j'k'}$ are obtained from the crossed lines by subtracting the two μ_{ij}s connected by a dashed line from the two μ_{ij}s connected by a solid line.

The sum of squares for the $A \times B$ interaction is

$$SSA \times B = (\mathbf{C}'_{A\times B}\hat{\mu})'[\mathbf{C}'_{A\times B}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_{A\times B}]^{-1}(\mathbf{C}'_{A\times B}\hat{\mu}) = 27.7885$$

with $(p-1)(q-1) = 2$ degrees of freedom.

The within-cell sum of squares is

$$SSWCELL = \mathbf{y}'\mathbf{y} - \hat{\mu}'(\mathbf{X}'\mathbf{y}) = 360.7500$$

where $\mathbf{y}'$ is the vector of weight-loss observations: [7 13 9 $\cdots$ 12]. The within-cell sum of squares has $N - h = 29 - 6 = 23$ degrees of freedom.

The total sum of squares is

$$SSTO = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{J}\mathbf{y}N^{-1} = 595.7931$$

where $\mathbf{J}$ is a 29×29 matrix of ones and $N = 29$, the number of weight-loss observations. The total sum of squares has $N - 1 = 28$ degrees of freedom.

The analysis of the weight-loss data is summarized in Table 1.8. The F statistics in Table 1.8 are identical to those in Table 1.6 where the regression model was used.

The cell means model approach is extremely versatile. It can be used when observations are missing and when entire cells are missing. It enables a researcher to test hypotheses about any linear combination of population cell means. It has an important advantage over the regression model: There is never any ambiguity about the hypothesis that is tested because an estimator of the null hypothesis, $\mathbf{C}'\mu - \mathbf{0}$, appears in the formula for a sum of squares. Lack of space prevents a discussion of the many other advantages of the model; the reader is referred to Kirk (2012, pp. 322–335, 411–431). However, before leaving the subject, the model will be used to test a null hypothesis for weighted means.

Occasionally, researchers collect data in which the sample sizes are proportional to the population sizes. This might occur, for example, in survey research. When cell ns are unequal, a researcher has a choice between computing unweighted means or weighted means. *Unweighted means* are simple averages of cell means. These are the means that were used in the previous analyses. *Weighted means* are weighted averages of cell means in which the weights are the sample cell sizes, n_{jk}. Consider again the weight-loss data in which observation Y_{322} is missing. Unweighted and weighted sample means for treatment level a_2 where observation Y_{322} is missing are, respectively,

$$\hat{\mu}_{2.} = \frac{\hat{\mu}_{21} + \hat{\mu}_{22}}{q} = \frac{9.00 + 10.75}{2} = 9.88$$

$$\hat{\bar{\mu}}_{2.} = \frac{n_{21}\hat{\mu}_{21} + n_{22}\hat{\mu}_{22}}{n_{j.}} = \frac{5(9.00) + 4(10.75)}{9} = 9.78$$

$n_{j.}$ is the number of observations in the jth level of Treatment A. The null hypothesis using weighted cell means for Treatment A is

$$H_0\colon \frac{n_{11}\mu_{11} + n_{12}\mu_{12}}{n_{1.}} - \frac{n_{21}\mu_{21} + n_{22}\mu_{22}}{n_{2.}} = 0$$

$$\frac{n_{21}\mu_{21} + n_{22}\mu_{22}}{n_{2.}} - \frac{n_{31}\mu_{31} + n_{32}\mu_{32}}{n_{3.}} = 0$$

The coefficient matrix for computing SSA is

$$\mathbf{C}'_A = \begin{bmatrix} \frac{5}{10} & \frac{5}{10} & -\frac{5}{10} & -\frac{4}{9} & 0 & 0 \\ 0 & 0 & \frac{5}{10} & \frac{4}{9} & -\frac{5}{10} & -\frac{5}{10} \end{bmatrix}$$

where the entries in $\mathbf{C}'_A$ are $\pm n_{jk}/n_{j.}$ and zero. The sum of squares and mean square for Treatment A are, respectively,

$$SSA = (\mathbf{C}'_A\hat{\mu})'[\mathbf{C}'_A(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_A]^{-1}(\mathbf{C}'_A\hat{\mu}) = 128.2375$$

$$MSA = SSA/(p-1) = 128.2375/(3-1) = 64.1188.$$

The F statistic and p value for Treatment A are

$$F = \frac{MSA}{MSWCELL} = \frac{64.1188}{15.6848} = 4.09 \quad p = .030$$

where $MSWCELL$ is obtained from Table 1.8. The null hypothesis is rejected. This is another example of the versatility of the cell means model. A researcher can test hypotheses about any linear combination of population cell means.

In most research situations, sample sizes are not proportional to population sizes. Unless a researcher has a compelling reason to weight the sample means proportional to the sample sizes, unweighted means *should be used*.

TABLE 1.8 Analysis of Variance Table for the Weight-Loss Data (Observation Y_{322} Is Missing)

Source	*SS*	*df*	*MS*	*F*	*p*
Treatment *A* (Diet)	127.4038	$p-1=2$	63.7019	4.06	.031
Treatment *B* (Exercise)	76.0500	$q-1=1$	76.0500	4.85	.038
$A \times B$	27.7885	$(p-1)(q-1)=2$	13.8943	0.89	.426
Within cell	360.7500	$N-h=23$	15.6848		
Total	595.7931	$N-1=28$			

RANDOMIZED BLOCK FACTORIAL DESIGN

Next I describe a factorial design that is constructed from two randomized block designs. The design is called a *randomized block factorial design* and is denoted by RBF-pq. The RBF-pq design is obtained by combining the levels of an RB-p design with those of an RB-q design so that each level of the RB-p design appears once with each level of the RB-q design and vice versa. The design uses the blocking procedure described in connection with an RB-p design to isolate variation attributable to a nuisance variable while simultaneously evaluating two or more treatments and associated interactions.

In discussing the weight-loss experiment, I hypothesized that ease of losing weight is related to the amount by which a girl is overweight. If the hypothesis is correct, a researcher can improve on the CRF-32 design by isolating this nuisance variable. Suppose that instead of randomly assigning 30 girls to the six treatment combinations in the diet experiment, the researcher formed blocks of six girls such that the girls in a block are overweight by about the same amount. One way to form the blocks is to rank the girls from the least to the most overweight. The six least overweight girls are assigned to Block 1. The next six girls are assigned to Block 2 and so on. In this example, five blocks of dependent samples can be formed from the 30 subjects. Once the girls have been assigned to the blocks, the girls in each block are then randomly assigned to the six treatment combinations. The layout for the experiment is shown in Figure 1.10.

The classical model equation for the experiment is

$$Y_{ijk} = \mu + \pi_i + \alpha_j + \beta_k + (\alpha\beta)_{jk} + (\alpha\beta\pi)_{jki}$$
$$(i = 1, \ldots, n;\ j = 1, \ldots, p;\ k = 1, \ldots, q),$$

where

Y_{ijk} is the weight loss for the subject in Block$_i$ and treatment combination a_jb_k.

μ is the grand mean of the six weight-loss population means.

π_i is the block effect for population i and is equal to $\mu_{i\cdot\cdot} - \mu_{\cdot}$. π_i reflects the effect of the nuisance variable in Block$_i$.

α_j is the treatment effect for population a_j and is equal to $\mu_{\cdot j\cdot} - \mu_{\cdot}$. α_j reflects the effect of diet a_j.

β_k is the treatment effect for population b_k and is equal to $\mu_{\cdot\cdot k} - \mu_{\cdot}$. β_k reflects the effects of exercise condition b_k.

$(\alpha\beta)_{jk}$ is the interaction effect for populations a_j and b_k and is equal to $\mu_{\cdot jk} - \mu_{\cdot j\cdot} - \mu_{\cdot\cdot k} + \mu_{\cdot}$.

$(\alpha\beta\pi)_{jki}$ is the residual error effect for treatment combination a_jb_k and Block$_i$.

The design enables a researcher to test four null hypotheses:

H_0: $\mu_{1\cdot\cdot} = \mu_{2\cdot\cdot} = \cdots = \mu_{5\cdot\cdot}$ (Block population means are equal.)

H_0: $\mu_{\cdot 1\cdot} = \mu_{\cdot 2\cdot} = \mu_{\cdot 3\cdot}$ (Treatment A population means are equal.)

H_0: $\mu_{\cdot\cdot 1} = \mu_{\cdot\cdot 2}$ (Treatment B population means are equal.)

H_0: $\mu_{\cdot jk} - \mu_{\cdot jk'} - \mu_{\cdot j'k} + \mu_{\cdot j'k'} = 0$ for all j and k (All $A \times B$ interaction effects equal zero.)

The hypothesis that the block population means are equal is of little interest because the blocks represent different amounts by which the girls are overweight, the nuisance variable.

The data for the RBF-32 design are shown in Table 1.9. The same data were analyzed earlier using a CRF-32 design. Each block in Table 1.9 contains six girls who at the beginning of the experiment were overweight by about

	Treat. Comb.	Dep. Var.	Treat. Comb.	Dep. Var.	⋯	Treat. Comb.	Dep. Var.	
Block$_1$	a_1b_1	Y_{111}	a_1b_2	Y_{112}	⋯	a_3b_2	Y_{132}	$\bar{Y}_{1\cdot\cdot}$
Block$_2$	a_1b_1	Y_{211}	a_1b_2	Y_{212}	⋯	a_3b_2	Y_{232}	$\bar{Y}_{2\cdot\cdot}$
Block$_3$	a_1b_1	Y_{311}	a_1b_2	Y_{312}	⋯	a_3b_2	Y_{332}	$\bar{Y}_{3\cdot\cdot}$
Block$_4$	a_1b_1	Y_{411}	a_1b_2	Y_{412}	⋯	a_3b_2	Y_{432}	$\bar{Y}_{4\cdot\cdot}$
Block$_5$	a_1b_1	Y_{511}	a_1b_2	Y_{512}	⋯	a_3b_2	Y_{532}	$\bar{Y}_{5\cdot\cdot}$
		$\bar{Y}_{\cdot 11}$		$\bar{Y}_{\cdot 12}$	⋯		$\bar{Y}_{\cdot 32}$	

Figure 1.10 Layout for a two-treatment randomized block factorial design (RBF-32 design). Each block contains six girls who are overweight by about the same amount. The girls in a block are randomly assigned to the six treatment combinations: a_1b_1, a_1b_2, a_2b_1, a_2b_2, a_3b_1, and a_3b_2.

TABLE 1.9 Weight-Loss Data for the Diet (a_j) and Exercise Conditions (b_k) (RBF-32 Design)

	a_1b_1	a_1b_2	a_2b_1	a_2b_2	a_3b_1	a_3b_2
Block_1	5	4	7	5	8	13
Block_2	7	7	4	7	5	16
Block_3	1	14	9	13	10	12
Block_4	9	9	11	15	12	20
Block_5	13	11	14	10	15	19

TABLE 1.10 Analysis of Variance Table for the Weight-Loss Data

Source	*SS*	*df*	*MS*	*F*	*p*
Blocks	209.3333	4	52.3333	6.43	.002
Treatments	234.1667	5			
Treatment *A* (Diet)	131.6667	2	65.8334	8.09	.003
Treatment *B* (Exercise)	67.5000	1	67.5000	8.30	.009
$A \times B$	35.0000	2	17.5000	2.15	.142
Residual	162.6667	20	8.1333		
Total	606.1667	29			

the same amount. The ANOVA for these data is given in Table 1.10. A comparison of Table 1.10 with Table 1.4 reveals that the RBF-32 design is more powerful than the CRF-32 design. Consider, for example, Treatment *A*. The F statistic for the randomized block factorial design is $F(2, 20) = 8.09,\ p = .003$; the F for the completely randomized factorial design is $F(2, 24) = 4.25,\ p = .026$. The randomized block factorial design is more powerful because the nuisance variable—amount by which the girls are overweight—has been removed from the residual error variance. A schematic partition of the total sum of squares and degrees of freedom for the two designs is shown in Figure 1.11. It is apparent from Figure 1.11 that the *SSRESIDUAL* will always be smaller than the *SSWCELL* if the *SSBLOCKS* is greater than zero. The larger the *SSBLOCKS* in a randomized block factorial design are, the greater the reduction in the *SSRESIDUAL*.

FACTORIAL DESIGNS WITH CONFOUNDING

In this section I describe three ANOVA designs that involve confounding. Confounding enables a researcher to reduce the size of blocks in a split-block factorial design and the number of treatment combinations in a fractional factorial design.

Split-Plot Factorial Design

As I have just shown, an important advantage of a randomized block factorial design relative to a completely randomized factorial design is greater power. However, if either p or q in a two-treatment randomized block factorial design is moderately large, the number of treatment combinations in each block can be prohibitively large. For example, an RBF-45 design has blocks of size $4 \times 5 = 20$. Obtaining blocks with 20 matched subjects or observing each subject 20 times is generally not feasible. In the late 1920s, Ronald A. Fisher and Frank Yates addressed the problem of prohibitively large block sizes by developing confounding schemes in which only a portion of the treatment combinations in an experiment are assigned to each block. Their work was extended in the 1940s by David J. Finney (1945, 1946) and Oscar Kempthorne (1947).

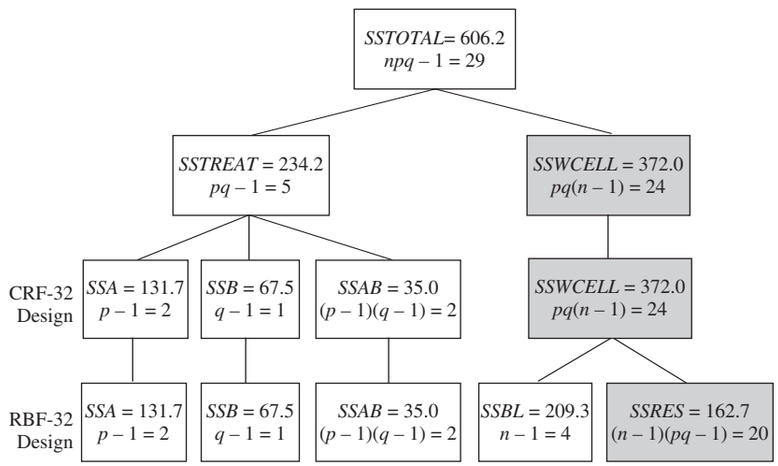

Figure 1.11 Schematic partition of the total sum of squares and degrees of freedom for CRF-32 and RBF-32 designs. The shaded rectangles indicate the sums of squares that are used to compute the error variance for each design: $MSWCELL = SSWCELL/pq(n-1)$ and $MSRES = SSRES/(n-1)(pq-1)$. If the nuisance variable (*SSBL*) in the RBF-32 design accounts for an appreciable portion of the total sum of squares, the design will have a smaller error variance and, hence, greater power than the CRF-32 design.

One design that achieves a reduction in block size is the two-treatment split-plot factorial design. The term *split-plot* comes from agricultural experimentation in which the levels of, say, Treatment A are applied to relatively large plots of land—the whole plots. The whole plots are then split or subdivided and the levels of Treatment B are applied to the subplots within each whole plot.

A two-treatment split-plot factorial design is constructed by combining two building block designs: a completely randomized design with p levels of Treatment A and a randomized block design with q levels of Treatment B. The assignment of subjects to the treatment combinations is carried out in two stages. Consider the weight-loss experiment again. Suppose that I ranked the 30 subjects from least to most overweight. The subjects ranked 1 and 2 are assigned to block 1, those ranked 3 and 4 are assigned to Block 2, and so on. This procedure produces 15 blocks each containing two girls who are similar with respect to being overweight. In the first stage of randomization, the 15 blocks of girls are randomly assigned to the three levels of Treatment A with five blocks in each level. In the second stage of randomization, the two girls in each block are randomly assigned to the two levels of Treatment B. An exception to this randomization procedure must be made when Treatment B is a temporal variable such as successive learning trials or periods of time. Trial 2, for example, cannot occur before Trial 1.

The layout for a split-plot factorial design with three levels of Treatment A and two levels of Treatment B is shown in Figure 1.12. Treatment A is called a *between-blocks treatment*; B is a *within-blocks treatment*. The

		b_1		b_2		
		Treat. Comb.	Dep. Var.	Treat. Comb.	Dep. Var.	
a_1 Group$_1$	Block$_1$	a_1b_1	Y_{111}	a_1b_2	Y_{112}	$\overline{Y}_{\cdot 1\cdot}$
	⋮	⋮	⋮	⋮	⋮	
	Block$_5$	a_1b_1	Y_{511}	a_1b_2	Y_{512}	
a_2 Group$_2$	Block$_1$	a_2b_1	Y_{121}	a_2b_2	Y_{122}	$\overline{Y}_{\cdot 2\cdot}$
	⋮	⋮	⋮	⋮	⋮	
	Block$_5$	a_2b_1	Y_{521}	a_2b_2	Y_{522}	
a_3 Group$_3$	Block$_1$	a_3b_1	Y_{131}	a_3b_2	Y_{132}	$\overline{Y}_{\cdot 3\cdot}$
	⋮	⋮	⋮	⋮	⋮	
	Block$_5$	a_3b_1	Y_{531}	a_3b_2	Y_{532}	
			$\overline{Y}_{\cdot\cdot 1}$		$\overline{Y}_{\cdot\cdot 2}$	

Figure 1.12 Layout for a two-treatment split-plot factorial design (SPF-3.2 design). The $np = 15$ blocks are randomly assigned to the $p = 3$ levels of Treatment A with the restriction that $n = 5$ blocks are assigned to each level of A. The 5 blocks assigned to each level of Treatment A constitute a group of blocks. In the second stage of randomization, the two matched subjects in a block are randomly assigned to the $q = 2$ levels of Treatment B.

designation for a two-treatment split-plot factorial design is SPF-$p{\cdot}q$. The p in front of the dot denotes the number of levels of the between-blocks treatment; the q after the dot denotes the number of levels of the within-blocks treatment. Hence, the design in Figure 1.12 is an SPF-3.2 design.

An RBF-32 design contains $3 \times 2 = 6$ treatment combinations and has blocks of size six. The SPF-3.2 design in Figure 1.12 contains the same six treatment combinations, but the block size is only two. The advantage of the split-plot factorial—smaller block size—is achieved by confounding groups of blocks with Treatment A. Consider the sample means $\overline{Y}_{\cdot 1\cdot}$, $\overline{Y}_{\cdot 2\cdot}$, and $\overline{Y}_{\cdot 3\cdot}$ in Figure 1.12. The differences among the means reflect the differences among the three groups of girls as well as the differences among the three levels of Treatment A. To put it another way, we cannot tell how much of the differences among the three sample means is attributable to the differences among Group$_1$, Group$_2$, and Group$_3$ and how much is attributable to the differences among treatments levels a_1, a_2, and a_3. For this reason, the three groups and Treatment A are said to be *completely confounded*.

The use of confounding to reduce the block size in an SPF-$p{\cdot}q$ design involves a trade-off that needs to be made explicit. The RBF-32 design uses the same error variance, *MSRESIDUAL*, to test hypotheses for Treatments A and B and the $A \times B$ interaction. The two-treatment split-plot factorial design, however, uses two error variances. *MSBLOCKS* within A, denoted by *MSBL*(A), is used to test Treatment A; a different and usually much smaller error variance, *MSRESIDUAL*, is used to test Treatment B and the $A \times B$ interaction. As a result, the power of the tests for B and the $A \times B$ interaction is greater than that for A. Hence, a split-plot factorial design is a good design choice if a researcher is more interested in Treatment B and the $A \times B$ interaction than in Treatment A. When both treatments and the $A \times B$ interaction are of equal interest, a randomized block factorial design is a better choice if the larger block size is acceptable. If a large block size is not acceptable and the researcher is primarily interested in Treatments A and B, an alternative design choice is the confounded factorial design. This design, which is described later, achieves a reduction in block size by confounding groups of blocks with the $A \times B$ interaction. As a result, tests of Treatments A and B are more powerful than the test of the $A \times B$ interaction.

Earlier, an RBF-32 design was used for the weight-loss experiment because the researcher was interested in tests of Treatments A and B and the $A \times B$ interaction. For purposes of comparison, I analyze the same weight-loss

data as if an SPF-3.2 design had been used even though, as we will see, this is not a good design choice. But first, I describe the classical model equation for a two-treatment split-plot factorial design.

The classical model equation for the weight-loss experiment is

$$Y_{ijk} = \mu + \alpha_j + \pi_{i(j)} + \beta_k + (\alpha\beta)_{jk} + (\beta\pi)_{ki(j)}$$
$$(i = 1, \ldots, n;\ j = 1, \ldots, p;\ k = 1, \ldots, q),$$

where

Y_{ijk} is the weight loss for the subject in $\text{Block}_{i(j)}$ and treatment combination $a_j b_k$.

μ is the grand mean of the six weight-loss population means.

α_j is the treatment effect for population a_j and is equal to $\mu_{.j.} - \mu$. α_j reflects the effect of diet a_j.

$\pi_{i(j)}$ is the block effect for population i and is equal to $\mu_{ij.} - \mu_{.j.}$. The ith block effect is nested within a_j.

β_k is the treatment effect for population b_k and is equal to $\mu_{..k} - \mu$. β_k reflects the effects of exercise condition b_k.

$(\alpha\beta)_{jk}$ is the interaction effect for populations aj and b_k and is equal to $\mu_{.jk} - \mu_{.j.} - \mu_{..k} + \mu$.

$(\beta\pi)_{ki(j)}$ is the residual error effect for treatment level b_k and $\text{Block}_{i(j)}$ and is equal to $Y_{ijk} - \mu - \alpha_j - \pi_{i(j)} - \beta_k - (\alpha\beta)_{jk}$

The design enables a researcher to test three null hypotheses:

H_0: $\mu_{.1.} = \mu_{.2.} = \mu_{.3.}$ (Treatment A population means are equal.)

H_0: $\mu_{..1} = \mu_{..2}$ (Treatment B population means are equal.)

H_0: $\mu_{.jk} - \mu_{.jk'} - \mu_{.j'k} + \mu_{.j'k'} = 0$ for all j and k

(All $A \times B$ interaction effects equal zero.)

The weight-loss data from Tables 1.2 and 1.9 are recast in the form of an SPF-3.2 design in Table 1.11. The ANOVA for these data is given in Table 1.12.

The null hypothesis for Treatment B can be rejected. However, the null hypothesis for Treatment A and the $A \times B$ interaction cannot be rejected. The denominator of the F statistic for Treatment A [$MSBL(A) = 20.1667$] is almost twice as large as the denominator for the tests of B and $A \times B$ ($MSRES = 10.8333$). A feeling for the relative power of the test of Treatment A for the SPF-3.2, CRF-32, and RBF-32 designs can be obtained by comparing their F statistics and p values.

TABLE 1.11 Weight-Loss Data for the Diet (a_j) and Exercise Conditions (b_k) (SPF-3.2 Design)

			Treatment Level b_1	Treatment Level b_2
Group_1	a_1	Block_1	5	4
		Block_2	7	7
		Block_3	1	14
		Block_4	9	9
		Block_5	13	11
Group_2	a_2	Block_1	7	5
		Block_2	4	7
		Block_3	9	13
		Block_4	11	15
		Block_5	14	10
Group_3	a_3	Block_1	8	13
		Block_2	5	16
		Block_3	10	12
		Block_4	12	20
		Block_5	15	19

TABLE 1.12 Analysis of Variance Table for the Weight-Loss Data

Source	*SS*	*df*	*MS*	*F*	*p*
1. Between blocks	373.6667	14			
2. Treatment *A* (Diet)	131.6667	2	65.8334	[2/3][a] 3.26	.074
3. Blocks within *A*	242.0000	12	20.1667		
4. Within blocks	232.5000	15			
5. Treatment *B* (Exercise)	67.5000	1	67.5000	[5/7] 6.23	.028
6. *A* × *B*	35.0000	2	17.5000	[6/7] 1.62	.239
7. Residual	130.0000	12	10.8333		
8. Total	606.1667	29			

[a] The fraction [2/3] indicates that the F statistic was obtained by dividing the mean square in row two by the mean square in row three.

Treatment A

SPF-3·2 design $F = \dfrac{131.6667/2}{242.0000/12} = \dfrac{65.8334}{20.1667}$

$= 3.26 \quad p = .074$

CRF-32 design $F = \dfrac{131.6667/2}{372.0000/24} = \dfrac{65.8334}{15.5000}$

$= 4.25 \quad p = .026$

RBF-32 design $F = \dfrac{131.6667/2}{162.6667/20} = \dfrac{65.8334}{8.1333}$

$= 8.09 \quad p = .003$

For testing Treatment A, the SPF-3.2 design is the least powerful. Clearly, if one's primary interest is in the effectiveness of the three diets, the SPF-3.2 design is a poor choice. However, the SPF-3.2 design fares somewhat better if one's primary interests are in Treatment B and the $A \times B$ interaction.

Treatment B

SPF-3·2 design $F = \dfrac{67.5000/1}{130.0000/12} = \dfrac{67.5000}{10.8333} = 6.23 \quad p = .028$

CRF-32 design $F = \dfrac{67.5000/1}{372.0000/24} = \dfrac{67.5000}{15.5000} = 4.35 \quad p = .048$

RBF-32 design $F = \dfrac{67.5000/1}{162.6667/20} = \dfrac{67.5000}{8.1333} = 8.30 \quad p = .009$

$A \times B$ interaction

SPF-3·2 design $F = \dfrac{35.0000/2}{130.0000/12} = \dfrac{17.5000}{10.8333} = 1.62 \quad p = .239$

CRF-32 design $F = \dfrac{35.0000/2}{372.0000/24} = \dfrac{17.5000}{15.5000} = 1.13 \quad p = .340$

RBF-32 design $F = \dfrac{35.0000/2}{162.6667/20} = \dfrac{17.5000}{8.1333} = 2.15 \quad p = .142$

The SPF-3.2 design is the first design I have described that involves two different building block designs: a CR-p design and an RB-q design. Also, it is the first design that has two error variances: one for testing the between-blocks effects and another for testing the within-blocks effects. A weighted average of the two error variances is equal to *MSWCELL* in a CRF-pq design, where the weights are the degrees of freedom of the two error variances. This can be shown using the mean squares from Tables 1.4 and 1.12:

$$\frac{p(n-1)MSBL(A) + p(n-1)(q-1)MSRESIDUAL}{p(n-1) + p(n-1)(q-1)} = MSWCELL$$

$$\frac{3(5-1)20.1667 + 3(5-1)(2-1)10.8333}{3(5-1) + 3(5-1)(2-1)} = 15.5000$$

A schematic partition of the total sum of squares and degrees of freedom for the CRF-32 and SPF-3.2 designs is shown in Figure 1.13.

Confounded Factorial Designs

As I have shown, an SPF-$p.q$ design is not the best design choice if a researcher's primary interest is in testing Treatments A and B. The RBF-pq design is a better choice if blocks of size $p \times q$ are acceptable. If this block size is too large, an alternative choice is a two-treatment confounded factorial design. This design confounds an

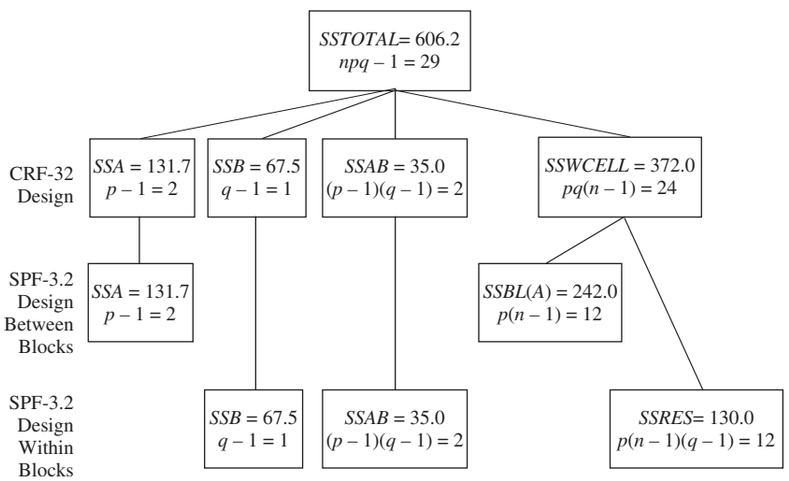

Figure 1.13 Schematic partition of the total sum of squares and degrees of freedom for CRF-32 and SPF-3.2 designs. The shaded rectangles indicate the sums of squares that are used to compute the error variance for each design. The SPF-3.2 design has two error variances: $MSBL(A) = SSBL(A)/p(n-1)$ is used to test Treatment A; $MSRES = SSRES/p(n-1)(q-1)$ is used to test Treatment B and the $A \times B$ interaction. The within-blocks error variance, *MSRES*, is usually much smaller than the between-blocks error variance, *MSBL*(A). As a result, tests of Treatment B and the $A \times B$ interaction are more powerful than the test of Treatment A.

interaction with groups of blocks. As a result, the test of the interaction is less powerful than the tests of Treatments A and B. Confounded factorial designs are constructed from either a randomized block design or a Latin square design. The designs are denoted by, respectively, RBCF-p^k and LSCF-p^k, where RB and LS identify the building block design, C indicates that the interaction is completely confounded with groups of blocks, F indicates a factorial design, and p^k indicates that the design has k treatments each having p levels. The simplest randomized block confounded factorial design has two treatments with two levels each. Consider the RBCF-22 design in Figure 1.14.

The $A \times B$ interaction is completely confounded with Group_1 and Group_2 as I now show. An interaction effect for Treatments A and B has the general form $\mu_{jk} - \mu_{jk'} - \mu_{j'k} + \mu_{j'k'}$. Let μ_{ijkz} denote the population mean for the ith block, jth level of A, kth level of B, and zth group. For the design in Figure 1.14, the $A \times B$ interaction effect is

$$\mu_{\cdot 111} - \mu_{\cdot 122} - \mu_{\cdot 212} + \mu_{\cdot 221}$$

$$\text{or} \quad (\mu_{\cdot 111} + \mu_{\cdot 221}) - (\mu_{\cdot 122} + \mu_{\cdot 212})$$

The difference between the effects of Group_1 and Group_2,

$$(\mu_{\cdot 111} + \mu_{\cdot 221}) - (\mu_{\cdot 122} + \mu_{\cdot 212}),$$

involves the same contrast among means as the $A \times B$ interaction effect. Hence, the two sets of effects are completely confounded because we cannot determine how much of the difference $(\mu_{\cdot 111} + \mu_{\cdot 221}) - (\mu_{\cdot 122} + \mu_{\cdot 212})$ is attributable to the $A \times B$ interaction and how much is attributable to the difference between Group_1 and Group_2.

The RBCF-p^k design, like the SPF-$p.q$ design, has two error variances: one for testing the between-blocks effects and a different and usually much smaller error variance for testing the within-blocks effects. In the RBCF-p^k design, Treatments A and B are within-block treatments and are evaluated with greater power than the $A \times B$ interaction, which is a between-block component. Researchers need to understand the trade-off that is required when a treatment or interaction is confounded with groups to reduce the size of blocks. The power of the test of the confounded effects is generally less than the power of tests of the unconfounded effects. Hence, if possible, researchers should avoid confounding effects that are the major focus of an experiment. Sometimes, however, confounding is necessary to obtain a reasonable block size. If the power of the confounded effects is not acceptable, the power can be increased by using a larger number of blocks.

One of the characteristics of the designs that I have described so far is that all of the treatment combinations appear in the experiment. The fractional factorial design that is described next does not share this characteristic. As the name suggests, a fractional factorial design includes only a fraction of the treatment combinations of a complete factorial design.

Fractional Factorial Designs

Two kinds of confounding have been described thus far: group–treatment confounding in an SPF-$p.q$ design and group–interaction confounding in an RBCF-p^k design. A third form of confounding, treatment–interaction confounding, is used in a fractional factorial design. This kind of confounding reduces the number of treatment combinations that must be included in a multitreatment experiment to some fraction—$\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{9}$, and so on—of the total number of treatment combinations. A CRF-22222 design has 32 treatment combinations. By using a $\frac{1}{2}$ or $\frac{1}{4}$ fractional factorial design, the number of treatment combinations that must be included in the experiment can be reduced to, respectively, $\frac{1}{2}(32) = 16$ or $\frac{1}{4}(32) = 8$.

The theory of fractional factorial designs was developed for 2^k and 3^k designs by Finney (1945, 1946) and extended by Kempthorne (1947) to designs of the type p^k, where p is a prime number that denotes the number of levels of each treatment and k denotes the number

		a_jb_k Treat. Comb.	Dep. Var.	a_jb_k Treat. Comb.	Dep. Var.	
$(AB)_{jk}$ Group$_1$	Block_1	a_1b_1	Y_{1111}	a_2b_2	Y_{1221}	$\overline{Y}_{\cdot 111} + \overline{Y}_{\cdot 221}$
	Block_{n_1}	a_1b_1	Y_{n_1111}	a_2b_2	Y_{n_1221}	
$(AB)_{jk}$ Group$_2$	Block_1	a_1b_2	Y_{1122}	a_2b_1	Y_{1212}	$\overline{Y}_{\cdot 122} + \overline{Y}_{\cdot 212}$
	Block_{n_2}	a_1b_2	Y_{n_2122}	a_2b_1	Y_{n_2212}	
			$\overline{Y}_{\cdot 111} + \overline{Y}_{\cdot 122}$		$\overline{Y}_{\cdot 221} + \overline{Y}_{\cdot 212}$	

Figure 1.14 Layout for a two-treatment randomized block confounded factorial design (RBCF-2^2 design). A score in the ith block, jth level of Treatment A, kth level of Treatment B, and zth group is denoted by Y_{ijkz}.

of treatments. Fractional factorial designs are most useful for pilot experiments and exploratory research situations that permit follow-up experiments to be performed. Thus, a large number of treatments, typically six or more, can be investigated efficiently in an initial experiment, with subsequent experiments designed to focus on the most promising independent variables.

Fractional factorial designs have much in common with confounded factorial designs. The latter designs achieve a reduction in the number of treatment combinations that must be included in a block. Fractional factorial designs achieve a reduction in the number of treatment combinations in the experiment. The reduction in the size of an experiment comes at a price, however. Considerable ambiguity may exist in interpreting the results of an experiment when the design includes only one half or one third of the treatment combinations. Ambiguity occurs because two or more names can be given to each sum of squares. For example, a sum of squares might be attributed to the effects of Treatment *A* and the *BCDE* interaction. The two or more names given to the same sum of squares are called *aliases*. In a one-half fractional factorial design, all sums of squares have two aliases. In a one-third fractional factorial design, all sums of squares have three aliases, and so on. Treatments are customarily aliased with higher-order interactions that are assumed to equal zero. This helps to minimize but does not eliminate ambiguity in interpreting the outcome of an experiment.

Fractional factorial designs are constructed from completely randomized, randomized block, and Latin square designs and denoted by, respectively, CRFF-p^{k-1}, RBFF-p^{k-1}, and LSFF-p^k. Let's examine the designation CRFF-2^{5-1}. The letters CR indicate that the building block design is a completely randomized design, FF indicates that it is a fractional factorial design, and 2^5 indicates that each of the five treatments has two levels. The -1 in 2^{5-1} indicates that the design is a one-half fraction of a complete 2^5 factorial design. This follows because the designation for a one-half fraction of a 2^5 factorial design can be written as $\frac{1}{2}2^5 = 2^{-1}2^5 = 2^{5-1}$. A one-fourth fraction of a 25 factorial design is denoted by CRFF-p^{5-2} because $\frac{1}{4}2^5 = \frac{1}{2^2}2^5 = 2^{-2}2^5 = 2^{5-2}$.

To conserve space, I describe a small CRFF-2^{3-1} design. A fractional factorial design with only three treatments is unrealistic, but the small size simplifies the presentation. The layout for the design is shown in Figure 1.15. On close inspection of Figure 1.15, it is apparent that the CRFF-2^{3-1} design contains the four treatment combinations of a CRF-22 design. For example, if I ignore Treatment *C*, the design in Figure 1.15 has the following

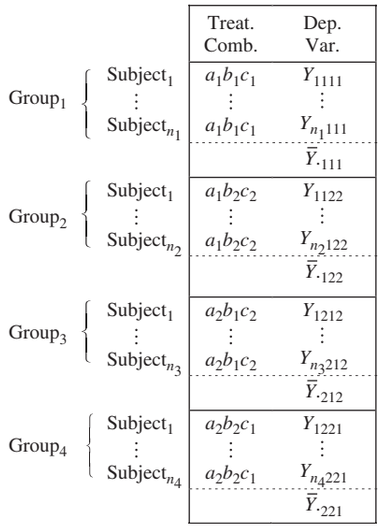

Figure 1.15 Layout for a three-treatment completely randomized fractional factorial design (CRFF-2^{3-1} design). A score for the ith subject in treatment combination $a_jb_kc_l$ is denoted by Y_{ijkl}. The $4n$ subjects are randomly assigned to the treatment combinations with the restriction that n subjects are assigned to each combination. The mean for the subjects in the four groups is denoted by $\overline{Y}_{\cdot 111}$, $\overline{Y}_{\cdot 122}$, $\overline{Y}_{\cdot 212}$, $\overline{Y}_{\cdot 221}$.

combinations of Treatments *A* and *B*: a_1b_1, a_1b_2, a_2b_1, and a_2b_2. The correspondence between the treatment combinations of the CRF-22 and CRFF-2^{3-1} designs suggests a way to compute sums of squares for the latter design—ignore Treatment *C* and analyze the data as if they came from a CRF-22 design.

Earlier, I observed that all sums of squares in a one-half fractional factorial design have two aliases. It can be shown (see Kirk, 2006; Kirk, 2012, pp. 805–809) that the alias pattern for the design in Figure 1.15 is as follows:

Alias (Name)	Alias (Alternative name)
A	$B \times C$
B	$A \times C$
$A \times B$	C

The labels—Treatment *A* and the $B \times C$ interaction—are two names for the same source of variation. Similarly, *B* and the $A \times C$ interaction are two names for another source of variation, as are $A \times B$ and *C*. Hence, the *F* statistics

$$F = \frac{MSA}{MSWCELL} \quad \text{and} \quad F = \frac{MSB \times C}{MSWCELL}$$

test the same sources of variation. If $F = MSA/MSWCELL$ is significant, a researcher doesn't know whether it is because Treatment *A* is significant, or the $B \times C$ interaction is significant, or both.

At this point you are probably wondering why anyone would use such a design—after all, experiments are supposed to help researchers resolve ambiguity, not create it. In defense of fractional factorial designs, recall that they are typically used in exploratory research situations where a researcher is interested in six or more treatments. In addition, it is customary to limit all treatments to either two or three levels thereby increasing the likelihood that higher-order interactions are small relative to treatments and lower-order interactions. Under these conditions, if a source of variation labeled Treatment A and its alias, the $BCDEF$ interaction, is significant, it is reasonable to assume that the significance is probably due to the treatment rather than the interaction.

Continuing the defense, a fractional factorial design can dramatically decrease the number of treatment combinations that must be run in an experiment. Consider a researcher who is interested in determining whether any of six treatments that have two levels each is significant. An experiment with six treatments and two subjects assigned to each treatment combination would have 64 combinations and require $2 \times 64 = 128$ subjects. By using a one-fourth fractional factorial design, CRFF-2^{6-2} design, the researcher can reduce the number of treatment combinations in the experiment from 64 to 16 and the number of subjects from 128 to 32. Suppose that the researcher ran the 16 treatment combinations and found that none of the F statistics in the fractional factorial design is significant. The researcher has answered the research questions with one fourth of the effort. On the other hand, suppose that F statistics for treatments C and E and associated aliases are significant. The researcher has eliminated four treatments, A, B, D, F, their aliases, and certain other interactions from further consideration. The researcher can then follow-up with a small experiment to determine which aliases are responsible for the significant F statistics.

In summary, the main advantage of a fractional factorial design is that it enables a researcher to investigate efficiently a large number of treatments in an initial experiment, with subsequent experiments designed to focus on the most promising lines of investigation or to clarify the interpretation of the original analysis. Many researchers would consider ambiguity in interpreting the outcome of the initial experiment a small price to pay for the reduction in experimental effort.

The description of confounding in a fractional factorial design completes a cycle. I began the cycle by describing group–treatment confounding in a split-plot factorial design. I then described group–interaction confounding in a confounded factorial design, and, finally, treatment–interaction confounding in a fractional factorial design. The three forms of confounding achieve either a reduction in the size of a block or the size of an experiment. As I have shown, confounding always involves a tradeoff. The price we pay for reducing the size of a block or an experiment is lower power in testing a treatment or interaction or ambiguity in interpreting the outcome of an experiment. In the next section, I describe hierarchical designs in which one or more treatments are nested.

HIERARCHICAL DESIGNS

All of the multitreatment designs that I have discussed so far have had crossed treatments. Treatments A and B are crossed, for example, if each level of Treatment B appears once with each level of Treatment A and vice versa. Treatment B is *nested* in Treatment A if each level of Treatment B appears with only one level of Treatment A. The nesting of Treatment B within Treatment A is denoted by $B(A)$ and is read, "B within A." A *hierarchical design* has at least one nested treatment; the remaining treatments are either nested or crossed.

Hierarchical Designs With One or Two Nested Treatments

Hierarchical designs are constructed from completely randomized or randomized block designs. A two-treatment hierarchical design that is constructed from CR-p and CR-q designs is denoted by CRH-$pq(A)$, where $pq(A)$ indicates that the design has p levels of Treatment A and q levels of Treatment B that are nested in Treatment A. A comparison of nested and crossed treatments for a CRH-24(A) design and a CRF 22 design is shown in Figure 1.16.

Experiments with one or more nested treatments are well suited to research in education, industry, and the

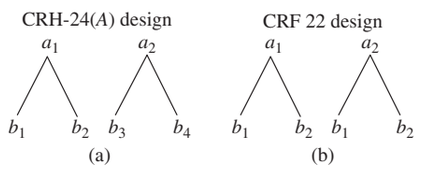

Figure 1.16 Comparison of designs with nested and crossed treatments. In (a), Treatment $B(A)$ is nested in Treatment A because b_1 and b_2 appear only with a_1 while b_3 and b_4 appear only with a_2. In (b), Treatments A and B are crossed because each level of Treatment B appears once and only once with each level of Treatment A and vice versa.

behavioral and medical sciences. Consider an example from education in which two approaches to introducing long division (treatment levels a_1 and a_2) are to be evaluated. Four schools (treatment levels $b_1, \ldots, b_4$) are randomly assigned to the two levels of Treatment A, and eight teachers (treatment levels $c_1, \ldots, c_8$) are randomly assigned to the four schools. Hence, this is a three-treatment CRH-24(A)8(AB) design: schools, Treatment $B(A)$, are nested in Treatment A and teachers, Treatment $C(AB)$, are nested in both A and B. A diagram of the nesting of treatments for this design is shown in Figure 1.17.

A second example is from medical science. A researcher wants to compare the efficacy of a new drug denoted by a_1 with the currently used drug denoted by a_2. Four hospitals, Treatment $B(A)$, are available to participate in the experiment. Because expensive equipment is needed to monitor the side effects of the new drug, it was decided to use the new drug in two of the four hospitals and the current drug in the other two hospitals. The hospitals are randomly assigned to the drug conditions with the restriction that two hospitals are assigned to each drug. Patients are randomly assigned to the hospitals. Panel A of Figure 1.16 illustrates the nesting of Treatment B within Treatment A.

As is often the case, the nested treatments in the drug and education examples resemble nuisance variables. The researcher in the drug example probably would not conduct the experiment just to find out whether the dependent variable is different for the two hospitals assigned to drug a_1 or the hospitals assigned to a_2. The important question for the researcher is whether the new drug is more effective than the currently used drug. Similarly, the educational researcher wants to know whether one approach to teaching long division is better than the other. The researcher might be interested in knowing whether some schools or teachers perform better than others, but this is not the primary focus of the research. The distinction between a treatment and a nuisance variable is in the mind of the researcher—one researcher's nuisance variable can be another researcher's treatment.

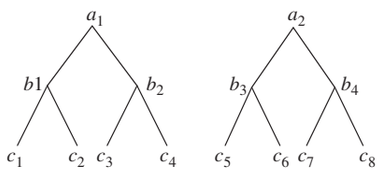

Figure 1.17 Diagram of a three-treatment completely randomized hierarchical design (CRH-24(A)8(AB) design). The four schools, $b_1, \ldots, b_4$, are nested in the two approaches to introducing long division, Treatment A. The eight teachers, $c_1, \ldots, c_8$, are nested in the schools and teaching approaches. The levels of Treatment $B(A)$ are randomly assigned to Treatment A. The levels of Treatment $C(AB)$ are randomly assigned to Treatment $B(A)$. Students are randomly assigned to the $pq_{(j)}r_{(jk)} = (2)(2)(2) = 8$ treatment combinations with the restriction that n students are assigned to each combination.

Consider the experiment to evaluate the efficacy of the two drugs. The classical model equation for the two-treatment hierarchical design is

$$Y_{ijk} = \mu + \alpha_j + \beta_{k(j)} + \varepsilon_{i(jk)} \qquad (i = 1, \ldots, n;\ j = 1, \ldots, p;\ k = 1, \ldots, q)$$

where

Y_{ijk} is the observation for subject i in treatment levels a_j and $b_{k(j)}$.
μ is the grand mean of the population means.
α_j is the treatment effect for population a_j and is equal to $\mu_{j.} - \mu_{..}$ α_j reflects the effect of drug a_j.
$\beta_{k(j)}$ is the treatment effect for population $b_{k(j)}$ and is equal to $\mu_{jk} - \mu_{j.}$. $\beta_{k(j)}$ reflects the effects of hospital $b_{k(j)}$ that is nested in a_j.
$\varepsilon_{i(jk)}$ is the within-cell error effect associated with Y_{ijk} and is equal to $Y_{ijk} - \mu - \alpha_j - \beta_{k(j)}$. $\varepsilon_{i(jk)}$ reflects all effects not attributable to treatment levels a_j and $b_{k(j)}$.

Notice that because Treatment $B(A)$ is nested in Treatment A, the model equation does not contain an $A \times B$ interaction term.

The design enables a researcher to test two null hypotheses:

H_0: $\mu_{1.} = \mu_{2.}$ (Treatment A population means are equal.)

H_0: $\mu_{11} = \mu_{12}$ or $\mu_{23} = \mu_{24}$ (Treatment $B(A)$ population means are equal.)

If the second null hypothesis is rejected, the researcher can conclude that the dependent variable is not the same for the populations represented by hospitals b_1 and b_2, that the dependent variable is not the same for the populations represented by hospitals b_3 and b_4, or both. However, the test of Treatment $B(A)$ does not address the question of whether, for example, $\mu_{11} = \mu_{23}$ because hospitals b_1 and b_3 were assigned to different levels of Treatment A.

Hierarchical Design With Crossed and Nested Treatments

In the educational example, Treatments $B(A)$ and $C(AB)$ were both nested treatments. Hierarchical designs with three or more treatments can have both nested and crossed treatments. Consider the partial hierarchical design shown in Figure 1.18.

The classical model equation for this design is

$$Y_{ijkl} = \mu + \alpha_j + \beta_k + \gamma_{l(k)} + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl(k)} + \varepsilon_{i(jkl)}$$
$$(i = 1, \ldots, n;\ j = 1, \ldots, p;\ k = 1, \ldots, q;\ l = 1, \ldots, r)$$

where

Y_{ijkl} is an observation for subject i in treatment levels a_j, b_k, and $c_{l(k)}$.

μ is the grand mean of the population means.

α_j is the treatment effect for population a_j and is equal to $\mu_{j..} - \mu$.

β_k is the treatment effect for population b_k and is equal to $\mu_{.k.} - \mu$.

$\gamma_{l(k)}$ is the treatment effect for population $c_{l(k)}$ and is equal to $\mu_{.kl} - \mu_{.k.}$.

$(\alpha\beta)_{jk}$ is the interaction effect for populations a_j and b_k and is equal to $\mu_{jk.} - \mu_{jk'.} - \mu_{j'k.} + \mu_{j'k'.}$.

$(\alpha\gamma)_{jl(k)}$ is the interaction effect for populations a_j and $c_{l(k)}$ and is equal to $\mu_{jkl} - \mu_{jkl'.} - \mu_{j'kl.} + \mu_{j'kl'.}$.

$\varepsilon_{i(jkl)}$ is the within-cell error effect associated with Y_{ijkl} and is equal to $Y_{ijkl} - \mu - \alpha_j - \beta_k - \gamma_{l(k)} - (\alpha\beta)_{jk} - (\alpha\gamma)_{jl(k)}$.

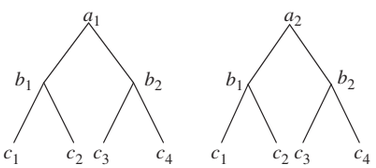

Figure 1.18 Diagram of a three-treatment completely randomized partial hierarchical design (CRPH-*pqr*(B) design). The letter P in the designation stands for "partial" and indicates that not all of the treatments are nested. In this example, Treatments A and B are crossed; Treatment $C(B)$ is nested in Treatment B because c_1 and c_2 appear only with b_1 while c_3 and c_4 appear only with b_2. Treatment $C(B)$ is crossed with Treatment A because each level of Treatment $C(B)$ appears once and only once with each level of Treatment A and vice versa.

Notice that because Treatment $C(B)$ is nested in treatment B, the model equation does not contain $B \times C$ and $A \times B \times C$ interaction terms.

This design enables a researcher to test five null hypotheses:

H_0: $\mu_{1..} = \mu_{2..}$ (Treatment A population means are equal.)

H_0: $\mu_{.1.} = \mu_{.2.}$ (Treatment B population means are equal.)

H_0: $\mu_{.11} = \mu_{.12}$ or $\mu_{.23} = \mu_{.24}$
(Treatment $C(B)$ population means are equal.)

H_0: $\mu_{jk.} - \mu_{jk'.} - \mu_{j'k.} + \mu_{j'k'.} = 0$ for all j and k
(All $A \times B$ interaction effects equal zero.)

H_0: $\mu_{jkl} - \mu_{jkl'} - \mu_{j'kl} + \mu_{j'kl'} = 0$ for all $j, k,$ and l
(All $A \times C(B)$ interaction effects equal zero.)

If the last null hypothesis is rejected, the researcher knows that Treatments A and C interact at one or more levels of Treatment B.

Lack of space prevents me from describing other partial hierarchical designs with different combinations of crossed and nested treatments. The interested reader is referred to the extensive treatment of these designs in Kirk (2012, Chapter 11).

EXPERIMENTAL DESIGNS WITH A COVARIATE

The emphasis so far has been on designs that use *experimental control* to reduce error variance and minimize the effects of nuisance variables. Experimental control can take various forms such as random assignment of subjects to treatment levels, stratification of subjects into homogeneous blocks, and refinement of techniques for measuring a dependent variable. In this section, I describe an alternative approach to reducing error variance and minimizing the effects of nuisance variables. The approach is called *analysis of covariance* (ANCOVA) and combines regression analysis and analysis of variance.

Analysis of covariance involves measuring one or more *concomitant variables* (also called *covariates*) in addition to the dependent variable. The concomitant variable represents a source of variation that was not controlled in the experiment and one that is believed to affect the dependent variable. Analysis of covariance enables a researcher to (a) remove that portion of the dependent-variable error

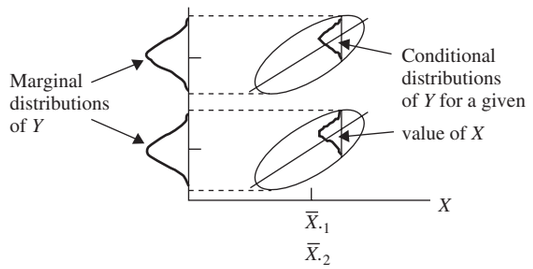

Figure 1.19 Scatterplots showing the relationship between the dependent variable, Y, and the concomitant variable, X, for the two treatment levels. The size of the error variance in ANOVA is determined by the dispersion of the marginal distributions. The size of the error variance in ANCOVA is determined by the dispersion of the conditional distributions. The higher the correlation between X and Y, the greater is the reduction in the error variance due to using analysis of covariance.

variance that is predictable from knowledge of the concomitant variable thereby increasing power and (b) adjust the dependent variable so that it is free of the linear effects attributable to the concomitant variable thereby reducing bias.

Consider an experiment with two treatment levels a_1 and a_2. The dependent variable is denoted by Y_{ij} and the concomitant variable by X_{ij}. The relationship between X and Y for a_1 and a_2 might look like that shown in Figure 1.19.

Each subject in the experiment contributes one data point to the figure as determined by his or her X_{ij} and Y_{ij} scores. The points form two scatterplots—one for each treatment level. These scatterplots are represented in Figure 1.19 by ellipses. Through each ellipsis a line has been drawn representing the regression of Y on X. In the typical ANCOVA model it is assumed that each regression line is a straight line and that the lines have the same slope. The size of the error variance in ANOVA is determined by the dispersion of the marginal distributions (see Figure 1.19). The size of the error variance in ANCOVA is determined by the dispersion of the *conditional distributions* (see Figure 1.19). The higher the correlation between X and Y, in general, the narrower are the ellipses and the greater is the reduction in the error variance due to using analysis of covariance.

Figure 1.19 depicts the case in which the concomitant-variable means, $\overline{X}_{\cdot 1}$ and $\overline{X}_{\cdot 2}$, are equal. If subjects are randomly assigned to treatment levels, in the long run the concomitant-variable means should be equal. However, if random assignment is not used, differences among the means can be sizable, as in Figure 1.20.

This figure illustrates what happens to the dependent variable means when they are adjusted for differences in the concomitant-variable means. In Panels A and B the absolute difference between adjusted dependent-variable means $\left|\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2}\right|$ is smaller than that between unadjusted means $\left|\overline{Y}_{\text{adj}\cdot 1} - \overline{Y}_{\text{adj}\cdot 2}\right|$. In Panel C the absolute difference between adjusted means is larger than that between unadjusted means.

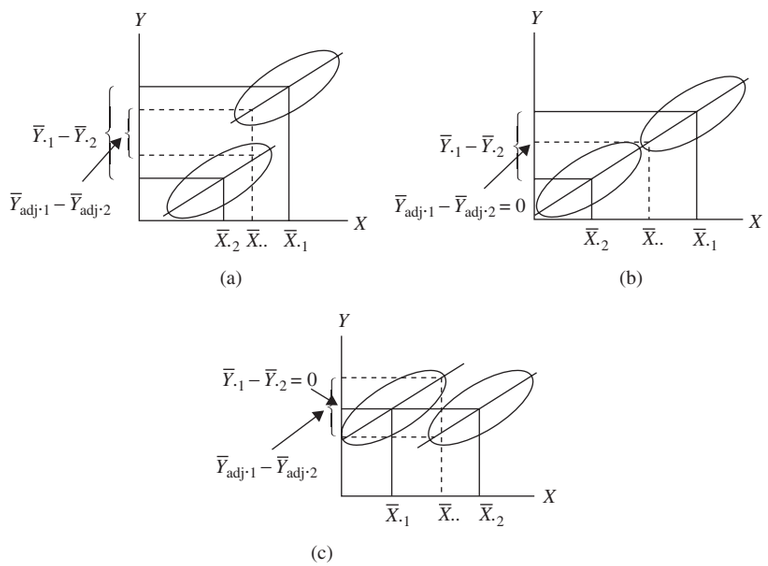

Figure 1.20 Analysis of covariance adjusts the concomitant-variable means, $\overline{X}_{\cdot 1}$ and $\overline{X}_{\cdot 2}$, so that they equal the concomitant-variable grand mean, $\overline{X}$. When the concomitant-variable means differ, the absolute difference between adjusted means for the dependent variable, $|\overline{Y}_{\text{adj}\cdot 1} - \overline{Y}_{\text{adj}\cdot 2}|$, can be less than that between unadjusted means, $|\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2}|$, as in Panels A and B, or larger, as in Panel C.

Analysis of covariance is often used in three kinds of research situations. One situation involves the use of intact groups with unequal concomitant-variable means and is common in educational and industrial research. Analysis of covariance statistically equates the intact groups so that their concomitant variable means are equal. Unfortunately, a researcher can never be sure that the concomitant variable used for the adjustment represents the only nuisance variable or the most important nuisance variable on which the intact groups differ. Random assignment is the best safeguard against unanticipated nuisance variables. In the long run, over many replications of an experiment, random assignment will result in groups that are, at the time of assignment, similar on all nuisance variables.

A second situation in which analysis of covariance is often used is when it becomes apparent that even though random assignment was used, the subjects were not equivalent on some relevant variable at the beginning of the experiment. For example, in an experiment designed to evaluate the effects of different drugs on stimulus generalization in rats, the researcher might discover that the amount of stimulus generalization is related to the number of trials required to establish a stable bar-pressing response. Analysis of covariance can be used to adjust the generalization scores for differences among the groups in learning ability.

Analysis of covariance is useful in yet another research situation in which differences in a relevant nuisance variable occur during the course of an experiment. Consider the experiment to evaluate two approaches toward introducing long division that was described earlier. It is likely that the daily schedules of the eight classrooms provided more study periods for students in some classes than in others. It would be difficult to control experimentally the amount of time available for studying long division. However, students could record the amount of time they spent studying long division. If test scores on long division were related to amount of study time, analysis of covariance could be used to adjust the scores for differences in this nuisance variable.

Statistical control and experimental control are not mutually exclusive approaches for reducing error variance and minimizing the effects of nuisance variables. It may be convenient to control some variables by experimental control and others by statistical control. In general, experimental control involves fewer assumptions than does statistical control. However, experimental control requires more information about the subjects before beginning an experiment. Once data collection has begun, it is too late to assign subjects randomly to treatment levels or form blocks of dependent subjects. The advantage of statistical control is that it can be used after data collection has begun. Its disadvantage is that it involves a number of assumptions such as a linear relationship between the dependent and concomitant variables and equal within-groups regression coefficients that may prove untenable in a particular experiment.

In this chapter, I have given a short introduction to those experimental designs that are potentially the most useful in the behavioral and social sciences. For a full discussion of the designs, the reader is referred to the many excellent books on experimental design: Kirk (2012), Maxwell and Delaney (2004), Montgomery (2009), and Wu and Hamada (2009). Experimental designs differ in a number of ways: (a) randomization procedures, (b) number of treatments, (c) use of independent samples or dependent samples with blocking, (d) use of crossed and nested treatments, (e) presence of confounding, and (6) use of covariates. Researchers have many design decisions to make. I have tried to make the researcher's task easier by emphasizing two related themes throughout the chapter. First, complex designs are constructed from three simple building block designs. Second, complex designs share similar layouts, randomization procedures, and assumptions with their building block designs.

REFERENCES

Box, J. F. (1978). *R. A. Fisher: The life of a scientist*. New York, NY: Wiley.

Cochran, W. G., & Cox, G. M. (1957). *Experimental designs* (2nd ed.). New York, NY: Wiley.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *12*, 115–159.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Federer, W. T. (1955). *Experimental design: Theory and application*. New York, NY: Macmillan.

Finney, D. J. (1945). The fractional replication of factorial arrangements. *Annals of Eugenics*, *12*, 291–301.

Finney, D. J. (1946). Recent developments in the design of field experiments. III. Fractional replication. *Journal of Agricultural Science*, *36*, 184–191.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, *52*, 399–433.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver & Boyd.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture*, *33*, 503–513.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, UK: Oliver & Boyd.

Fisher, R. A., & Mackenzie, W. A. (1923). Studies in crop variation. II. The manurial response of different potato varieties. *Journal of Agricultural Science*, *13*, 311–320.

Hays, W. L. (1963). *Statistics*. Fort Worth, TX: Holt, Rinehart and Winston.

Hedges, L. V. (1981). Distributional theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128.

Hocking, R. R. (1985). *The analysis of linear models*. Pacific Grove, CA: Brooks/Cole.

Hocking, R. R., & Speed, F. M. (1975). A full rank analysis of some linear model problems. *Journal of the American Statistical Association*, *70*, 706–712.

Kempthorne, O. (1947). A simple approach to confounding and fractional replication in factorial experiments. *Biometrika*, *34*, 255–272.

Kirk, R. E., (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746–759.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, *61*, 213–218.

Kirk, R. E. (2005). Effect size measures. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 2, pp. 532–542). Hoboken, NJ: Wiley.

Kirk, R. E. (2006). Fractional randomized block design. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 375–379). Thousand Oaks, CA: Sage.

Kirk, R. E. (2007). Effect magnitude: A different focus. *Journal of Statistical Planning and Inference*, *137*, 1634–1646.

Kirk, R. E. (2012). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Thousand Oaks, CA: Sage.

Kirk, R. E., & Bekele, B. N. (2009). Cell means analyses for ANCOVA designs. *Journal of Statistical Theory and Practice*, *3* (4), 891–906.

Leonard, W. H., & Clark, A. G. (1939). *Field plot techniques*. Minneapolis, MN: Burgess.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

Montgomery, D. C. (2009). *Design and analysis of experiments* (7th ed.). Hoboken, NJ: Wiley.

Pearson, K. (1901). On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London*, *195*, 1–47.

Searle, S. R. (1987). *Linear models for unbalanced data*. New York, NY: Wiley.

Snedecor, G. W. (1934). *Analysis of variance and covariance*. Ames, IA: Collegiate Press.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, *6*, 100–116.

Urquhart, N. S., Weeks, D. L., & Henderson, C. R. (1973). Estimation associated with linear models, a revisitation. *Communications in Statistics*, *1*, 303–330.

Wu, C. F. J., & Hamada, Michael S. (2009). *Experiments: Planning, analysis, and optimization* (2nd ed.). Hoboken, NJ: Wiley.