Keppel, G. (1991) Design and analysis: a researcher's handbook. Upper Saddle River, NJ: Prentice Hall.

treatments that produce large estimates of statistical association. On the other hand, if we based *all* our actions on the size of $\hat{\omega}_A^2$ alone, we would be making a mistake. This is because a small statistical association may often be theoretically important.

4.2 CONTROLLING TYPE I AND TYPE II ERRORS

As discussed in the last chapter, we are able to control the magnitude of type I error (false rejection of the null hypothesis) through our choice of a rejection region for the F distribution (the α level). The control of the type II error (failing to reject the null hypothesis when it is false) and of power, unfortunately, is not this simple because power depends on several factors, including the size of the treatment effects, sample size, degree of error variance, and significance level.

You have already seen the reciprocity between the two types of error in Fig. 3–3 (p. 55)—that is, you know that any change in the size of the type I error will produce a change of opposite direction in the type II error. Given this reciprocity, then, one obvious way to decrease type II error (and to increase power) would be to increase the probability of a type I error. Unfortunately, however, we seem to be stuck with a rigidly set α level since rarely will it be set at any value greater than .05.

Why has type I error become fixed over the years? One answer is the absence of agreement concerning the relative seriousness of the two types of errors; in the absence of agreement, we must be arbitrary. It may be possible to establish a rational balancing of type I and type II errors in certain applied fields, however. In the medical sciences, for instance, researchers might be able to place a numerical value on the consequences of failing to recognize a new wonder drug (a type II error) and contrast this with the value placed on the consequences of switching to a new drug that is not better than the original one (a type I error). But in most research areas of the behavioral sciences, we are without such guidance. How serious is it if a new hypothesis is not recognized or if an old one is incorrectly rejected? Without explicit values to guide us, we must proceed by conventions. Thus, we fix the type I error at a level that will be acceptable to most researchers—that is, $\alpha = .05$ or lower—and allow the type II error (and power) to be what it has to be. We cannot answer the question of what is an acceptable level of power since we are usually not in a position to give weight to the relative consequences of the two types of error.

Why should we be concerned with controlling power? The answer is that power reflects the degree to which we can *detect* the treatment differences we expect and the chances that others will be able to *duplicate* our findings when they attempt to repeat our experiments. These are compelling reasons for us to pay strict attention to controlling power in our experiments. In spite of these arguments, however, the reality is that most researchers appear to pay little attention to power and that most experiments in the behavioral sciences are surprisingly lacking in power. To illustrate, Cohen (1962) surveyed all the research published in Volume 61 (1960) of the *Journal of Abnormal and Social Psychology* and found the studies to be substantially deficient in power. More specifically, if we assume that the overall effect size of studies reported in this journal was of "medium"

strength (that is, $\hat{\omega}_A^2 = .06$), the average power calculated by Cohen was **.48**. You should note carefully exactly what this finding means: The significant effects reported in Volume 61 of this journal would have on average a 50–50 chance of being detected by others trying to duplicate these findings—a pretty dismal prognosis. A subsequent analysis 10 years later by Brewer (1972), who reviewed studies in a number of research journals, echoed Cohen's conclusion. Even more recently, an analysis by Sedlmeier and Gigerenzer (1989), who duplicated Cohen's examination of the same journal 24 years later, yielded almost exactly the same conclusion—the average power for detecting medium effects was **.50**.

The unfortunate conclusion from these findings is that research in the behavioral sciences is woefully lacking in power. This statement implies that a substantial number of research projects have been undertaken and then discarded when they failed to produce results at the accepted significance level of $\alpha = .05$. If power is truly equal to .50, as the evidence suggests, half of the research undertaken will not yield significant results even though there are real differences among the treatment means that should have been detected. This finding also means that the research outcomes that are published are unlikely to be duplicated by others who may attempt to repeat or to replicate these studies.

The puzzling aspect of these estimates of the power of published findings is that procedures are readily available to assist researchers in designing studies with respectable power, thus permitting them to avoid perpetuating this somewhat discouraging and dismal state of affairs. Let me state the problem once more. Why should we waste time and resources undertaking a project that has a relatively low probability of detecting treatment effects and producing significant results? We should be designing experiments that stand an excellent chance of detecting differences—power of .80 or higher—rather than repeat the actions taken by researchers in the past. Why haven't experimenters learned from the analyses of Cohen and others? Sedlmeier and Gigerenzer (1989) explore this puzzling question in some detail. I suspect that the answer lies in the training that researchers in the behavioral sciences have received. Most books on statistics and methodology tend to place more emphasis on significance testing and design issues than on considerations of statistical power. As Kraemer and Thiemann (1987) put it, "although, in principle, deriving power is as straightforward as deriving a significance level, researchers are routinely trained to deal with significance level, but rarely with power" (p. 16). Cohen (1988) is more optimistic, however, as he comments on the problem in the Preface to the second edition of *Statistical Power Analysis for the Behavioral Sciences* (see pp. xiii-xiv). He reports admittedly hearsay evidence that funding agencies are beginning to require power analyses to be included as an integral part of research proposals submitted to them. If true, I suspect that researchers will quickly correct any deficiencies they may have had in their statistical educations concerning power.

Another reason for conducting a power analysis is to avoid wasting resources by performing experiments with *too much* power. Not only does an experiment with an excessively large sample size cost more in time and money than one with a more appropriate sample size, but also the conclusions drawn are often misleading. Since the size of F depends in part on sample size, a particularly large sample

size will produce a large F, which frequently is thought to reflect a "large" effect (see p. 64). This potential misinterpretation is one of the reasons why measures of relative treatment magnitude were developed and are reported in research articles. In addition, contemporary concerns for the rights of animals used as subjects in experiments usually involve the recommendation that animal studies be designed with the smallest sample sizes commensurate with acceptable power to keep any possible suffering and loss of life to a minimum.

In the remainder of this chapter, I will emphasize ways to facilitate the determination of power during the planning stage of an experiment. There is really no excuse for omitting this critical step in the design of an experiment. The stakes are simply too high for researchers to continue ignoring power when they design and interpret experiments.

4.3 REDUCING ERROR VARIANCE

It is of interest to see how the sensitivity of an experiment is related to the size of the error component. With treatment effects of a given magnitude, any increase in the size of the error variance reduces the size of the F ratio and lessens our chances of rejecting the null hypothesis; any decrease in error increases these chances.

There are three major sources of error variance: random variation in the actual treatments, unanalyzed control factors, and individual differences ("permanent" or "temporary" factors affecting a subject's performance during the course of the experiment). All these sources are reflected in a subject's score on the dependent variable and thus contribute to error variance, although certain steps can be taken to reduce their magnitude.

Reducing Treatment Variability

I have noted previously that no experimental treatment is exactly alike for every subject in a particular condition. The calibration of the equipment may change from session to session; the experimenter will not be perfectly consistent in the conduct of the experiment; and environmental factors such as noise level, illumination, and temperature will not be identical for each subject. To the extent that these factors influence the behavior under study, their variation from subject to subject contributes to the estimate of experimental error. I should add to this list any error of measurement and of recording that appear randomly in the data collection.

We can take certain steps to minimize these sources of variability: carefully calibrated equipment, automation, well-trained experimenters, and special testing rooms. In essence, this solution attempts to hold constant the specific conditions of testing in the experiment.

Unanalyzed Control Factors

Control factors are nuisance variables that are introduced into an experiment for a variety of reasons, but primarily for the removal of possible bias and for an increase in the generality of the results. For example, suppose an experimenter plans an experiment that requires far too many subjects for one assistant to test. If more than one assistant is employed, the researcher should make sure that each runs an equal number of subjects in each of the treatment conditions. To do otherwise would introduce a potential confounding of assistants and treatments into the experiment.

Most researchers would disregard these control factors—different laboratory assistants in the example—in the analysis of their experiments and simply analyze the results as a completely randomized single-factor design. Often this is a mistake because any variability associated with control factors contributes directly to error variance. That is, the variability of subjects within any group will now include differences associated with laboratory assistants, say, in addition to the usual factors contributing to experimental error. In short, a nonrandom source of variance, such as assistants, that is spread equally over all treatment conditions, as in this example, does not bias the treatment effects, but it does contribute to the size of the error term. The obvious solution to such a situation is to include control factors in the statistical analysis, using procedures that I will consider in later chapters of this book.

Reducing Subject Variability

Undoubtedly the major source of error variance in the behavioral sciences is that contributed by individual differences. The fact that subjects differ widely in performance on laboratory tasks means that when they are assigned randomly to the treatment conditions, this variability becomes an important source of error variance. The most obvious way of reducing subject variability is to select subjects who are relatively similar on some important and relevant characteristic, for example, IQ in a learning test, visual acuity in a perception experiment, socioeconomic status in an attitude-change study, and so on. A second type of matching is accomplished in small sets that consist of subjects matched *within* a set while generally differing widely *between* sets. Neither procedure is widely used in psychology, however, perhaps because there are more effective methods of reducing subject variability. These preferred procedures include the use of the same subject in all the treatment conditions and a statistical technique called the **analysis of covariance**, which adjusts estimates of error variance and of treatment effects on the basis of information obtained before the start of the study. Both of these procedures will be considered in subsequent chapters.

4.4 USING SAMPLE SIZE TO CONTROL POWER

The power of an experiment is determined by the interplay of three factors, namely, significance level α, the magnitude or size of the treatment effects, and sample size n. We will consider the influence of each factor on power in a moment. From a practical point of view, however, only one of these—sample size—is normally used to control power. This is because the α level is effectively fixed at

$p = .05$ by most researchers in the behavioral sciences and the effect size is frequently assumed to be as large as possible, given the specific interests of the researcher and the conditions surrounding his or her experimental design.

The relationship between sample size and power is presented in Table 4–1 for an experiment with $a = 4$ treatment conditions.[4] The entries in the upper half of the table are based on $\alpha = .05$, and those in the lower half are based on $\alpha = .01$. The three rows within each half of the table give the sample sizes n needed to achieve varying degrees of power for three different assumed or expected effect sizes, $\omega^2 = .01$ ("small"), .06 ("medium"), and .15 ("large"). We will examine this table carefully as it reveals some sobering facts about the sample sizes needed to obtain respectable amounts of power.

First, let's consider the sample sizes needed when the significance level is set at $p = .05$. Look carefully at the numbers appearing in the first row—that is, when the expected effect size is "small." As you can see, sample sizes are outrageously large if we want to obtain reasonable amounts of power. For example, if power is set at .80, we must assign 271 subjects to each of the four treatment conditions (a total of $(a)(n) = (4)(271) = 1{,}084$ subjects); if power is set at .90, we need 354 subjects (a total of 1,416). Consider next the sample sizes required when the expected effect size is "medium." The sample size we need for a power of .80 is 44 subjects (a total of 176 subjects) and for a power of .90 is 57 subjects (a total of 228 subjects). Obviously the situation is improved in the sense that there is a substantial drop in the sample sizes when the expected effect size increases from $\omega^2 = .01$ to .06, but still, these numbers will probably seem alarmingly large to most seasoned researchers. Finally, the sample sizes we need when we expect to find a "large" effect are more like those that we typically find in psychology research journals, namely, $n = 17$ for a power of .80 and $n = 22$ for a power of .90.

What effect will a more stringent significance level ($\alpha = .01$) have on the sample sizes, assuming that we want to maintain power at the same level we achieved at $\alpha = .05$? If you remember the relationship between significance level

Table 4–1 Sample Size (n) as a Function of Power, Effect Size (ω^2), and Significance Level (α)

EFFECT SIZE (ω^2)	*POWER*								
	.10	.20	.30	.40	.50	.60	.70	.80	.90
					$\alpha = .05$				
.01	21	53	83	113	144	179	219	271	354
.06	5	10	14	19	24	30	36	44	57
.15	3	5	6	8	10	12	14	17	22
					$\alpha = .01$				
.01	70	116	156	194	232	274	323	385	478
.06	13	20	26	32	38	45	53	62	77
.15	6	8	11	13	15	18	20	24	29

[4]These values were calculated by a computer program called PC-SIZE, described by Dallal (1986).

and power diagrammed in Fig. 3–3 (p. 55), you will realize that reducing the size of the rejection region guarantees that we will reject fewer null hypotheses when H_0 is true—we reduce type I error— but it also guarantees that we will reject fewer null hypotheses when H_0 is false (and H_1 is true), thus increasing type II error (and decreasing power). As a consequence, we will need to increase sample size to maintain power. You can see the effect of decreasing the rejection region from .05 to .01 simply by comparing the entries in the upper half of Table 4–1 with those in the lower half. To achieve power of .80 in an experiment with an expected effect size of $\omega^2 = .06$, for example, we need to increase sample size from $n = 44$ (for $\alpha = .05$) to $n = 62$ (for $\alpha = .01$).

What you have just observed has been called the power-sample size "facts of life" (Kraemer, 1985):

1. Increasingly larger sample sizes are needed to increase power a fixed amount.
2. Relatively small expected effect sizes require substantial sample sizes to achieve a reasonable power.
3. Adopting a more stringent significance level leads to a hefty increase in sample size to maintain power at the same level with a less stringent criterion.

Consider again how we reached this point in our discussion. Low power is poor science—we waste time, energy, and resources whenever we conduct an experiment that has a low probability of producing a significant result. What is the point of initiating an experiment that has low power? I have already indicated that we can directly translate the estimated power of an experiment into a probability statement that we will successfully reject the null hypothesis when it is false. When we design an experiment with an estimated power of .50, we stand a 50–50 chance of obtaining a significant F. Would you bet any money in a game of chance with these odds? Science must be based on solid research findings, findings that others can depend on and duplicate if they were to repeat the experiments. Experiments with low power do not produce reliable findings.

Estimating Effect Size

To estimate power or to achieve a certain power by selecting an appropriate sample size, we need to specify the sort of experimental outcome we wish to detect. This statement is usually expressed as a ratio that relates the variation of the anticipated population treatment means to an estimate of the variation within these populations. In many cases, researchers try to make realistic guesses of the expected outcomes by conducting preliminary or pilot studies in which a few selected treatment conditions are compared or from similar research published by others. One of my colleagues, for example, was able to determine the sample size he would need for an extensive series of related experiments on the basis of several preliminary studies in which two of the key conditions were compared. This, then, is the ideal situation—an educated guess about the specific patterns of differences that theory and a knowledge of the field suggest should occur. You should note that we do not have to estimate the absolute values of the population means, only the expected

differences among them. Finally, if we estimate the population treatment effects from the means of an experiment, we should adjust these estimates for the unavoidable presence of random error. We can accomplish this most easily by calculating estimated omega squared ($\hat{\omega}_A^2$) from the sample data and substituting this value in one of the formulas I will soon present to begin the process of estimating power.

On the other hand, if we are unable to specify the exact pattern of differences, we might be able to specify simply the *range* of the population treatment means—the difference between the largest and smallest population mean—as well as a general pattern that the population means might take. Cohen (1977, pp. 276–280), for example, offers the following three patterns that a researcher might specify:

1. *Minimum variability*: One mean is at each extreme and the others at the midpoint between them.
2. *Intermediate variability*: The means are spaced equally over the range.
3. *Maximum variability*: Half the means fall at each extreme.

Cohen then shows how we can combine this information with an estimate of population variability to produce a useful measure of effect size.[5]

Finally, researchers might find it acceptable to specify the relative size of the expected treatment effects. A measure such as omega squared can provide the information needed to determine power and to fix sample size. We could, for example, estimate omega squared from previous research studies. Alternatively, we could simply specify the relative size of the expected effects by using Cohen's (1977) suggested labels, namely, "small," "medium," or "large," and his numerical values (ω^2 = .01, .06, and .15, respectively) to provide the quantitative estimate. There is a potential danger if we use Cohen's values indiscriminately, of course, as he points out: ". . . these qualitative adjectives . . . may not be reasonably descriptive in any specific area. Thus, what a sociologist may consider a small effect size may well be appraised as medium by a clinical psychologist" (p. 285).[6] On the other hand, Cohen also reminds us that researchers simply cannot shirk their responsibility to make at least a stab at estimating effect size. Again, in his words, "The investigator who insists that he has absolutely no way of knowing how large an [effect size] to posit fails to appreciate that this necessarily means that he has no rational basis for deciding whether he needs to make ten observations or ten thousand" (p. 285).

Usually, we base our power estimates on the *minimum* effect size that we wish to detect. Cohen (1977) suggests that a "small" effect size (ω^2 = .01) is the minimum for experimental research in the behavioral sciences. Most researchers would probably not adopt this value for their research because it is too small and would require far more subjects than they might wish to commit to a study (see Table 4–1). Our choice of effect size should represent a *realistic* estimate, one that is based on earlier research. We do not benefit by casually overestimating the expected effect size. The result of overestimating is an estimate of sample size that is far too small to allow us to detect the actual effect that may be present. If anything, we should be a bit cautious and *under*estimate the effect size so that our choice of sample size will be sure to afford reasonable power for our proposed study.

[5]Kirk (1982, pp. 144–145) describes a related procedure in which we estimate the *largest* difference we expect to detect, relate that difference to an estimate of error variance, and refer this information to a set of special tables.

[6]Cohen (1988) indicates that these values were chosen on the basis of his intuition and were intended only to serve as a guide to researchers who were unable to provide more accurate estimates of expected effect sizes. For additional views on this issue, see, for example, Feldt (1973), Hinkle and Oliver (1983), and O'Grady (1982).

Choosing a Reasonable Value for Power

In addition to an estimate of effect size, we also need to select the degree of power we want our experiment to have. Although I have implied that a power of .50 is too low for the behavioral sciences, I have said nothing about what might be a reasonable or even a desirable level of power. Certainly there is no presumed agreement among researchers on the issue of what defines reasonable power, as there is with regard to significance level. In fact, one could conclude that many researchers tend to *ignore* the power of their experiments (see Sedlmeier & Gigerenzer, 1989). Interestingly, methodologists are beginning to agree that a power of about .80 represents a reasonable and realistic value for research in the behavioral sciences (Cohen, 1965, 1977; Hinkle & Oliver, 1983; Kirk, 1982, p. 144). A power of .80 is reasonable in the sense that it reflects a presumed general sentiment among researchers that type I errors are more serious than type II errors and that a 4:1 ratio of type II to type I error is probably appropriate (see Chase & Tucker, 1976).[7] This value is also realistic, particularly when we consider the sharp increase in the sample size required to increase power from .80 to .90 or higher (see Table 4–1).

The ultimate decision is ours, of course. No one wants to make any error of statistical inference, which is why we agree to set α at a fairly low level—that is, .05—although there are those who advocate reducing the probability of this error lower still (see, for example, Ryan, 1985). One of the main purposes of this chapter is to call attention to the importance of measuring and controlling the other error of statistical inference, β. By using relatively small sample sizes, researchers have unknowingly given relatively greater emphasis to controlling type I error than to controlling type II error. To be more specific, the ratio of type II error to type I error reflected in most experiments reported in the psychological literature is substantially *greater* than 4:1, perhaps closer to a ratio of at least 20:1 (Rosenthal & Rubin, 1985; Rosnow & Rosenthal, 1989b).

What if we cannot perform an experiment with the appropriate number of subjects required to achieve the power we want for our experiment? Some have suggested that one way to cope with this problem is to relax our control of type I error as an additional way of increasing power in an experiment (see, for example,

[7]Assuming that we set α = .05 and power = .80, β thus becomes .20 and the ratio of type II to type I error is .20:.05 or 4:1.

Cohen, 1965, pp. 99–100; Cohen, 1977, pp. 15–16; Rotton & Schönemann, 1978; Stevens, 1986, p. 138). Simply by increasing the significance level from $\alpha = .05$ to .10, for instance, we can increase the probability with which we will reject the null hypothesis—because of this expanded rejection region, of course—and substantially increase power as a consequence.

You might find this option particularly attractive when you are entering a new research area and plan to replicate any significant result you may discover before you present your findings to the professional public. By relaxing your significance level to $\alpha = .10$, for example, you immediately increase your chances of discovering true differences among the treatment means, while the subsequent planned replication—based on this initial study and conducted at $\alpha = .05$—will help you guard against type I error. To elaborate, suppose you did commit a type I error in the first experiment, which is a real possibility because you set $\alpha = .10$ rather than at the more standard significance level of $\alpha = .05$. Because the probability is low that you will make the same type I error in two independent experiments, you stand a reasonable chance of catching this error by obtaining a nonsignificant F in the second experiment. Thus, by adopting a policy of replicating or repeating any experiment that we conducted with a "relaxed" significance level, we are able to protect ourselves from reporting a type I error while maintaining reasonable statistical power and sensitivity in an initial, exploratory investigation.

Using Power Charts to Determine Sample Size

Pearson and Hartley (1951, 1972) have constructed some helpful charts from which we can estimate a sample size that will ensure a particular degree of power. You will recall that four factors—power, significance level, effect size, and sample size—are interrelated and that fixing any three will fully determine the fourth. We estimate sample size with the Pearson-Hartley charts following a somewhat indirect procedure that starts with a trial sample size, which we will refer to as n' to distinguish it from the actual sample size (n) of an experiment. Next, we estimate the smallest effect size we wish to detect and select a significance level. We then use this information to determine the power associated with this particular combination of trial sample size, effect size, and significance level. If the estimated power is either too low or too high, we adjust n' accordingly and determine the power associated with this new combination. We continue this process until we find the sample size that produces the desired level of power. Let's see how this works in practice.

Suppose we propose to conduct an experiment with four treatment conditions and that we can make reasonable estimates of the minimum expected treatment effects, which we base on theoretical considerations and on data obtained from related studies. We will assume that the population treatment means μ_i are the following:

$$\mu_1 = 18,\ \mu_2 = 21,\ \mu_3 = 22, \text{ and } \mu_4 = 19$$

The grand mean μ_T is an average of the four population means ($\mu_T = 20$). In addition, we will assume that an accurate estimate of the common population variance is available, namely $\sigma^2_{S/A} = 16$. These two pieces of information—population treatment means and population variance—are converted into a statistic ϕ^2_A, which is calculated by the following formula:

$$\phi^2_A = n' \frac{\Sigma(\mu_i - \mu_T)^2/a}{\sigma^2_{S/A}} \tag{4–4}$$

where n' = the trial sample size
μ_i = the population treatment means
μ_T = the mean of the population treatment means
a = the number of treatment means
$\sigma^2_{S/A}$ = the average or common variance in the treatment populations

The expression to the right of n' is a ratio of treatment variance relative to error variance. Substituting in Eq. (4–4), we find

$$\phi^2_A = n' \frac{[(18 - 20)^2 + (21 - 20)^2 + (22 - 20)^2 + (19 - 20)^2]/4}{16}$$

$$= n' \frac{10/4}{16} = .1563\,n'$$

Taking the square root of ϕ^2_A gives us

$$\phi_A = \sqrt{.1563\,n'} = .395\sqrt{n'}$$

With ϕ_A expressed in this form, we can now begin the process of choosing a trial sample size, n'.

Let's start with $n' = 16$ as a trial sample size and assume that we want to set power at .90 and $\alpha = .05$. Solving for ϕ_A, we find

$$\phi_A = .395\sqrt{16} = (.395)(4) = 1.58$$

We are now ready to use one of the Pearson-Hartley charts, which is presented in Fig. 4–1. The first thing to notice is that there are two sets of power functions within the body of the chart, one for $\alpha = .05$ and the other for $\alpha = .01$. Within either set, there are 11 different power functions, each associated with a different value of the *denominator* degrees of freedom of the F ratio. As you can see, $df_{denom.}$ = 6, 7, 8, 9, 10, 12, 15, 20, 30, 60, and ∞. Intermediate values may be interpolated visually. The baseline is marked off in increasing values of ϕ; this baseline is used in conjunction with the set of power functions for $\alpha = .05$. (I have omitted the baseline used with the set of power functions for $\alpha = .01$ to simplify this example.) Once all these factors are coordinated on the chart—α, $df_{denom.}$, and ϕ—we can read the power value directly off the ordinate.

The particular power function we need is the one on the left ($\alpha = .05$) associated with $df_{denom.} = (a)(n' - 1) = (4)(16 - 1) = 60$. I have highlighted this function in Fig. 4–1. We now locate $\phi = 1.58$ on the baseline and visually extend a vertical line upward from this point until it intersects with the appropriate power curve. From this point of intersection, we visually extend a horizontal line to the

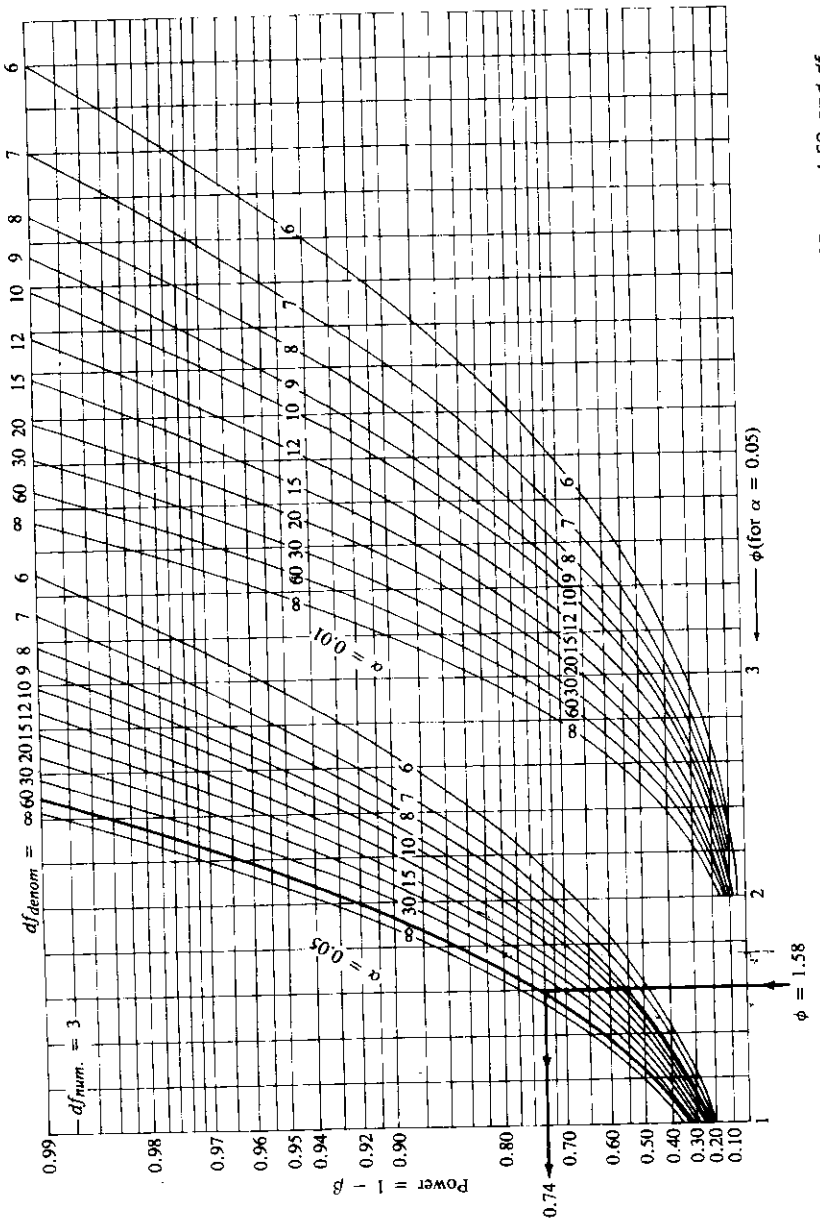


FIGURE 4–1 Illustrating the use of the Pearson-Hartley power charts for $a = 4$ ($df_{num.} = 3$), $\alpha = .05$, $\phi = 1.58$, and $df_{denom.} = 60$.

left until it intersects with the ordinate, where we can then read the estimated power for this combination of factors off the ordinate; I have illustrated these steps in the figure. As you can readily see, power = approximately .74. Since we were aiming for power = .90, however, we need to increase n' and repeat the process. If we try $n' = 24$, for example, we determine

$$\phi_A = .395\sqrt{24} = (.395)(4.899) = 1.94$$

To find out which power function to use with this new trial sample size, we again have to calculate the $df_{denom.}$. In this case, $df_{denom.} = (a)(n' - 1) = (4)(24 - 1) = 92$. This time we visually interpolate between the power functions for $df_{denom.} = 60$ and ∞ and find that the power associated with $\phi = 1.94$ is reasonably close to .90.

To recapitulate, estimating sample size with the Pearson-Hartley charts involves what amounts to a trial-and-error operation. We estimate the minimum treatment effects (or effect size) we wish to detect, decide on the α level, and use this information to solve for ϕ_A expressed in terms of the trial sample size, n'. We then estimate the power of the F test for the trial sample size by obtaining a numerical value for ϕ_A and entering this value appropriately in the relevant power chart. If the resultant power estimate is unsatisfactory (too low or too high), we change the trial sample size in the correct direction and repeat the operations.

We can reduce the number of repetitive calculations in this process by beginning our calculations with a realistic starting value for n'. I will describe how to determine this initial trial size as a series of steps.

1. We identify a particularly likely power function—one that is close to the final function we will use; I suggest using the power function for $df_{denom.} = 60$ for most situations.
2. From this function, we determine the value of ϕ that is associated with the power we seek. In our example, we were striving for a power of .90. Extending a line from .90 on the ordinate to the power function for $df_{denom.} = 60$ and then reading the value of ϕ on the baseline directly below the point of intersection, we find ϕ equal to approximately 1.92.
3. We take the value for ϕ_A^2 we obtained from Eq. (4–4), namely, $\phi_A^2 = .1563\,n'$, and then solve for n'; more specifically,

$$n' = \frac{\phi_A^2}{.1563}$$

4. We substitute our estimated value for ϕ (1.92) into this equation and calculate our first trial sample size as follows:

$$n' = \frac{(1.92)^2}{.1563} = 23.59$$

5. We would use a value of 23 or 24 as our starting point. Since these numbers are actually very close to the value we determined earlier, we have substantially reduced the number of times we would need to recycle through the calculations.

The entire set of Pearson-Hartley power charts is found in Appendix Table A–2. You will find 10 different charts, which are distinguished by the degrees of freedom in the *numerator* of the F ratio; these consist of $df_{num.} = 1, 2, 3, 4, 5, 6, 7,$

8, 12, and 24—a reasonable range for most experimental situations. Please note that there are two scales for ϕ on the baseline, one for $\alpha = .05$, of course, and the other for $\alpha = .01$. You should also note that for all charts *except the first*, the set of power curves on the left and the first set of ϕ values on the baseline are for $\alpha = .05$, whereas the charts on the right and the second set of ϕ values are for $\alpha = .01$. These relative positions are *reversed* for the first chart ($df_{num.} = 1$). Many users of these charts have misread critical information by failing to notice this reversal.

Using a Power Analysis as a Planning Tool

There are alternative ways of approaching the determination of an appropriate sample size that you may find useful during the planning stage of an experiment. One approach, for example, consists of working out the relationship between sample size and power over a useful range of values, rather than focusing on the sample size required for a certain degree of power. Suppose you started with a sample size that is typical for similar or related experiments in your research field and then calculated the power associated with a systematic variation of sample sizes around this value. Let's say that $n = 15$ is a reasonable number for the numerical example we have been using in this section. You could determine the power associated with various trial sample sizes around this value. Alternatively, you could start with the minimum sample size you would consider—the value below which you would begin to question the stability of your treatment means—and then determine power for different values of n above this number. Kraemer and Thiemann (1987, p. 28) define this number as "the minimum sample size necessary for the credibility of a study," which depends, of course, on one's research field.

Suppose that $n = 10$ is our minimum sample size and we will vary sample size in increments of 5 up to $n = 30$, which represents the largest sample size we can possibly afford. Table 4–2 summarizes the results of this analysis. The first row gives the values of ϕ_A for the different trial sample sizes, which we may easily obtain from the formula we calculated previously (see pp. 76–77), namely, $\phi_A = .395\sqrt{n'}$. The second row gives the $df_{denom.}$ for each value of n, which determines the power function we will consult in reading the power chart; the formula for $df_{denom.}$ is, of course, $df_{S/A} = (a)(n - 1)$. The third row contains the power estimates we obtain when we coordinate the values of each ϕ_A with the appropriate power functions in the chart. I have plotted these values in Fig. 4–2. As you can see, we can easily use this curve to obtain reasonable estimates of sample size over a wide range of power. If we need a more accurate determination of power, we can use this curve to make realistic choices for trial sample sizes.

Table 4–2 Power as a Function of Sample Size

	TRIAL SAMPLE SIZES (n')				
	10	**15**	**20**	**25**	**30**
ϕ_A	1.25	1.53	1.77	1.98	2.16
$df_{S/A}$	36	56	76	96	116
Power	.33	.68	.84	.92	.96

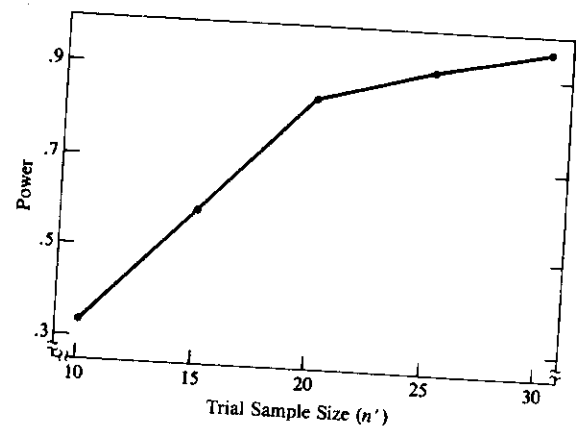


FIGURE 4–2 Power as a function of sample size.

Another way we can use a power analysis is to compare the relationship between sample size and power for a number of different plausible outcomes of our experiment or for different estimates of the common population variance. Under these circumstances, we would need to enter these new values into Eq. (4–4) and solve for ϕ_A as a function of n', duplicating the steps we followed in the numerical example (see pp. 76–80). Once we have calculated this formula, we can easily create a table or a graph depicting the relationship between power and sample size for these new estimates. You should find this information helpful in designing an experiment that is reasonably sensitive to several possible outcomes.

Another use of these power-sample size curves is to examine the trade-off between significance level and power over an extended range of trial sample sizes. Researchers generally have not considered relaxing the standard significance level ($\alpha = .05$) and adopting less stringent levels (for example, $\alpha = .10$ or .25), but this option most certainly is a reasonable possibility (see, for example, Cohen, 1977, pp. 15–16). The widespread use of .05 as the standard of statistical significance in the behavioral sciences is simply a convention and we should be willing to break with this convention when the situation calls for it. If careful analysis on our part reveals that we cannot attain reasonable power without relaxing the rejection standard, we should consider doing so. We can view a power analysis as a form of statistical "contract" in which we specify exactly the circumstances under which we will obtain a certain effect. The significance level is part of that contract. I am *not* recommending that we loosen our control of type I error in general, but only when we have integrated a power analysis into our experimental research plan and data analysis. If the effects we seek are significant, we can immediately plan a replication as a responsible response on our part to the use of a more liberal significance level. On the other hand, if the effects are not significant, we will at least have some guarantee—having set power at .80 or higher—that our research hypotheses have been given a fair and sensitive test.

The power charts of Pearson and Hartley cover only two levels of significance, $\alpha = .05$ and .01, and thus do not allow us to make power estimates with other levels. Fortunately, Rotton and Schönemann (1978) have prepared a detailed table for less stringent significance levels that supplements the Pearson-Hartley charts. More specifically, their table provides information on power for six levels of significance, namely, $\alpha = .05, .10, .20, .30, .40$, and .50. You use this table by locating your estimate of ϕ, obtained in the usual fashion, in the appropriate part of the table—which is determined by α, $df_{num.}$, and $df_{denom.}$—and then reading the value of power found at that point.[8]

Finally, we can sometimes increase power substantially by reducing the number of treatment conditions included in an experiment to increase the sample size for the remaining conditions (see Cohen, 1977, pp. 362–363, 402–403). We would not consider this option, of course, if all the conditions in the experiment are necessary and critical for the study. On the other hand, we may have been overly ambitious and have included additional conditions that provide either supplemental information or information of secondary importance to the main purpose of the study. Under these circumstances, then, we could drop these conditions without damaging the integrity of the experiment. A power analysis is essential for this decision, of course. We must estimate power with the original experimental design, including the original choice of sample size, and then determine the effects on power of discarding certain groups and redistributing the subjects among the conditions that remain. A power analysis is the only rational way to make this sort of decision.

Using Other Estimates of Effect Size

As I mentioned earlier in this section, we can estimate the anticipated effect size for a proposed experiment in a variety of ways. Once we have this estimate, we can translate this information into ϕ_A and then use this index in conjunction with the Pearson-Hartley charts to determine sample size.

Using Cohen's f Statistic. I illustrated the general process of determining sample size by estimating the population treatment means and error variance (the common within-group variance). You may find it difficult, however, to estimate your expected outcome with this degree of specification. If you are unable to provide such a detailed estimate, you might be able to estimate the range between the largest and smallest treatment means and to specify some expected pattern they might reflect. This information, in addition to an estimate of error variance, can be used to produce Cohen's (1977, pp. 276–280) index of effect size f^2, which is related to ϕ^2 as follows:

$$\phi^2 = n' f^2 \tag{4–5}$$

Central to Cohen's formulas for this index is d, which he defines as the range of

[8] You may have to use linear interpolation to determine power for combinations that fall between the values listed in the table.

the means divided by the standard deviation in the treatment populations.[9] More specifically,

$$d = \frac{\mu_{max.} - \mu_{min.}}{\sigma_{S/A}} \tag{4–6}$$

where $\mu_{max.}$ = the largest population treatment mean
$\mu_{min.}$ = the smallest population treatment mean
$\sigma_{S/A}$ = the standard deviation in the treatment populations

Armed with an estimate of d and some idea of how the treatment means are distributed between the two extremes, we can estimate Cohen's f. Table 4–3 gives the formulas for calculating f for the three different dispersion patterns of the means, which we described earlier in this section (p. 74).

To illustrate with the current example, suppose that we were able to estimate only the range of the means ($\mu_{max.} - \mu_{min.} = \mu_3 - \mu_1 = 22 - 18 = 4$) and the population standard deviation ($\sqrt{16} = 4$) and that we anticipated that the other two means would fall in the middle between them (Cohen's "minimum" variation among the means). Using Eq. (4–6), we find

$$d = \frac{4}{4} = 1.00$$

Substituting $d = 1.00$ into the formula listed in the first row of Table 4–3, we calculate

$$f = d\sqrt{\frac{1}{(2)(a)}} = 1.00\sqrt{\frac{1}{(2)(4)}}$$

$$= 1.00\ \sqrt{.125} = .354$$

We can now use Eq. (4–5) to calculate ϕ; that is,

$$\phi_A^2 = n'f^2 = n'(.354)^2 = .125\,n'$$

$$\phi_A = .354\sqrt{n'}$$

Table 4–3 Calculating Cohen's f Statistic for Three Different Patterns of the Population Means

Dispersion Pattern	Formula
Minimum variation	$f = d\sqrt{\frac{1}{(2)(a)}}$
Intermediate variation	$f = d\sqrt{\frac{a+1}{(12)(a-1)}}$
Maximum variation	$f = d\frac{\sqrt{a^2-1}}{(2)(a)}$

[9] Cohen calls this d statistic the **standardized range of the population means**, which is defined as the range of the means divided by the population standard deviation.

We would now begin the process of varying the trial sample size n' until we achieved the level of power we want. As a reminder, we first turn to the third power chart ($df_{num.} = 3$). Using the method described in the last section to select a starting value for n', we find that $\phi = 1.92$ when power is .90 and $df_{denom.} = 60$. Solving for the first trial sample size,

$$n' = \frac{\phi_A^2}{.125} = \frac{(1.92)^2}{.125} = 29.49$$

If we use $n' = 30$,

$$\phi_A = (.354)\sqrt{30} = (.354)(5.477) = 1.94$$

and power is approximately .90.[10]

Using Omega Squared. Even if we are unable to estimate the range of the means and the population standard deviation, we can still conduct a power analysis simply by estimating the population omega squared, ω_A^2. We could obtain such an estimate from pilot work we have conducted or from the work of others appearing in related studies reported in the literature. The relationship between ϕ_A^2 and ω_A^2 is as follows:

$$\phi_A^2 = n' \frac{\omega_A^2}{1 - \omega_A^2} \tag{4–7}$$

Even if we could find no estimate of ω_A^2, we could simply guess at a realistic effect size. Suppose we hoped to detect a "medium" effect, which happens to reflect the average effect size reported in psychological research (see Sedlmeier & Gigerenzer, 1989). You will recall that Cohen (1977) has defined a "medium" effect as $\hat{\omega}_A^2 = .06$. Substituting this value in Eq. (4–7), we find

$$\phi_A^2 = n' \frac{.06}{1 - .06} = n' \frac{.06}{.94}$$

$$= .0638\, n'$$

Taking the square root of ϕ_A^2 gives us $\phi_A = .253\sqrt{n'}$, which we can use to estimate the sample size required to achieve a particular degree of power. Without showing the steps, I determined that n must be approximately 58 to provide us with a power of .90 to detect a "medium" effect.[11]

Summary of Formulas. Table 4–4 provides a summary of the various formulas by which ϕ^2 may be calculated. These formulas will permit you to shift from one measure of effect size to another and to use information available in other power charts and tables that are based on one of these measures.

[10] If we simply cannot specify a pattern for the minimum treatment effects we wish to detect, we can use the d statistic alone to estimate sample size. Under these circumstances, we can turn to a convenient table prepared by Hinkle and Oliver (1983) for exactly this situation. In the present example, we would locate the column for $d = 1.00$, which is designated 1.0 σ in the table, and find the sample size for $a = 4$ groups, $\alpha = .05$, and power .90; the estimated value from their table is $n = 29$.

[11] See problem 6b in the exercises if you want to determine this value yourself.

Table 4–4 Calculating ϕ^2 for the Pearson-Hartley Power Charts

General formula:	$\phi^2 = n \frac{\Sigma (\mu_i - \mu_T)^2 / a}{\sigma^2_{S/A}}$
Relation to f:	$\phi^2 = nf^2$
Relation to ω^2:	$\phi^2 = n \frac{\omega^2}{1 - \omega^2}$

Using the Cohen Power Tables

Cohen (1977) provides a useful set of tables for calculating the sample size needed to control the power of an experiment (pp. 381–389).[12] He presents three basic tables, one for three different significance levels ($\alpha = .10$, .05, and .01). Each basic table consists of 11 subtables, which are differentiated by the number of degrees of freedom associated with the numerator of the F ratio ($df_{num.} = 1, 2, 3, 4, 5, 6, 8, 10, 12, 15$, and 24). For our example, we would need the subtable for $df_{num.} = 3$ (Table 8.4.4). Each of these subtables lists a range of values for Cohen's f statistic (.05, .10, .15, .20, .25, .30, .35, .40, .50, .60, .70, and .80) and a range of values for power (.10, .50, .70, .80, .90, .95, and .99). Since $f^2 = \phi_A^2 / n'$, we can use the value for ϕ_A^2 we calculated previously for our original example (see p. 77). More specifically, we found that $\phi_A^2 = .1563\, n'$. Thus, $f^2 = .1563$ and $f = \sqrt{.1563} = .395$. If we coordinate $f = .395$ with power = .90 for the $\alpha = .05$ significance level, using Cohen's Table 8.4.4 (p. 384), we find a sample size of approximately $n = 23$, which is identical to the one we obtained by using the Pearson-Hartley charts.

You may find Cohen's tables a useful way to estimate sample size. They encompass a wide variety of experimental situations and there is the added bonus of a less stringent significance level ($\alpha = .10$), if you are willing to increase type I error from $\alpha = .05$ to $\alpha = .10$ as a way of achieving greater power. One possible drawback is the need to use linear interpolation for entries that fall between those categories of f and power that are provided in the table.

Using Computer Programs to Estimate Sample Size

Software programs are beginning to appear that greatly facilitate the estimation of power and sample size. I can easily envision the time when tables and charts will be replaced by computer programs that provide the same information with greater speed and more accuracy. When we reach that time, there will be no good excuse

[12] The most recent version of Cohen's book (1988) provides updated references and includes some new material. The chapter dealing with the analysis of variance (Chap. 8) received only minor revisions, however, thus allowing researchers to use the material in the revised edition (1977) and the second edition (1988) interchangeably.

for researchers to fail to take power into consideration during the planning stages of their experiments. Although it is true that researchers will still have to provide estimates of the minimum effect sizes they wish to detect—for many, this is a serious stumbling block—the programs will inform them quickly what the sample size must be or what power will be achieved. Researchers will be able to try out different values for α, power, and n to determine an optimal course of action, that is, a feasible combination of significance level, power, and sample size for their studies. I find it difficult to offer specific recommendations here because information about the availability of programs is rapidly changing and any advice will soon be out of date. You will probably find out about useful programs from your friends and colleagues. This source of information is particularly appealing because you will have someone who may be willing to help you learn how to use these specialized programs, not a minor consideration when working with a new computer program.

Where else might you find out about statistical software programs? First, there are several software directories, which should be kept reasonably up to date. The American Psychological Association, for example, publishes a software directory for psychologists (Stoloff & Couch, 1988). A more useful directory, which lists over 200 statistical software packages, is also available (Woodward, Elliot, Gray, & Matlock, 1988). Second, certain professional journals provide a timely source of information. *Behavior Research Methods, Instruments, & Computers* and *Educational and Psychological Measurement*, for example, publish announcements of new programs developed by researchers, which are usually available from the author for a nominal sum. The *American Statistician* also publishes announcements of new programs as well as substantial reviews of established statistical software packages. Detailed reviews may also be found in two other journals, the *British Journal of Mathematical and Statistical Psychology* and *Multivariate Behavioral Research*. Finally, some computer magazines, such as *Byte* and *PC Magazine*, periodically publish in-depth reviews of statistical programs.

I do want to mention several programs that are currently available for power determination, although I cannot guarantee they will be available when you read this paragraph. Jacob Cohen, who has devoted much of his professional life to the power-sample size problem, has published a software package with Michael Borenstein entitled Statistical Power Analysis (Borenstein & Cohen, 1988). A useful program called Ganova 4—Power Computations has been developed by researchers in the Psychology Department at the University of California, Los Angeles (Brecht, Woodward, & Bonett, 1988). Another program, PC-Size, is available at nominal cost from its author (see Dallal, 1986). A final program, called STAT POWER, is written for the Apple computer (Anderson, 1981). Additional programs designed for determining sample size and power with MS/PC-DOS computers have been reviewed and compared by Goldstein (1989).

4.5 ESTIMATING THE POWER OF AN EXPERIMENT

Our focus up to this point has centered on the use of power to provide a rational basis for choosing an appropriate sample size. There is another use for power,

which usually takes place *after* an experiment has been completed, a procedure sometimes called a **post hoc power analysis**. The power analyses reported by Cohen (1962), Sedlmeier and Gigerenzer (1989), and others, which supported the conclusion that many of the studies reported in our journals are underpowered, are examples.

These analyses were conducted on studies reporting significant statistical tests. Post hoc power analyses are frequently useful when we attempt to interpret an F test that is not significant. Does a nonsignificant test mean that the null hypothesis is true in the sense that the differences are either trivial or nonexistent, or does a nonsignificant test mean that there may be differences present but that there was insufficient power to detect them? One way to resolve this ambiguity is to estimate omega squared or some other index of relative treatment magnitude. Since estimated omega squared is essentially uninfluenced by sample size, its magnitude will provide useful information regardless of the underlying power of the experiment on which it based. It is for this reason that editors and reviewers urge that we estimate relative treatment magnitude for *all* statistical tests we perform. A "small" but significant F might suggest the presence of a "trivial" effect that was detected by a particularly powerful experiment, whereas a "medium" but nonsignificant F might suggest the possible presence of an "important" effect that was not detected because of a serious lack of power.[13]

Estimating Power from the Pearson-Hartley Charts

We can easily estimate the power of an experiment from the Pearson-Hartley charts by estimating omega squared and then using other information to enter the charts and determine power. Let's see how this is done. Suppose that we came across an interesting experiment with the following characteristics:

$$F = 3.20, a = 3, \text{ and } n = 5$$

Since the critical value of $F(2, 12) = 3.89$, the F is not significant ($\alpha = .05$). Using Eq. (4–2), we find

$$\hat{\omega}_A^2 = \frac{(a-1)(F-1)}{(a-1)(F-1)+(a)(n)}$$

$$= \frac{(3-1)(3.20-1)}{(3-1)(3.20-1)+(3)(5)} = \frac{4.40}{4.40+15} = .227$$

Simply by estimating omega squared we have made an important discovery, namely, there is a potentially important effect (falling well within Cohen's "large" category) reflected in these data. The reason the researcher failed to reject the null hypothesis is probably because of the low power afforded by the relatively small sample size ($n = 5$).

[13]For an example of the importance of considering power when interpreting nonsignificant results, see Stevens (1986, p. 137).

To use the Pearson-Hartley charts, we need to calculate ϕ. From Table 4–4, we find that

$$\phi^2 = n\frac{\omega^2}{1-\omega^2}$$

Substituting in this equation the value for omega squared we estimated for this experiment ($\hat{\omega}_A^2 = .227$), we obtain

$$\hat{\phi}_A^2 = n\frac{\hat{\omega}_A^2}{1-\hat{\omega}_A^2} \qquad (4\text{–}8)$$

$$= 5\frac{.227}{1-.227} = (5)(.2937) = 1.469$$

and $\hat{\phi}_A = \sqrt{1.469} = 1.21$. We now turn to the second chart in Table A–2 ($df_{num.} = 2$) and find the power curve for $df_{denom.} = 12$. For $\phi = 1.21$, power is approximately .36. You can understand now why the F test was not significant—power was simply too low to detect an effect even this large.

What sort of sample size would we need to detect this effect? We can use this same information to estimate the sample size we would need to reject the null hypothesis at $\alpha = .05$, but at a more comfortable power of .80. From Eq. (4–8), we obtain

$$\hat{\phi}_A^2 = n'\frac{.227}{1-.227} = .294\,n' \text{ and } \hat{\phi}_A = .542\,\sqrt{n'}$$

If we try $n' = 12$ as a trial sample size,[14] we find

$$\hat{\phi}_A = .542\sqrt{12} = (.542)(3.464) = 1.88$$

Since $df_{denom.} = (a)(n-1) = (3)(12-1) = 33$, we will use the power function for $df_{denom.} = 30$. Locating $\phi = 1.88$ on this curve, I estimate power to be approximately .80. Assuming that this particular sample size does not exhaust our resources, we would design a new experiment with $n = 12$ as the sample size and know that the study will be reasonably sensitive (power = .80) for detecting an effect of this relative magnitude ($\hat{\omega}_A^2 = .227$).

Estimating Power from the Cohen Tables

Since Cohen's (1977, 1988) tables are familiar to many researchers, I will show how to estimate the power of an experiment by using his system of calculation. Cohen's extensive tables (pp. 289–354) require the calculation of f rather than $\hat{\phi}_A$ We can calculate f^2 with the following formula:

[14] I used the procedure I outlined earlier in this chapter for arriving at this starting sample size (see p. 79). To illustrate briefly, I first determined the value of ϕ associated with a power of .80 from the power function for $df_{denom.} = 60$; the value I found was $\phi = 1.83$. By expressing the formula, $\hat{\phi}_A^2 = .294\,n'$, in terms of n'—that is, $n' = \hat{\phi}_A^2/.294$—and substituting $\phi = 1.83$ in the new expression, I determined $n' = (1.83)^2/.294 = 11.39$. I chose to use $n' = 12$ as a convenient starting value.

$$f^2 = \frac{\hat{\omega}_A^2}{1-\hat{\omega}_A^2} \qquad (4\text{–}9)$$

Substituting $\hat{\omega}_A^2 = .227$ in Eq. (4–9), we find

$$f^2 = \frac{.227}{1-.227} = .294 \text{ and } f = \sqrt{.294} = .542$$

Consulting Cohen's Table 8.3.13 (p. 313) for $\alpha = .05$, $f = .54$, $df_{num.} = 3$, and $n = 5$, I find power to be approximately .37, very close to the value we obtained with the Pearson-Hartley charts.

4.6 "PROVING" THE NULL HYPOTHESIS

When researchers obtain a nonsignificant F, many are tempted to conclude that the independent variable produced no systematic effects on the dependent variable. Or stated more strongly, they are tempted to conclude that there are no treatment effects in the population, that they have in effect "proved" that the null hypothesis is true. What exactly can we conclude under these circumstances? Failing to reject the null hypothesis means that the differences we observed in our experiment were *too small* to permit us to conclude that the population treatment means are different. We do not conclude that treatment differences are absent or lacking, but that the experiment is not sufficiently sensitive to detect them if they did exist.

Is there any way we can determine whether treatment effects are truly absent? Technically, no—we cannot *prove* that the population treatment means are identical. On a more practical level, however, we could specify a band or set of values that we consider "trivial," "unimportant," or "negligible," which we are willing to assert is functionally equivalent to the complete absence of treatment effects. Some methodologists refer to this band as the **null range** (for example, Greenwald, 1975; Hays, 1973, pp. 850–853). Suppose we conducted an experiment with sufficient power to allow us to detect an effect just outside the null range. By creating an experiment with a low degree of risk for both type I *and* type II error, we could reasonably conclude that treatment effects are "trivial" (or functionally nonexistent) when we fail to reject the null hypothesis.[15]

It is important to note how this procedure I just described parallels the more familiar procedure of hypothesis testing. In both cases, we set up a situation in which we specify the risks involved when we reject the null hypothesis (α) and when we fail to reject the null hypothesis (β). We establish the risk of a type I error (α) by choosing an appropriate significance level and of a type II error (β) by estimating the size of the minimum treatment effects we wish to detect and choosing a sample size to achieve it. The "minimum treatment effects" in the more typical case of hypothesis testing refer to the smallest set of differences of any interest to us, whereas in the case of proving the null hypothesis they refer to the largest set of differences that we would still consider functionally equivalent to zero.

[15] The steps involved in accepting the null hypothesis are described in an interesting article by Greenwald (1975).

Experiments designed to prove the null hypothesis are rare in psychology. Moreover, many of these are deficient in the sense that they have failed to establish convincing degrees of power (Sedlmeier & Gigerenzer, 1989). As I have argued, it is not sufficient simply to fail to reject the null hypothesis to "prove" it, but you must do so under conditions of high power. Most researchers are unwilling to commit the resources necessary to achieve this experimental state of affairs. As an example, suppose we consider an effect size of $\omega^2 = .01$ as representing the transition between a negligible effect and one of some interest. Let's assume that we have only two conditions and that we would like to establish the same risk for both types of error, namely, $\alpha = .05$ and $\beta = .05$, which seems most appropriate under the circumstances. To use the Pearson-Hartley charts (Table A–2), we need to calculate ϕ_A, which we can obtain by means of Eq. (4–8):

$$\phi_A^2 = n' \frac{\omega_A^2}{1 - \omega_A^2} = n' \frac{.01}{1 - .01}$$

$$= .0101\, n'$$

Taking the square root of this value gives us $\phi_A = .1005\sqrt{n'}$. We can facilitate the process of finding the trial sample size n' by finding the value of ϕ that is associated with a power of .95; we will use the power function for $df_{denom.} = \infty$, which should give us a reasonable approximation for any large value of n'. From the first chart ($df_{num.} = 1$), we determine that for a power of .95, we will need a ϕ of 2.55 (please note that the power functions for $\alpha = .05$ are drawn on the right and the appropriate scale on the baseline is the lower one). Solving for n', we find

$$n' = \frac{\phi_A^2}{.0101} = \frac{(2.55)^2}{.0101} = 643.81$$

Thus, we will need a sample size of around $n' = 644$ subjects to provide the desired power for detecting this minimum effect. Perhaps you can understand now why few researchers ever take these steps to "prove" the null hypothesis. Even if we relax our target power to .90 or .80, we still would need large numbers of subjects ($n = 519$ and 384, respectively).

The purpose of this section is to point out the misconception that many researchers harbor that a nonsignificant difference between two groups implies that there is no difference between the two groups. I argued that the appropriate conclusion in such a situation, which revolves around a specification of and a concern for power, is that the experiment may have not been sufficiently sensitive to detect a true difference. I also indicated that an experiment that is specifically designed to prove the null hypothesis usually requires a huge commitment of subjects and, as a consequence, is likely to be undertaken only when it is vitally important to show that a particular effect does not exist (for an example of such an undertaking, see Gillig & Greenwald, 1974).

4.7 EXERCISES[16]

1. Table 3–2 (p. 45) summarizes the analysis of a single-factor experiment.
   a. Calculate $\hat{\omega}_A^2$ from these data, using both Eq. (4–1) and Eq. (4–2).
   b. Using the same data, calculate R^2.
   c. What do these two quantities tell you about the outcome of this experiment?
2. a. Calculate $\hat{\omega}_A^2$ with the data presented in problem 2 in the exercises for Chap. 3.
   b. What does this quantity tell you about the outcome of this experiment?
3. Consider the analysis of variance obtained from two single-factor experiments summarized in the accompanying table.
   a. Calculate $\hat{\omega}_A^2$ for these two studies.
   b. What has this index told you about the relative outcomes of these two experiments?

Summary Tables for Two Experiments

Source	*SS*	*df*	*MS*	*F*
A	233.33	2	116.67	6.52*
S/A	376.00	21	17.90	
Total	609.33	23		
A	233.33	2	116.67	6.52*
S/A	1,557.30	87	17.90	
Total	1,790.63	89		

*$p < .05$.

4. Suppose an experimenter is planning an experiment with $a = 5$ different treatments and is able to assume the following population data: $\mu_1 = 10$, $\mu_2 = 10$, $\mu_3 = 14$, $\mu_4 = 16$, $\mu_5 = 15$. On the basis of past research, the experimenter estimates the population error variance to be 15.
   a. What sample size will the researcher need to achieve power of .80 at $\alpha = .05$?
   b. What sample size is needed to achieve power of .90?
   c. Suppose the researcher prefers to work at the .01 significance level. What sample sizes would be needed to achieve power of .80 and .90?
5. Cohen (1977) proposed a simplified way of specifying a pattern of results, which we considered in Sec. 4.4. Let's see how this procedure works with a a numerical example. We will assume that $a = 8$, $n = 9$, and the population error variance 27.56. We next estimate the smallest and largest population means, $\mu_{min.} = 5$ and $\mu_{max.} = 12$, and then assign values to the remaining means according to one of three possible patterns, which follow. Estimate power for each of these patterns, assuming that $\alpha = .05$.
   a. The pattern of means reflects *minimum variability* (the six remaining means are placed at the midpoint between the two extremes).
   b. The pattern of means reflects *intermediate variability* (the six means are spaced equally between the two extremes).

[16]The answers to these problems are found in Appendix B.

EXPERIMENT CHECKLIST (Net Station 2.0)

Subject Number _____ **Gender: F M DOB:** _____ **Date:** _____

Experiment Code: _____ **Experimenter ID:** _____ **Time:** _____

Net Used _____ **Stimulus OFFSET:** _____

1. Turn on the equipment at least 30 minutes prior to the testing. _____
2. Make new electrolyte. _____
3. Set Net Station
    a. Click New Session _____
    b. Choose Standard Session (or 2nd run) _____
    c. Enter file name: <study>_<Subject#><f/m>_<experiment> _____
    d. Click "Rename Session" and navigate to the study folder _____
    e. Create new folder for subject (or place file into subject's folder) _____
    f. Click "New" to place the file into the selected folder _____
    g. Click "Begin Session" _____
    h. Wait until Net Station is finished calibrating the amplifiers _____
4. Check the sound level for auditory stimuli (80 dB) or response button assignment for visual experiments _____
5. Inform the nurse that you are taking the baby and wheel the bassinet to the testing room. _____
6. Dim the lights in the testing room. _____
7. Measure the circumference of the subject's head (in centimeters!) _____
   and locate vertex. nasion-inion _____ meatus-meatus _____
8. Apply the net and connect it to the amplifiers. _____
9. Measure Impedance and take a screen picture (Shift-Apple-3) _____
   Record bad channels and corresponding impedances:

1.	5.
2.	6.
3.	7.
4.	8.

10. Set E-prime experiment _____
11. Position the speaker over the subject's vertex and/or the monitor 3 feet in front of the subject. _____

partial omega squared is necessarily smaller than the one for Eq. (10–8), the partial omega squared will be larger.

Comparing the Sizes of Omega Squared

Researchers are frequently interested in comparing the sizes of estimated omega squared within the same factorial experiment. In these cases, they might ask if the main effect for factor A is larger than the main effect for factor B or for the interaction. Although it is tempting to compare these $\hat{\omega}^2_{effect}$'s directly, any conclusion about differences must be based on appropriate statistical tests. It is not sufficient to conclude that the estimates of omega squared for two sources are different when one source is significant and the other is not (Rosenthal & Rubin, 1982). Ronis (1981) introduced a statistical procedure for comparing treatment magnitudes, but it can be applied only to designs in which all factors consist of only two levels. Fowler (1987) provides a more general method, which corrects this limitation, but it does so with a procedure that most researchers will find difficult to apply.

Even if we were able to compare two estimates of treatment magnitude statistically, however, we still must take care in how we interpret any significant difference we may find. As Ronis (1981) puts it, "to draw clear conclusions about the relative impact of two independent variables, we must have some assurance that the manipulations of the two variables are of comparable magnitudes" (p. 998). Applying this caution to our numerical example, we could ask how meaningful a comparison is between the treatment magnitudes of a main effect based on different drugs (factor A) and one based on two levels of food deprivation (factor B). Ronis discusses several ways in which a researcher might select comparable manipulations for the two independent variables, but these are often difficult to achieve in practice.

In short, as tempting as such comparisons may be, we should keep in mind that comparisons of treatment magnitude within an experiment must be coupled with an appropriate statistical test and that they will be difficult to interpret unambiguously unless we can show that the two independent variables represent comparable manipulations.

10.8 DETERMINING SAMPLE SIZE

We considered power and design sensitivity in Chap. 4, where you were shown how we can use power determinations to choose an adequate sample size. (See pp. 76–80 for a review of these procedures.) We can use power estimates to serve the same important function in factorial designs as well.[7]

Using Population Deviations

To estimate sample size, we have to specify the nature of the population treatment effects we are interested in detecting and to guess at the magnitude of the error

[7]An extension of power estimates to factorial designs is given comprehensive treatment by Cohen (1977, Chap. 8; see especially pp. 376–379 and 400–403 for application to the two-factor design).

variance we expect to find in the experiment. These and other values, including a trial sample size (n'), are entered in a formula that provides a value for ϕ^2_{effect}; we then refer the square root of this value, ϕ_{effect}, to appropriate power charts to estimate the power associated under these circumstances. If the power is inadequate, a new trial sample size is used to calculate ϕ^2_{effect} again and to determine the level of power achieved by this increase in sample size. We repeat this trial-and-error procedure until we achieve the level of power we want.

A general formula for ϕ^2_{effect} is given by

$$\phi^2_{effect} = \frac{(\text{no. obsn})[\Sigma(\text{dev.})^2]}{(df_{effect} + 1)(\sigma^2_{error})} \qquad (10\text{–}10)$$

where (no. obsn) = the number of observations that will contribute to each basic deviation
$\Sigma(\text{dev.})^2$ = the basic population deviations constituting the treatment effects in question
df_{effect} = the df associated with the treatment effects, calculated in the usual fashion
σ^2_{error} = the population error variance

Table 10–10 illustrates how Eq. (10–10) is adapted for the $A \times B$ factorial design. The basic deviations for each factorial effect are listed in column 1 of the table. These deviations come from the formal statement of the structural model (see Sec. 10.5). The numbers of observations on which estimates of these deviations would be based in an actual experiment appear in column 2. For the main effect of factor A, $(b)(n')$ observations are available to estimate the deviation for any given mean; for the main effect of factor B, $(a)(n')$ are available; and for the interaction, n' observations are available. (You will recall that n′ represents the trial sample size we systematically adjust when using the power charts to estimate the sample size required to achieve certain power with a new experiment.) The formulas for ϕ^2_{effect}, which result from substituting relevant values into Eq. (10–10), are presented in column 3 of the table.

Consider these three formulas. In determining the sample size to be used in

Table 10–10 Formulas for ϕ^2_{effect} in the $A \times B$ Design

Source	(1) Deviation	(2) Number of Observations	(3) ϕ^2_{effect}
A	$\alpha_i = \mu_i - \mu_T$	$(b)(n')$	$\frac{(b)(n')[\Sigma(\alpha_i)^2]}{(a)(\sigma^2_{S/AB})}$
B	$\beta_j = \mu_j - \mu_T$	$(a)(n')$	$\frac{(a)(n')[\Sigma(\beta_j)^2]}{(b)(\sigma^2_{S/AB})}$
$A \times B$	$(\alpha\beta)_{ij} = \mu_{ij} - \mu_i - \mu_j + \mu_T$	n'	$\frac{n'[\Sigma(\alpha\beta_{ij})^2]}{[(a-1)(b-1)+1](\sigma^2_{S/AB})}$

any given experiment, we will vary n' since the levels of factors A and B (a and b, respectively) are determined by the nature of the experimental questions we want to ask and thus are presumably fixed at this stage of the planning. If we are interested in achieving a certain power for all three factorial effects, the final sample size will be determined by the *largest* estimate of n'. Generally, the largest estimate will come from the interaction because power is in part a function of the actual number of observations contributing to the different means; due to the nature of the factorial design, fewer observations contribute to the cell means (that is, interaction) than to either set of marginal means (that is, the main effects). Thus, if we are interested primarily in interaction, which often will be the case with a factorial design, we only need to work with the corresponding relevant formula in estimating sample size.

An Example. Let's consider an example based on an actual experiment.[8] A researcher was interested in the possibility that a certain drug administered after learning would enhance memory for the task 24 hours later. On Day 1, laboratory rats were placed in an apparatus, in which they were administered an electric shock if they failed to enter a distinctive adjoining chamber within 30 seconds; each rat was given 2 chances to avoid the shock. Immediately following training, one group was administered a drug (the experimental condition) and another some inert substance (the control condition). On Day 2, all animals were tested 8 more times in the avoidance apparatus; the dependent variable consisted of the number of times each animal avoided the shock over the 8 trials. After successfully replicating the facilitating property of the drug in a number of related experiments, the researcher planned a more elaborate series of experiments designed to pinpoint the locus of the effect of the drug in the brain. This would be accomplished by introducing a second independent variable (operation) in which animals either had a particular area of the brain removed surgically or were given a control or sham operation. The basic design is a factorial in which factor A consists of two levels (control and drug) and factor B consists of two levels (sham and actual operation).

The earlier research provided the experimenter with stable estimates of the sorts of drug effects he could expect in this experimental setting and of error variance. He predicted the following idealized outcome:

	No Drug (a_1)	Drug (a_2)	Mean
Sham operation (b_1)	4.2	5.8	5.0
Actual operation (b_2)	3.0	3.0	3.0
Mean	3.6	4.4	4.0

His estimate of error variance was $\sigma^2_{S/AB} = 2.5$. As you can see from the predicted means, the researcher expected to obtain the usual enhancement of performance

[8] I wish to thank Dr. Joe Martinez and Patricia Janak for providing me with this illustration.

by the drug for the animals receiving the sham operation (4.2 versus 5.8 avoidances) and to eliminate the effect completely for the animals receiving the actual operation (3.0 versus 3.0)—an interaction; he also expected a depression in performance for both groups of animals receiving the operation—a main effect.

We begin by calculating the deviations representing the interaction effect. Using the cell and marginal means from the factorial matrix, we find

$$(\alpha\beta)_{1,1} = \mu_{A_1B_1} - \mu_{A_1} - \mu_{B_1} + \mu_T = 4.2 - 3.6 - 5.0 + 4.0 = -.4$$
$$(\alpha\beta)_{2,1} = \mu_{A_2B_1} - \mu_{A_2} - \mu_{B_1} + \mu_T = 5.8 - 4.4 - 5.0 + 4.0 = .4$$
$$(\alpha\beta)_{1,2} = \mu_{A_1B_2} - \mu_{A_1} - \mu_{B_2} + \mu_T = 3.0 - 3.6 - 3.0 + 4.0 = .4$$
$$(\alpha\beta)_{2,2} = \mu_{A_2B_2} - \mu_{A_2} - \mu_{B_2} + \mu_T = 3.0 - 4.4 - 3.0 + 4.0 = -.4$$

Next, we sum the squared deviations:

$$\Sigma(\alpha\beta_{ij})^2 = (-.4)^2 + (.4)^2 + (.4)^2 + (-.4)^2 = .64$$

Substituting in the formula for $\phi^2_{A \times B}$ in Table 10–10, we have

$$\phi^2_{A \times B} = \frac{n'(.64)}{[(2-1)(2-1)+1](2.5)} = \frac{n'(.64)}{2(2.5)} = .128\,n' \qquad (10\text{–}11)$$

To use the power charts, we need the square root of ϕ^2:

$$\phi_{A \times B} = \sqrt{.128\,n'} = .358\sqrt{n'}$$

We are now in a position to use the Pearson-Hartley power charts in Table A–2. Since the $df_{num.} = 1$, we will use the first power chart. We now begin the process of trying different sample sizes until we find one that produces reasonable power (.80). Suppose we try $n' = 31$ as our trial sample size.[9] Then,

$$\phi_{A \times B} = .358\sqrt{31} = (.358)(5.57) = 1.99$$

This calculation establishes a location on the baseline of the power chart ($\alpha = .05$) of $\phi = 1.99$. We now need to determine which power function is appropriate:

$$df_{denom.} = (a)(b)(n-1) = (2)(2)(31-1) = 120$$

Using the curve for $df_{denom.} = \infty$, we find power to be slightly greater than our targeted value of .80. It appears, therefore, that a sample size of about $n = 30$, to pick a round number, will provide reasonable power to detect this interaction effect. If the resulting power were appreciably lower (or higher) than .80, we would have to raise (or lower) the trial sample size and assess the consequences of this change. We would continue this process until the desired level of power was achieved.

[9] A strategy for selecting the initial trial sample size was introduced in Chap. 4 (p. 79). Briefly, we begin with Eq. (10–11) and solve for n'; that is, $n' = \phi^2_{A \times B}/.128$. We now find the value for ϕ associated with power .80; using the curve at $df_{denom.} = 60$ from the first power chart, we find $\phi = 2.0$. Finally, we substitute this value in the equation we just obtained and solve for n'. This gives us $n' = (2.0)^2/.128 = 31.25$. I used $n' = 31$ as the trial sample size.

Determining Sample Size for an Exact Replication

Researchers often wish to repeat an experiment exactly but with a sample size that will assure significance when it is conducted a second time. The first experiment may have been a pilot study, or the results may have been unexpected but interesting and worthy of replication. Although it is usually not appropriate to base a power analysis on sample data, which reflect to some extent error variability, we can use a procedure that circumvents this problem by using the F ratios obtained in the original study.[10] (An F ratio, in which MS_{effect} is divided by MS_{error}, adjusts treatment effects for error variability.)

Let's assume that our interest centers primarily on the $A \times B$ interaction, which is a reasonable assumption given the nature of the design. In the specific case, in which an exact replication is being attempted, differing only in the number of subjects that are run, we can use the following formula to estimate the ϕ^2 for the interaction:

$$\phi^2_{A \times B} = \left[\frac{n'}{n_{old}}\right]\left[\frac{(a-1)(b-1)}{(a-1)(b-1)+1}\right]\left[F_{A \times B} - 1\right] \tag{10–12}$$

where n' is the trial sample size for the *new* experiment and n_{old} is the actual sample size for the original experiment. An example of this general procedure is given in problem 4 in the exercises.

Reemphasizing the Need to Control Power

Conducting a power analysis can be a sobering experience, particularly when you discover that you will need an unacceptably large sample size to obtain reasonable assurance (that is, power = .80) that you will be able to detect the differences you have hypothesized. On the other hand, what is the sense of conducting the experiment if the power is low? Low power means that you have a poor chance of detecting these differences when they are real, that is, actually present in the treatment populations. There are other ways to increase power, which we discussed in Chap. 4 (pp. 80–82). One of these, which deserves thoughtful consideration, is adopting a more flexible significance level ($\alpha = .10$, for example). An increase in the rejection region produces an increase in power.

Any relaxation of our protection against type I error will need to be defended, of course. The argument is greatly strengthened, in my opinion, when theory predicts certain specific outcomes. In the example we have been considering in this section, the researcher predicted a particular pattern of results. He was not interested in a significant interaction in general but one in which the drug effect is eliminated by the removal of specific brain tissue. Any other sort of interaction would work against his theory. I believe that explicit planned outcomes such as this one are ideal candidates for more flexible significance levels, but only when the significance level is selected in conjunction with a power analysis conducted during the planning stages of an experiment.

[10] I am indebted to Dr. Thomas D. Wickens, who called my attention to this procedure.

Behavioral scientists are generally not familiar with this quite sensible option. By ignoring power when we design experiments, we have no real choice except to focus on the control of type I error—we choose a reasonable and affordable sample size and set the significance level (usually $\alpha = .05$). Power (and type II error) is virtually uncontrolled. Numerous reviews by methodologists (for example, Cohen, 1962; Sedlmeier & Gigerenzer, 1989) reveal that our experiments are substantially underpowered. On the other hand, this is exactly what can happen when we conduct experiments with no information about power. In contrast, choosing to set power in advance represents a responsible way of achieving a rational balance between type I and type II errors. If the results are not significant or if we fail to obtain the predicted pattern of results, we at least know what the risks were of committing a type II error. We are in a much stronger position to accept a negative outcome under these circumstances than if we made no attempt to control power.

10.9 EXERCISES[11]

1. A two-variable factorial experiment is designed in which factor A consists of $a = 5$ equally spaced levels of shock intensity and factor B consists of b 3 discrimination tasks of different difficulty (b_1 = easy, b_2 = medium, and b_3 = hard). There are $n = 5$ rats assigned to each of the $(a)(b) = (5)(3) = 15$ treatment conditions. The animals are to learn to avoid the shock by solving the discrimination task within a 10-second period. The response measure consists of the number of learning trials needed to reach the criterion of an avoidance of the shock on three consecutive trials. The data are given in the following data matrix.

a_1	a_1	a_1	a_2	a_2	a_2	a_3	a_3	a_3	a_4	a_4	a_4	a_5	a_5	a_5
b_1	b_2	b_3	b_1	b_2	b_3	b_1	b_2	b_3	b_1	b_2	b_3	b_1	b_2	b_3
6	14	15	5	12	14	8	11	16	13	14	16	15	15	17
7	18	18	11	10	17	11	10	20	12	19	18	19	12	15
3	12	14	6	15	15	13	15	17	10	17	19	13	16	19
4	13	13	5	14	11	9	17	13	14	12	11	17	18	14
9	11	15	7	11	14	7	12	16	9	13	14	12	13	16

   a. Conduct an analysis of variance on these data. Reserve your calculations for problem 4 in the exercises for Chap. 11.
   b. Construct a 95 percent confidence interval based on the following means: one of the cell means ($\bar{Y}_{A_1B_1}$) and two of the marginal means, one contributing to the main effect of factor A ($\bar{Y}_{A_2}$) and the other to the main effect of factor B ($\bar{Y}_{B_3}$).
   c. Estimate standard omega squared for each of the factorial effects.
   d. Estimate partial omega squared for each of the factorial effects.
2. Consider the factorial design displayed in the following data matrix and the scores produced by the $n = 3$ subjects in each of the treatment conditions.

[11] The answers to these problems are found in Appendix B.

$$= \frac{14.93}{15.67} \times 100 = 95.3 \text{ percent}$$

A value greater than 100 percent would indicate greater efficiency with the within-subjects design. The value of 95.3 percent is not what we would expect since a value of less than 100 percent implies that the within-subjects design is less efficient than the between-subjects design.

What happened in this example? The data for this numerical example were created to demonstrate that the sensitivity of a within-subjects design may be masked by the presence of sizable *practice effects*. I will discuss this problem in Sec. 16.7 and show how the sensitivity may be restored by a simple, supplementary analysis. For the time being, I will concentrate on the standard analysis and return to this complication in Sec. 16.7.

16.4 STATISTICAL MODEL AND ASSUMPTIONS

In this section, we will consider the statistical model and special assumptions that underlie the analysis of the single-factor within-subjects design. As you will see, even small deviations from these assumptions complicate the interpretation of the overall F test. I will describe the statistical model first.

Linear Model

The linear model underlying the analysis of variance is usually specified by expressing the basic score Y_{ij} as a sum of a number of quantities:

$$Y_{ij} = \mu_T + \alpha_i + \pi_j + (\alpha\pi)_{ij} + \epsilon_{ij}$$

where μ_T = the overall mean of the population
α_i = the treatment effect at level a_i $(\mu_i - \mu_T)$
π_j = the subject effect for the jth subject $(\mu_j - \mu_T)$
$(\alpha\pi)_{ij}$ = the interaction of treatment and subject at $a_i s_j$ $(\mu_{ij} - \mu_i - \mu_j + \mu_T)$
ϵ_{ij} = experimental error

From this basic statement, expected values of the mean squares for the sources we normally extract in the analysis are written in terms of population variance components. The error term for evaluating the main effect of factor A is found by locating a mean square the expectation of which matches the expected value of the main effect (except for the population treatment component, of course). More specifically, the expected values for these two mean squares are

$$E(MS_A) = \sigma^2_{error} + \sigma^2_{A \times S} + n(\theta^2_A)$$

$$E(MS_{A \times S}) = \sigma^2_{error} + \sigma^2_{A \times S}$$

You will note that the expected value of the interaction mean square contains two

quantities, σ^2_{error} and $\sigma^2_{A \times S}$, the first of which refers to uncontrolled sources of variability—for example, variations in the testing conditions or measurement error—and the second of which refers to differential reactions of subjects to the treatment conditions (interaction). Although we can distinguish between these two sources of variability theoretically, we cannot disentangle their separate contributions in this particular design.[4]

The Homogeneity Assumptions

The statistical analysis of within-subjects designs operates under the same distribution assumptions required of completely randomized designs, namely, normality, homogeneity of within-treatment variances, and independence. In addition, however, certain assumptions are made concerning the correlations between the multiple measures obtained from the same subjects. I will describe these new assumptions first and then indicate the consequences when they are violated, as they often are in the behavioral sciences. Finally, I will discuss various ways to deal with these problems.

The Assumptions. Suppose we arrange the data from the $(A \times S)$ design into a set of smaller AS matrices formed by isolating pairs of treatment conditions. There would be three such matrices for our numerical example, one consisting of levels a_1 and a_2, another of levels a_1 and a_3, and a third of levels a_2 and a_3. For each pair of treatments, suppose we subtract the two scores for each of the subjects and then calculate the variances based on these three sets of difference scores. The assumption is that these three variances of differences are equal in the population. This assumption is more formally stated in terms of population within-treatment variances and of correlations between pairs of treatments and is referred to as the **sphericity assumption**.[5]

Tests of the sphericity assumption are described by Huynh and Feldt (1970), Huynh and Mandeville (1979), and Rouanet and Lépine (1970), but they are complicated and beyond the scope of this book. Some statistical computer programs provide tests of this assumption, but most of the tests have been questioned because of assumptions of their own that complicate any interpretation of their outcome (see, for example, Keselman, Rogan, Mendoza, & Breen, 1980). The safest course of action is to assume that the sphericity assumption does not hold for most experiments in the behavioral sciences and to direct your efforts instead to dealing directly with the problems resulting from these violations.

[4]The linear model underlying the analysis we have just completed is sometimes called the **nonadditive model**, which emphasizes the fact that the treatment × subject interaction is included in the equation. The **additive model**, in which this interaction is absent, is not a reasonable model for within-subjects designs in the behavioral sciences. Comprehensive presentations of the linear models normally adopted with repeated-measures designs can be found in Kirk (1982), Myers (1979), and Winer (1971).

[5]The sphericity assumption, which is also called the **circularity assumption**, has a more formal definition, which is slightly less restrictive than the one I have given. Kirk (1982, pp. 253–266) and Myers (1979, pp. 163–174) provide useful discussions of the statistical model and its underlying assumptions.

Implications of Violating the Sphericity Assumption. Violating the assumption of homogeneity of within-group variances in the completely randomized designs does not affect our evaluation of F tests unless the ratio of the largest to the smallest variance is greater than 3. Variance heterogeneity becomes more of a problem, however, with unequal sample sizes or when single-*df* comparisons are involved. We discussed problems of variance heterogeneity for experiments with equal sample sizes in Chap. 5 (pp. 98–99), for experiments with unequal sample sizes in Chap. 13 (pp. 283–284) and for single-*df* comparisons in Chap. 6 (pp. 123–128).

In contrast, even minor violations of the sphericity assumption in repeated-measures designs can seriously affect our interpretation of F ratios (see, for example, Boik, 1981). More specifically, these violations produce sampling distributions of the F ratio that are not distributed as F when the null hypothesis is true, which means that the standard F tables cannot be directly used to judge the significance of an observed F. Since it is known that when violations are present the actual sampling distribution shifts to the *right* of the central F distribution, the critical values of F obtained from Table A–1 are *too small*. That is, the actual critical values we should be using are larger than those listed in the F table.

Under these circumstances, the F test is said to be biased in a *positive* direction. It could be the case, for example, that the tabled value of F at $\alpha = .05$ actually represents a significance level that is greater than .05—for example, $\alpha = .10$. If we do not make an adjustment in our rejection procedure, we will in effect be operating at a more "lenient" significance level than we had set originally. As a consequence, we will reject the null hypothesis falsely a greater percentage of the time than our statements of significance would imply.

Correcting the Positive Bias. Several ways of solving the problem of positive bias have been proposed in the literature. One solution is to perform the usual analysis of variance but to evaluate the observed F ratios against a new critical value that for statistical convenience assumes the presence of *maximal heterogeneity*. In practice, this is accomplished easily by evaluating the F in this design with $df_{num.} = 1$ and $df_{denom.} = n - 1$, instead of $df_{num.} = a - 1$ and $df_{num.} = (a - 1)(n - 1)$. Applied to our numerical example, in which $a = 3$ and $n = 6$, we would use $F(1, 5) = 6.61$ as our critical value, rather than $F(2, 10) = 4.10$. Since the observed value of F was 4.72, we would not have declared the overall F significant if we had performed this corrective test.

This procedure is known as the **Geisser-Greenhouse correction** (Geisser & Greenhouse, 1958). It is important to note that the mean squares obtained from the analysis are still calculated with the usual *df*'s and not on these corrected ones. The corrected *df*'s are used only when we turn to the F table to find the critical value.

The main difficulty with the Geisser-Greenhouse correction is that it tends to overcorrect, reducing the type I error below the desired level. That is, the significance level may actually be $\alpha = .02$ rather than the value planned ($\alpha = .05$). Only if the heterogeneity is at its theoretical maximum will the new statistical test reflect the correct significance level. In other words, the F ratios are now biased in a *negative* direction. Thus, if we proceed in the normal fashion and use an uncorrected value of F when there is heterogeneity, the test is positively biased; if we use the correction, the test is probably negatively biased.

The Box Correction. Box (1954b) introduced a method for adjusting numerator and denominator *df*'s by a factor that reflects the degree of heterogeneity actually present in an experiment. This factor, $\hat{\epsilon}$, is estimated from the data. Examples of the calculations required to obtain the adjustment factor $\hat{\epsilon}$, which are complex but manageable, can be found in Kirk (1982, p. 262), who calls the factor $\hat{\theta}$; Myers (1978, pp. 173–174); and Winer (1971, pp. 523–524). Huynh and Feldt (1976) introduced a related correction factor, $\tilde{\epsilon}$, which they recommend should be used when $\hat{\epsilon}$ is greater than .75 (see pp. 75–76 of their article for the formulas; see also Kirk, 1982, p. 262). In either case, the procedure is the same, the $df_{num.}$ is found by multiplying df_A by one of the correction factors—that is, either $\hat{\epsilon}$ or $\tilde{\epsilon}$; the $df_{denom.}$ is found by multiplying $df_{A \times S}$ the same way. The resulting corrections, reflected by the new degrees of freedom, will not be as great as those imposed by the Geisser-Greenhouse procedure, unless maximal heterogeneity is present.

A number of methodologists recommend using a testing strategy advocated by Greenhouse and Geisser (1959), which was designed to avoid calculating $\hat{\epsilon}$ except when logically necessary:

Evaluate F at $df_{num.} = a - 1$ and $df_{denom.} = (a - 1)(n - 1)$. If the F is not significant, we retain the null hypothesis; if it is significant, we turn to the next step.
Apply the Geisser-Greenhouse correction and evaluate F at $df_{num.} = 1$ and $df_{denom.} = n - 1$. If the F is significant under this overly stringent criterion, we reject the null hypothesis; if it is not significant, we turn to the next step.
Calculate $\hat{\epsilon}$ and apply the Box correction. If the F is significant, we reject the null hypothesis; if it is not, we retain the null hypothesis.

You can also avoid these complicated calculations by using a computer. The Box correction factor $\hat{\epsilon}$ is available from a number of comprehensive statistical packages, including SPSSX, Version 4, and BMDP4V.

Missing Data

Occasionally there will be missing data; one or more Y scores sometimes will be missing for a subject (or some subjects). The least ambiguous course of action is to replace such subjects entirely and to duplicate the testing conditions for the new subject(s). But perhaps this is not feasible. Procedures are available by which missing data can be estimated from the data available in the AS matrix. These require the assumption that the data loss is unrelated to the differences in the treatment conditions. Kirk (1982, pp. 268–270) and Myers (1979, pp. 177–178) discuss methods for estimating missing data under these circumstances.

16.5 EFFECT SIZE AND POWER

In my discussions of the completely randomized designs, I stressed the value of estimating treatment effects and of using power estimates to help us choose an

appropriate sample size to use in a proposed experiment. The arguments hold true for within-subjects designs as well, but certain complications arise. I will consider the problems associated with estimating effect size first.

Estimating Treatment Effects (Omega Squared)

There are two major problems in estimating omega squared in the within-subjects design. First, there is the definition itself. The standard treatments of the concept define omega squared as

$$\omega_A^2 = \frac{\sigma_A^2}{\sigma_T^2}$$

where σ_T^2 is the sum of the relevant variance components. The problem lies in defining these relevant components. In the completely randomized single-factor design, there were two components, σ_A^2 and σ_{error}^2 and no ambiguity. In the completely randomized factorial design, however, there were two possibilities for defining σ_T^2, namely, $\sigma_A^2 + \sigma_{error}^2$ or $\sigma_A^2 + \sigma_{error}^2$ plus the variance components associated with all other main effects and interactions. I argued in favor of the first definition, which defines a *partial* omega squared, because it gives the same meaning to the concept regardless of the nature of the design (see pp. 223–224). Because the $(A \times S)$ is a factorial design with respect to subjects, there are again two ways of defining σ_T^2, specifically,

$$\sigma_T^2 = \sigma_A^2 + \sigma_{error}^2 \text{ and } \sigma_T^2 = \sigma_A^2 + \sigma_S^2 + \sigma_{error}^2$$

The two definitions will give different values for ω_A^2.

The second problem concerns estimating the different variance components from an experiment. Vaughan and Corballis (1969), for example, give a formula that admittedly overestimates σ_T^2 and, consequently, underestimates ω_A^2. Myers (1979, pp. 178–179) offers two estimates of omega squared, which together set a range of values within which the "actual" ω_A^2 will fall.

I have argued that estimates of relative treatment provide useful information that supplements the actual significance test. But because of these two theoretical uncertainties, however—the definition of σ_T^2 and the estimates of the components themselves—any recommendation must be provisional. Until there is more work on these problems and methodologists develop a consensus concerning the definition of a more useful index, I offer a definition that is continuous with the concept of partial omega squared:

$$\hat{\sigma}_A^2 = \frac{df_A(MS_A - MS_{A \times S})}{(a)(n)} \quad (16\text{–}1)$$

$$\hat{\sigma}_{error}^2 = MS_{A \times S} \quad (16\text{–}2)$$

$$\hat{\omega}_A^2 = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}_{error}^2} \quad (16\text{–}3)$$

Using Power Estimates to Choose Sample Size

In previous chapters, you have seen why power estimates are vital in the realistic planning of completely randomized designs. The same arguments hold for the within-subjects designs as well. Again you will need to estimate the minimum treatment effects you wish to detect and the error variance you expect to be present. You then combine these estimates with other information about your experiment to calculate the statistic ϕ_A^2, which is then translated into an estimate of power. For the $(A \times S)$ design,

$$\phi_A^2 = n' \frac{\Sigma (\text{dev.})^2}{(a)(\sigma_{error}^2)} \quad (16\text{–}4)$$

where n' = the trial sample size
$\Sigma (\text{dev.})^2$ = the basic population deviations $(\mu_i - \mu_T)$
a = the number of treatment conditions
σ_{error}^2 = the relevant population error variance

The last quantity, σ_{error}^2, may be estimated from a pilot study or from previous research you have conducted or has been reported in the literature.

At this point, you turn to the Pearson-Hartley power charts (Table A–2) and fiddle with different trial sample sizes (n') until you achieve the power you want for your proposed study. If you need to counterbalance the orders of the treatments, you would select a sample size that is some multiple of the number of treatment conditions (a). For our numerical example, in which $a = 3$, our sample size would be some multiple of 3—that is, 3, 6, 9, and so on. A nonmultiple will not permit the counterbalancing procedure to work properly.

Violating Homogeneity Assumptions. The procedures I have outlined for estimating power and setting sample size assume that the sphericity assumption, which we discussed in Sec. 16.4, is upheld. In most cases, the data will depart from the form specified by the assumptions underlying the statistical model, which will have a direct consequence on power, particularly if we will apply some sort of correction to reduce the positive bias associated with the standard F test. This correction is achieved by making it more difficult to reject the null hypothesis, which is what is intended when the focus is type I error but which has an opposite effect when the focus is type II error. That is, power will be reduced when any of these correction techniques is applied. Possible solutions to this problem are only beginning to be proposed (see, for example, Muller & Barton, 1989), leaving us in an ambiguous position with regard to planning sample size.

What we need is a convenient and realistic way to estimate the sample size necessary for a within-subjects experiment that will not satisfy the sphericity assumption. I propose we use the Geisser-Greenhouse correction in the power determinations. The only change necessary is in the selection of power charts and the specific power function. Assuming perfect sphericity, we would use the chart for $df_{num.} = a - 1$ and $df_{denom.} = (a - 1)(n - 1)$; this is the procedure I outlined previously. If it seems likely that the sphericity assumption will not be met and that

we will probably use the Geisser-Greenhouse correction in analyzing the data, the chart we should now use is one that takes that correction into account, namely, $df_{num.} = 1$ and $df_{denom.} = n - 1$. If we have any way of obtaining a realistic estimate of $\hat{\epsilon}$, which seems unlikely in most practical situations, we would use the power chart appropriate to that new combination of df's.

16.6 COMPARISONS INVOLVING THE TREATMENT MEANS

All of the types of comparisons discussed in Part II for the completely randomized single-factor design are available for the corresponding within-subjects design. There is one important difference, however, and this again lies in the selection of the error term.

Error Terms for Comparisons

You have seen that the mean square used to test the main effect of factor A, $MS_{A \times S}$, is influenced by two components: experimental error and the treatment $\times$ subject interaction. In our evaluation of the overall treatment effect, in which all treatment means are compared, it makes intuitive sense to use an error term based on all the scores in the experiment. However, this overall interaction mean square is generally *not* appropriate for evaluating the significance of individual comparisons. The $MS_{A \times S}$ is an *average* of a set of individual *comparison* $\times$ subject interactions and, as such, may not provide an appropriate estimate of the specific interaction reflecting itself in the particular set of treatment means we are considering. Let's consider an example from an actual experiment, consisting of $a = 6$ treatments and $n = 8$ subjects.[6] The overall error term was $MS_{A \times S} = 2.79$, and the separate error terms for five single-df comparisons were 1.48, 3.90, 1.03, 2.51, and 5.05—a range of nearly 5 to 1. If the overall error term had been used to evaluate these comparisons, the resultant F's would have been too small for comparisons 1 and 3, too large for comparisons 2 and 5, and about right for comparison 4.

Research has shown that even minor violations of the sphericity assumption can produce sizable differences among separate error terms (see, for example, Keselman, Rogan, & Games, 1981). The safest strategy is to construct separate error terms for *all* comparisons. The computational procedures we will follow in this section have been specifically devised to facilitate the calculation of the individual error terms needed for within-subjects designs. Consequently, the analysis differs considerably from that presented in Chap. 6 for the completely randomized design, which worked directly with the treatment means. The approach taken here conducts what amounts to a within-subjects analysis on the information relevant for the single-df comparison; we will work with scores and sums in a specialized

[6]Keppel, Postman, and Zavortink (1968).

matrix I will call an ***AS* comparison matrix**. We can then with minor modifications employ the formulas for the overall within-subjects analysis.

Forming the *AS* Comparison Matrix

The key to the analysis is the AS comparison matrix, which captures the information relevant to the specific comparison under consideration. For a pairwise comparison, the procedure is simple: We extract the two relevant scores for each subject and place them directly in the new matrix. For a complex comparison, the procedure involves an intermediate step in which each Y score is weighted by the relevant coefficient before being placed in the comparison matrix. This procedure is best illustrated with an example, which we will base on the data from the earlier numerical example.

Suppose we wanted to compare the mean at level a_2 with the mean of the other two levels combined (a_1 and a_3). The coefficients I will use are the set c_j: $\frac{1}{2}$, -1, $\frac{1}{2}$, although any appropriate set will do (the set 1, -2, 1, for example, will avoid decimal numbers). The matrix on the left side of Table 16–5 contains the Y scores weighted by the appropriate coefficients. To illustrate the procedure for the first subject, we take the three original Y scores for this subject, which were 8, 12, and 9 (see Table 16–3, p. 348); when weighted by the coefficients, the three scores become $(+\frac{1}{2})(8) = +4.0$; $(-1)(12) = -12$; and $(+\frac{1}{2})(9) = +4.5$, respectively. The scores for the other subjects are constructed in the same way. We now begin to construct the AS comparison matrix by adding together for each subject the scores weighted by a *positive* coefficient and entering them in column 1 of the comparison matrix, which appears on the right side of Table 16–5; for s_1, we would place 8.5 (4.0 + 4.5) in the first column. The weighted sums in this column represent the "positive" part of the single-df comparison, which I have labeled $a_{(+)}$. The sums of the scores weighted by a *negative* coefficient are next entered in column 2 of the comparison matrix. (We delete the negative signs at this point as they are not relevant for the rest of the analysis.) Since there is only one negative coefficient for this comparison, the "sum" for s_1 is 12. The sums appearing in this column are labeled $a_{(-)}$.

From this point on, we treat this comparison matrix exactly as if it were an

Table 16–5 Constructing an *AS* Comparison Matrix

	Weighted Y Scores		
	a_1	a_2	a_3
s_1	+4.0	−12	+4.5
s_2	+4.0	−13	+7.0
s_3	+4.5	−15	+3.0
s_4	+0.0	−18	+6.0
s_5	+6.5	−15	+9.0
s_6	+5.0	−17	+3.5

	AS Comparison Matrix		
	$a_{(+)}$	$a_{(-)}$	Sum
s_1	8.5	12.0	20.5
s_2	11.0	13.0	24.0
s_3	7.5	15.0	22.5
s_4	6.0	18.0	24.0
s_5	15.5	15.0	30.5
s_6	8.5	17.0	25.5
Sum	57.0	90.0	147.0

partial omega squared since power may have been low in this example. Using Eq. (19–6), we find

$$\hat{\omega}^2_{A\times B\times C} = \frac{df_{A\times B\times C}(MS_{A\times B\times C} - MS_{S/ABC})}{(a)(b)(c)(n)}$$

$$= \frac{2(4.32 - 1.75)}{(3)(2)(2)(5)} = \frac{5.14}{60} = .0857$$

and from Eq. (19–7) we have

$$\hat{\sigma}^2_{S/ABC} = MS_{S/ABC} = 1.75$$

Substituting in Eq. (19–8), we obtain

$$\hat{\omega}^2_{A\times B\times C} = \frac{\hat{\sigma}^2_{A\times B\times C}}{\hat{\sigma}^2_{A\times B\times C} + \hat{\sigma}^2_{S/ABC}} = \frac{.0857}{.0857 + 1.75} = .0467$$

This means that the three-way interaction produces an effect somewhere between "small" ($\hat{\omega}^2_{effect} = .01$) and "medium" ($\hat{\omega}^2_{effect} = .06$) and suggests that a three-way interaction may be present and that our experiment simply had insufficient power to detect it.[5]

19.6 USING POWER TO DETERMINE SAMPLE SIZE

Power determinations, as you will recall, provide a rational way of setting sample size (n). We considered procedures for determining sample size in Chap. 4 (pp. 76–80) for the single-factor design and in Chap. 10 (pp. 224–229) for the two-way factorial. These procedures may be adapted for the three-factor design. We will consider three possible approaches.

Using Population Deviations

The first procedure works with population means and deviations derived from them, which we will then use to estimate ϕ. We then use this quantity, you may recall, to determine power from the power charts for different trial values for sample size, which I call n'. The general formula for this estimate is given by

$$\phi^2_{effect} = \frac{(\text{no. obsn.})[\Sigma\,(\text{dev.})^2]}{(df_{effect} + 1)(\sigma^2_{error})} \qquad (19\text{–}9)$$

where (no. obsn.) = the number of observations that will contribute to each basic deviation
$\Sigma\,(\text{dev.})^2$ = the basic population deviations constituting the treatment effect in question

[5]The standard estimated omega squared is .0224, less than half the size of the value we obtained with the formula for partial estimated omega squared.

df_{effect} = the df associated with the treatment effects, calculated in the usual fashion
σ^2_{error} = the population error variance

The biggest stumbling block, of course, will be estimating the relevant population means from which the deviations are derived. If you are able to make realistic guesses for the population means and can estimate the population error variance, you can calculate the relevant deviations from formulas presented in Winer (1971, pp. 335–340).[6] I illustrated this procedure for the two-way interaction in Chap. 10 (pp. 226–227). After this point, you simply follow the trial-and-error procedure described in the earlier chapters for determining sample size.

Using a Pilot Study

A realistic alternative is to base your estimate of sample size on a pilot study that you wish to replicate. We can base our estimate on a formula presented in Chap. 10 (p. 228). With the three-way interaction in mind,

$$\phi^2_{A\times B\times C} = \left(\frac{n'}{n_{old}}\right)\left(\frac{df_{A\times B\times C}}{df_{A\times B\times C} + 1}\right)(F_{A\times B\times C} - 1) \qquad (19\text{–}10)$$

where n' = the trial sample size
n_{old} = the sample size from the pilot study
$df_{A\times B\times C}$ = the df associated with the three-way interaction
$F_{A\times B\times C}$ = the F associated with that effect in the pilot study

Since we were particularly interested in the three-way interaction, which was present but not significant in the overall analysis, we might consider repeating the experiment with a new set of subjects. We can use the information from our first study to help us to determine an appropriate sample size for this replication.

To illustrate, we can use the information in Table 19–8 to calculate $\phi^2_{A\times B\times C}$. Substituting in Eq. (19–10), we find

$$\phi^2_{A\times B\times C} = \left(\frac{n'}{5}\right)\left(\frac{2}{2+1}\right)(2.47 - 1) = \frac{(n')(2.94)}{15} = .1960\,n' \qquad (19\text{–}11)$$

Suppose we want to set power at .80. To start the process with a reasonable trial sample size, we can rearrange Eq. (19–11) in terms of n'; that is,

$$n' = \frac{\phi^2_{A\times B\times C}}{.1960} \qquad (19\text{–}12)$$

[6]The deviation for the three-way interaction, for example, is given by

$$(\alpha\beta\gamma)_{ijk} = \mu_{ijk} - \mu_{ij} - \mu_{ik} - \mu_{jk} + \mu_i + \mu_j + \mu_k - \mu_T$$

where μ_{ijk} = a given cell mean
μ_{ij}, μ_{ik}, and μ_{jk} = the relevant marginal means from the two-way matrices
μ_i, μ_j, and μ_k = the relevant means reflecting the main effects
μ_T = the grand mean

and then use the Pearson-Hartley power charts to determine a reasonable value for ϕ. We obtain our first trial sample size by substituting this ϕ into Eq. (19–12) and solving for n'.

Let's see how this is done. We turn to the second power chart in Table A–4 ($df_{num.} = 2$) and find the ϕ associated with a = .05, power = .80, and $df_{denom.} = 60$. This value is $\phi = 1.82$. Entering this value in Eq. (19–12), we find

$$n' = \frac{(1.82)^2}{.1960} = 16.9$$

We will use 17 for our first trial sample size. Substituting this value into Eq. (19–11), we find

$$\phi^2_{A \times B \times C} = (.1960)(17) = 3.332$$

$$\phi_{A \times B \times C} = \sqrt{3.332} = 1.825$$

We now need to determine the power associated with $\phi = 1.825$. Returning to the power chart, but using the power function at dfdenom. = ∞ because the df for this sample size would be $df_{S/ABC} = 192$, we find power slightly higher than .80. A sample size of $n = 16$ should be about right.

Using Partial Omega Squared

Cohen (1977) presents a method for determining sample size that is based on some estimate of relative treatment magnitude (pp. 396–400). Suppose we wanted to detect a three-way interaction in this example that has a partial omega squared of at least $\omega^2 = .06$ ("medium"). In order to use Cohen's special tables we need to calculate f, where

$$f^2 = \frac{\omega^2}{1 - \omega^2} = \frac{.06}{1 - .06} = .0638$$

$$f = \sqrt{.0638} = .253$$

With this information and Cohen's tables, I find $n = 13.75$, or 14.

19.7 EXERCISES[7]

1. Following are the outcomes of 10 three-way factorial experiments. The design in each example is the same—a 2 × 2 × 2 factorial. The main intent of this problem is to test your ability to identify three-way interactions in a set of data. Indicate the presence or absence of a triple interaction for each example. (Please assume for this problem that the means are "error-free.") When the three-way interaction is not present, indicate which, if any, of the two-way interactions are present. Finally, are there any main effects that may be interpreted unambiguously in any of these examples?

[7]The answers to these problems are found in Appendix B.

	TREATMENT CONDITIONS							
EXAMPLES	a_1 b_1 c_1	a_1 b_1 c_2	a_1 b_2 c_1	a_1 b_2 c_2	a_2 b_1 c_1	a_2 b_1 c_2	a_2 b_2 c_1	a_2 b_2 c_2
1	1	1	1	1	3	3	3	3
2	2	2	1	1	3	3	2	2
3	3	2	2	1	4	3	3	2
4	1	1	3	3	3	3	1	1
5	1	2	2	3	2	3	1	2
6	2	0	1	1	4	2	3	3
7	2	4	1	3	0	3	1	4
8	2	3	1	4	0	4	1	3
9	2	2	1	1	4	4	3	4
10	1	0	2	3	2	3	1	0

2. Consider the following results of a 3 × 3 × 2 factorial experiment in which $n = 4$ subjects are randomly assigned to each of the treatment conditions:

TREATMENT CONDITIONS																		
a_1 b_1 c_1	a_1 b_1 c_2	a_1 b_2 c_1	a_1 b_2 c_2	a_1 b_3 c_1	a_1 b_3 c_2	a_2 b_1 c_1	a_2 b_1 c_2	a_2 b_2 c_1	a_2 b_2 c_2	a_2 b_3 c_1	a_2 b_3 c_2	a_3 b_1 c_1	a_3 b_1 c_2	a_3 b_2 c_1	a_3 b_2 c_2	a_3 b_3 c_1	a_3 b_3 c_2
7	7	2	2	4	1	10	6	4	1	7	1	13	12	9	7	8	7
4	5	4	3	3	3	7	5	6	3	4	3	10	13	8	6	5	7
5	5	3	4	0	2	6	5	3	4	5	3	13	11	9	7	6	4
6	6	3	1	3	2	8	6	5	5	5	0	8	12	10	6	6	6

   a. Conduct an analysis of variance on these data. Save your calculations for problem 1 of the exercises for Chap. 20.
   b. Given the outcome of the analysis, to what sources of variance would you now give close attention?

3. Suppose a 4 × 3 × 2 factorial experiment is conducted. The sums for the treatment conditions, which are based on $n = 5$ subjects, follow.

	c_1				c_2			
	a_1	a_2	a_3	a_4	a_1	a_2	a_3	a_4
b_1	35	46	49	55	37	49	56	65
b_2	42	51	55	62	40	38	44	42
b_3	29	40	45	49	27	23	19	14

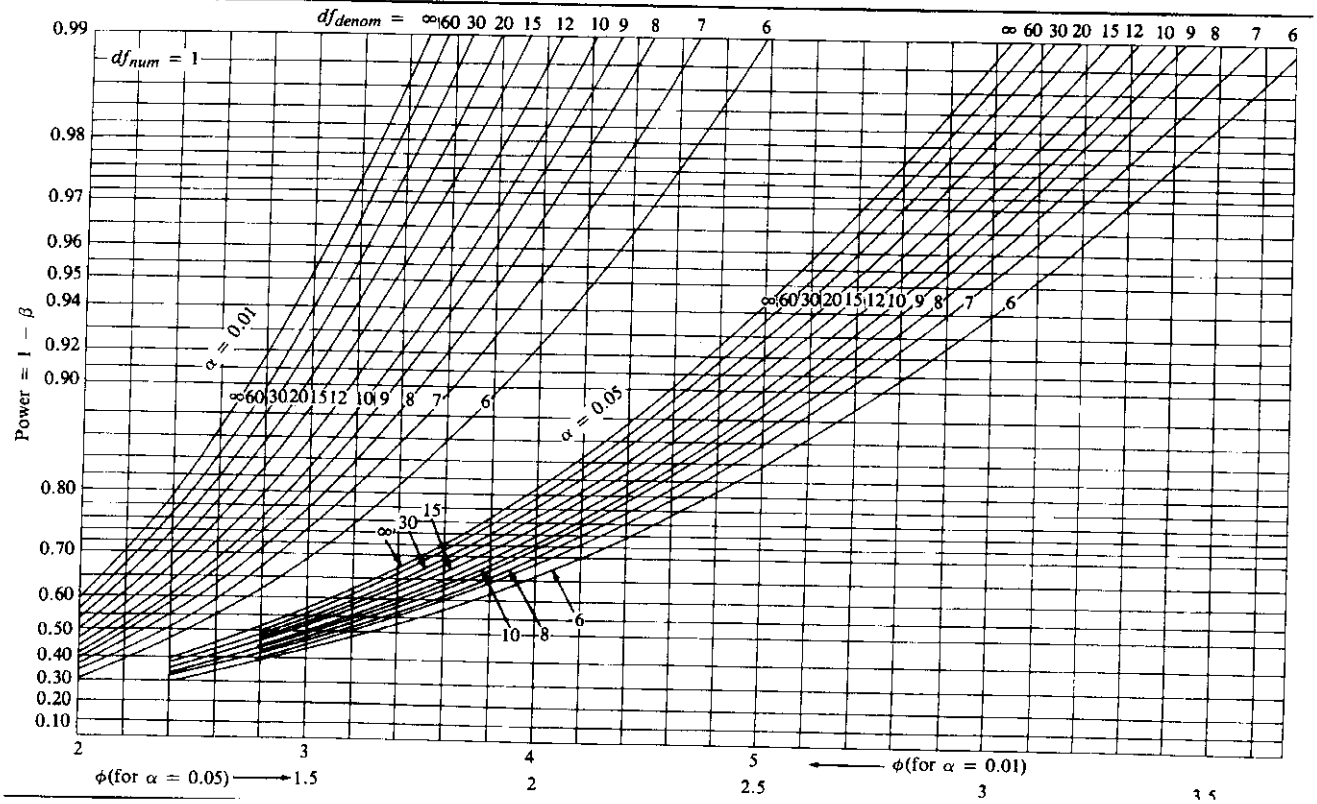
   a. Conduct an analysis of variance on these data. (Assume $MS_{S/ABC} = 1.86$.) Save your calculations for problem 2 of the exercises for Chap. 20.
   b. Given the outcome of the analysis, to what sources of variance would you now give close attention?

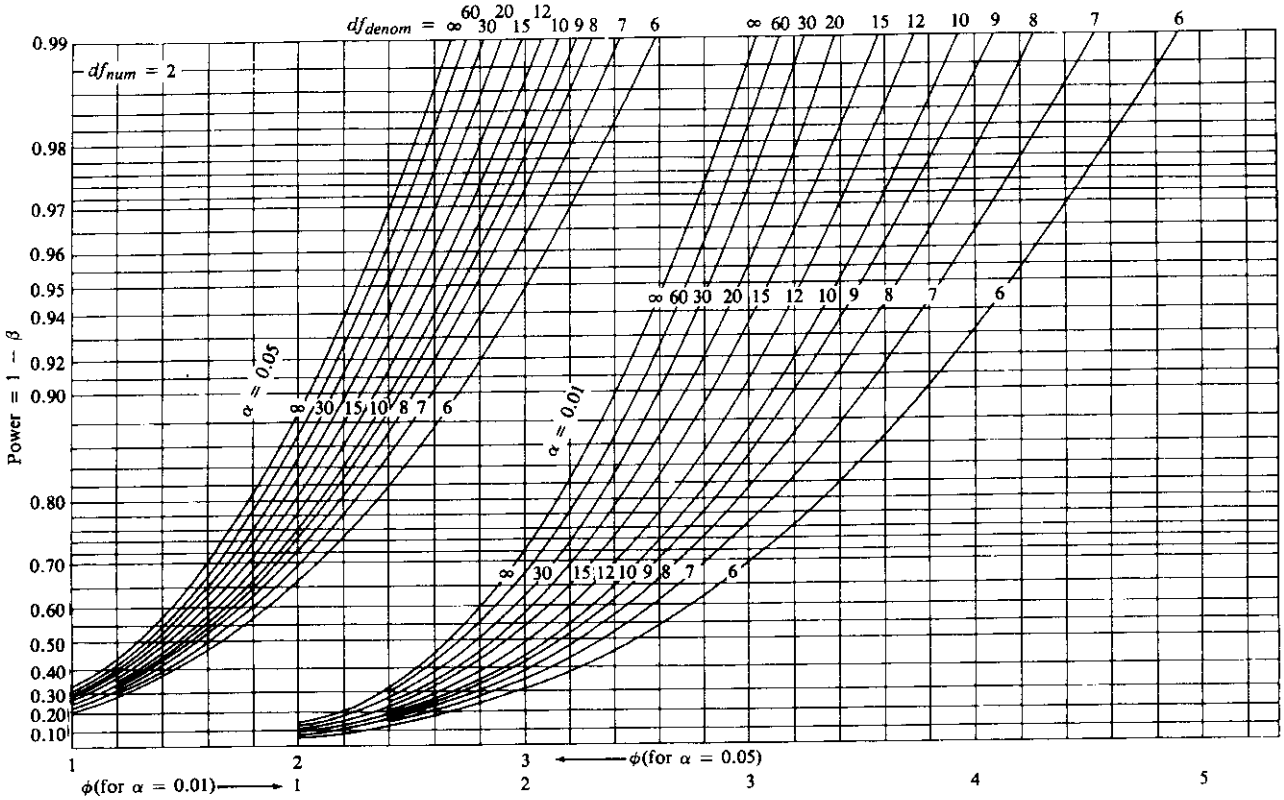
Table A–1 (*Cont.*)

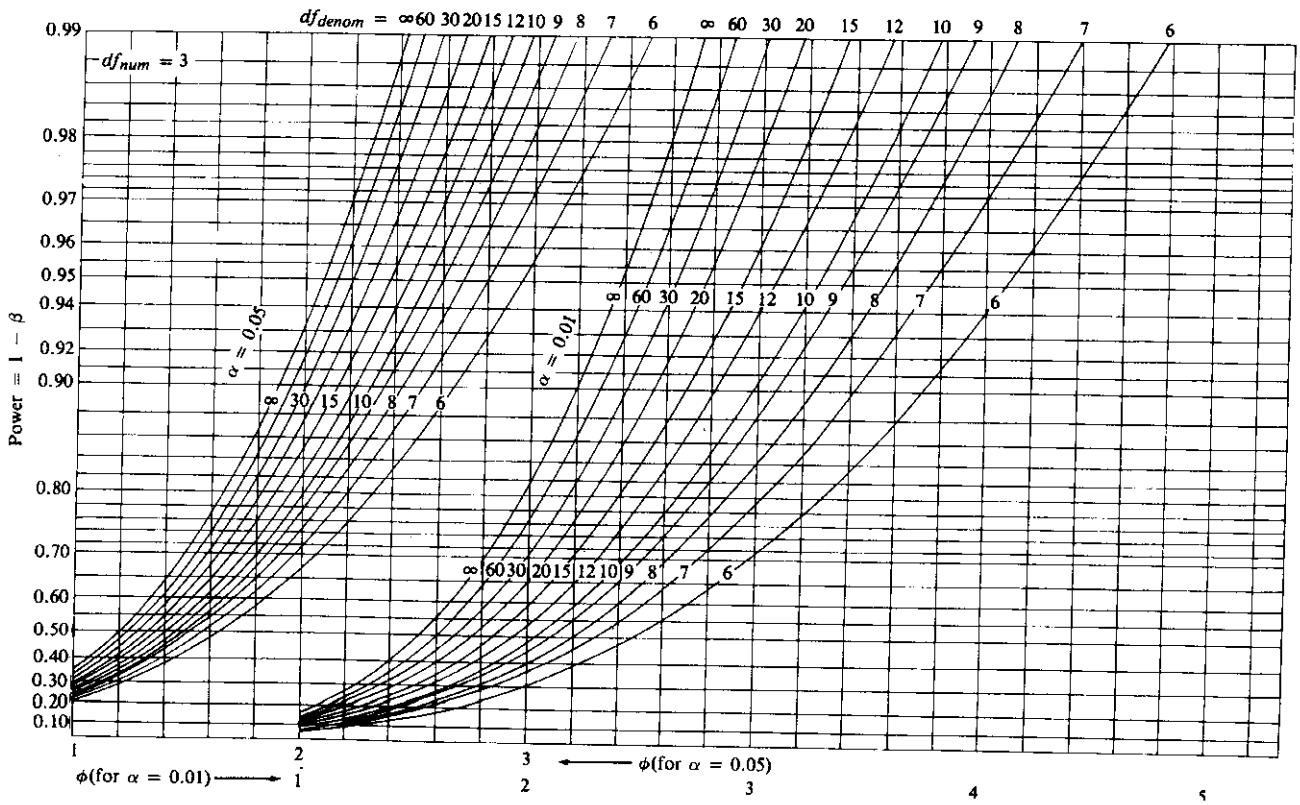
df FOR DENOM.	α	*df* FOR NUMERATOR																	
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	∞
120	.25	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28	1.26	1.24	1.22	1.21	1.19	1.18	1.16	1.10
	.10	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.19
	.05	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.25
	.025	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.31
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.38
	.001	11.4	7.32	5.79	4.95	4.42	4.04	3.77	3.55	3.38	3.24	3.02	2.78	2.53	2.40	2.26	2.11	1.95	1.54
∞	.25	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25	1.24	1.22	1.19	1.18	1.16	1.14	1.12	1.00
	.10	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.00
	.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.00
	.025	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.00
	.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.00
	.001	10.8	6.91	5.42	4.62	4.10	3.74	3.47	3.27	3.10	2.96	2.74	2.51	2.27	2.13	1.99	1.84	1.66	1.00

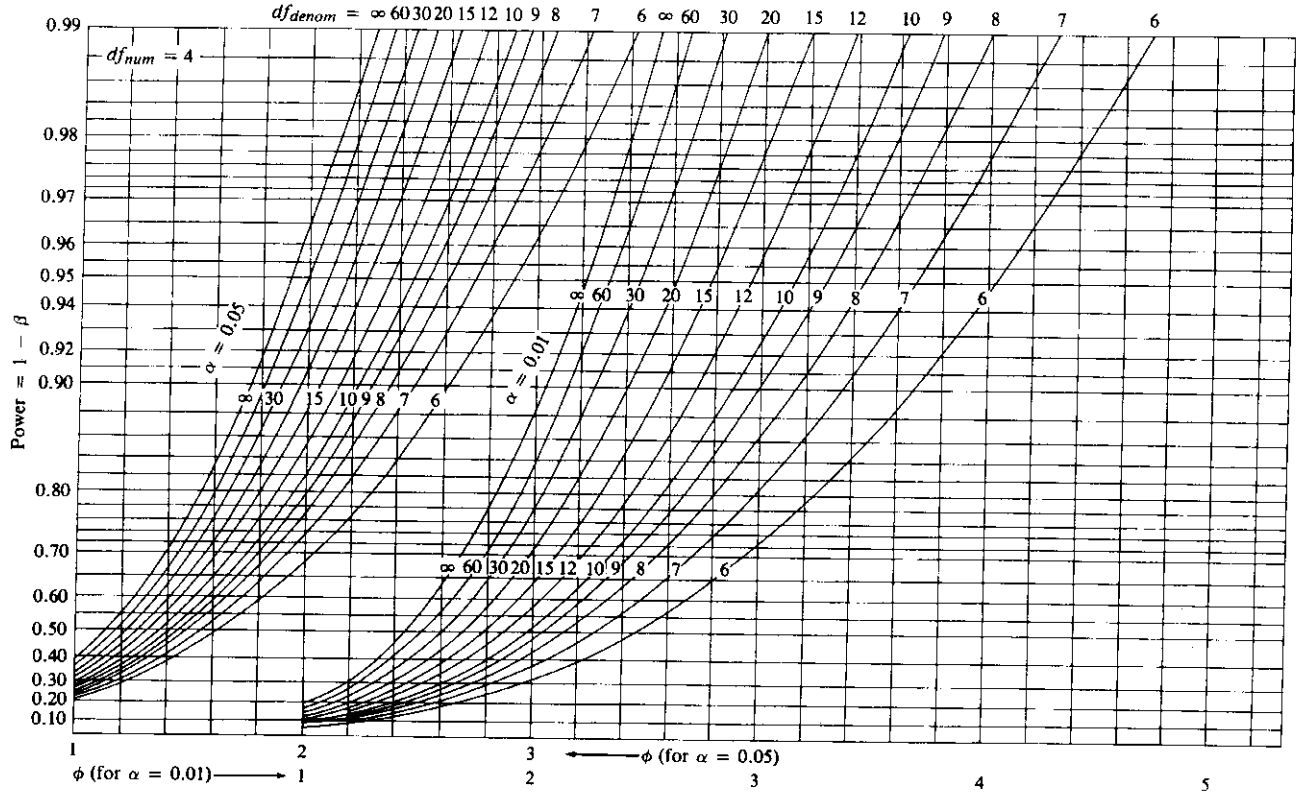
This table is abridged from Table 18 in E. S. Pearson and H. O. Hartley (Eds.), *Biometrika tables for statisticians* (3rd ed., Vol. 1), Cambridge University Press, New York, 1970, by permission of the *Biometrika* Trustees.

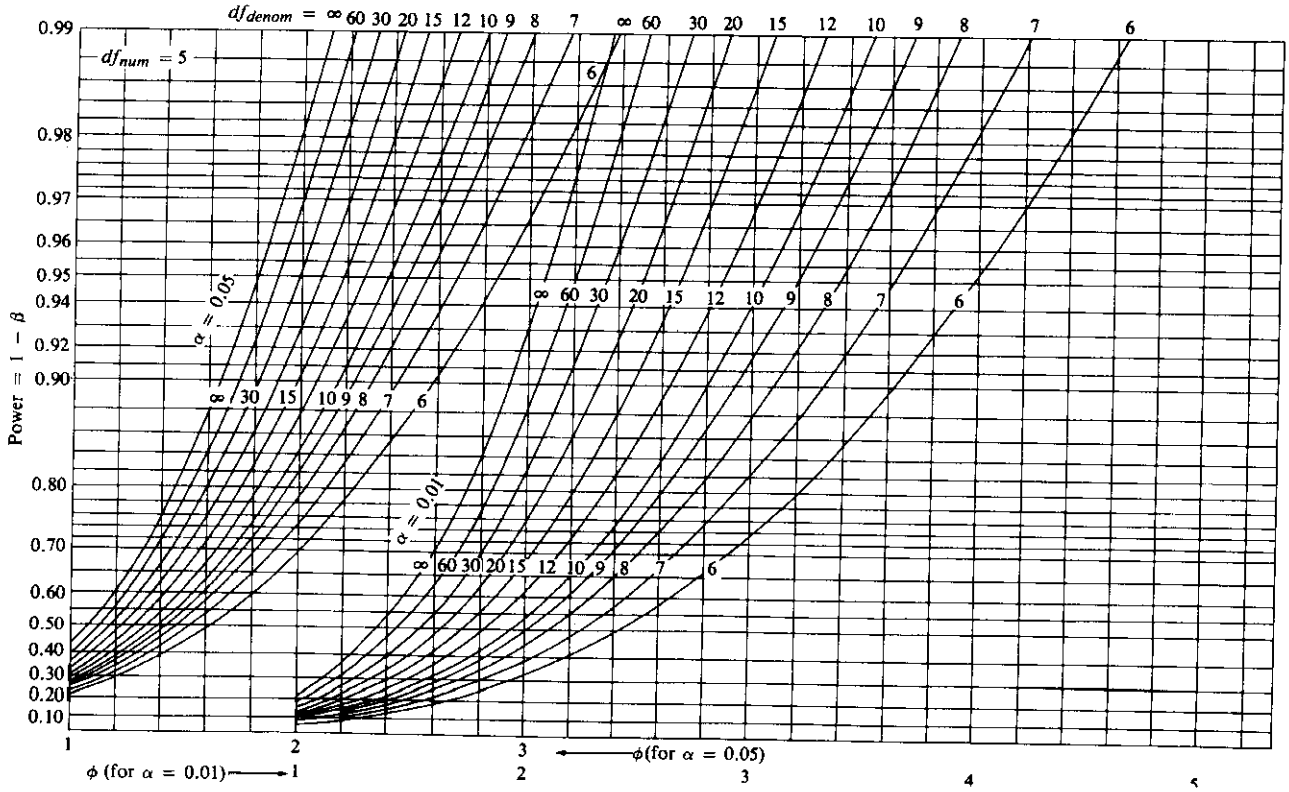
Table A–2 Power Function for Analysis of Variance (Fixed-Effects Model

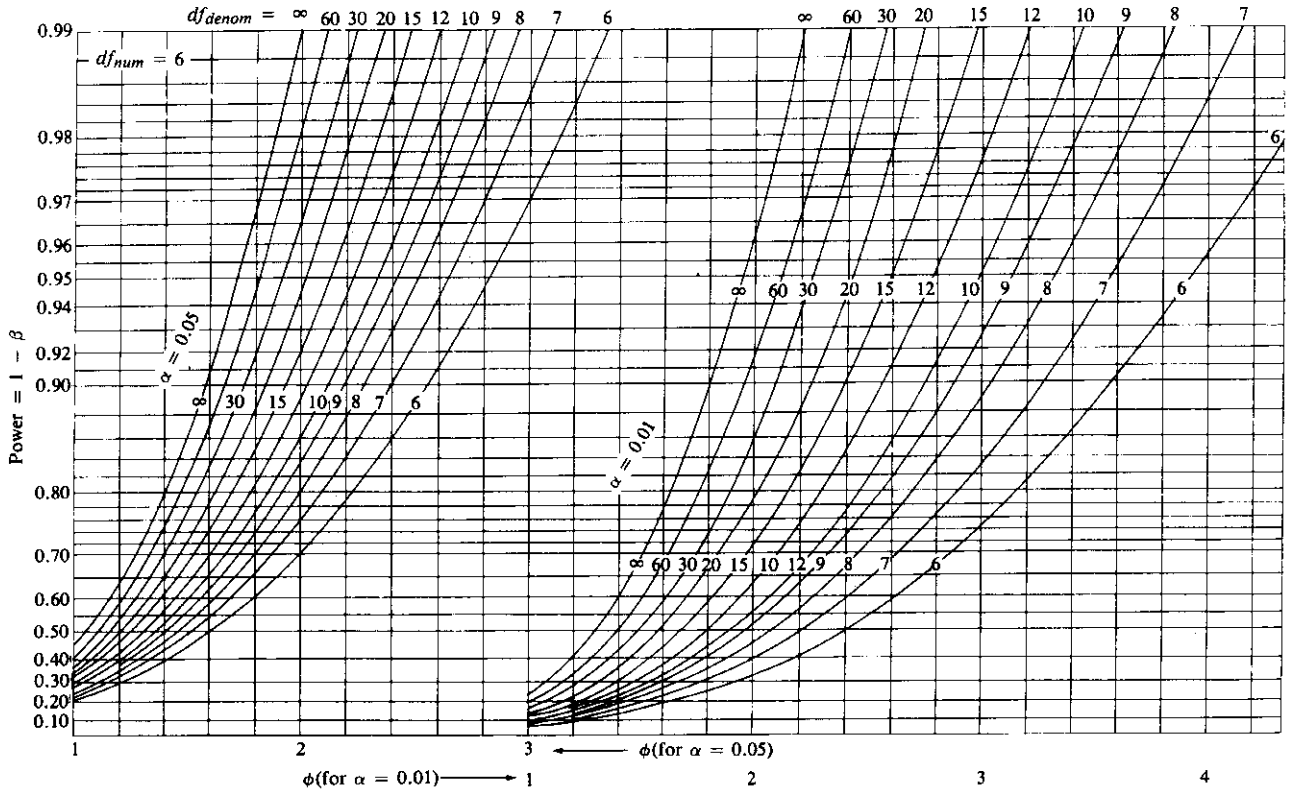


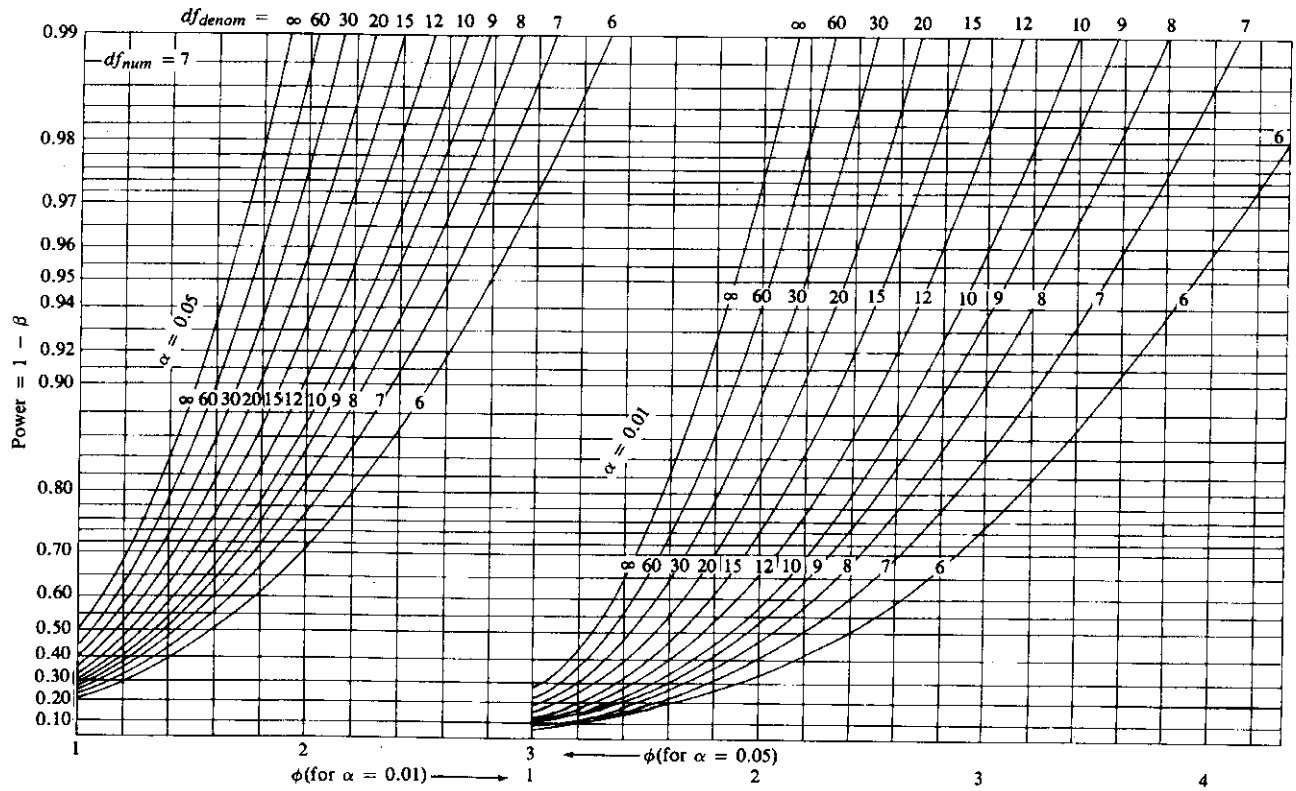

df_{denom} = ∞ 60 30 20 15 12 10 9 8 7 6
df_{num} = 2
Power = 1 − β
α = 0.05
α = 0.01
φ(for α = 0.01)
φ(for α = 0.05)

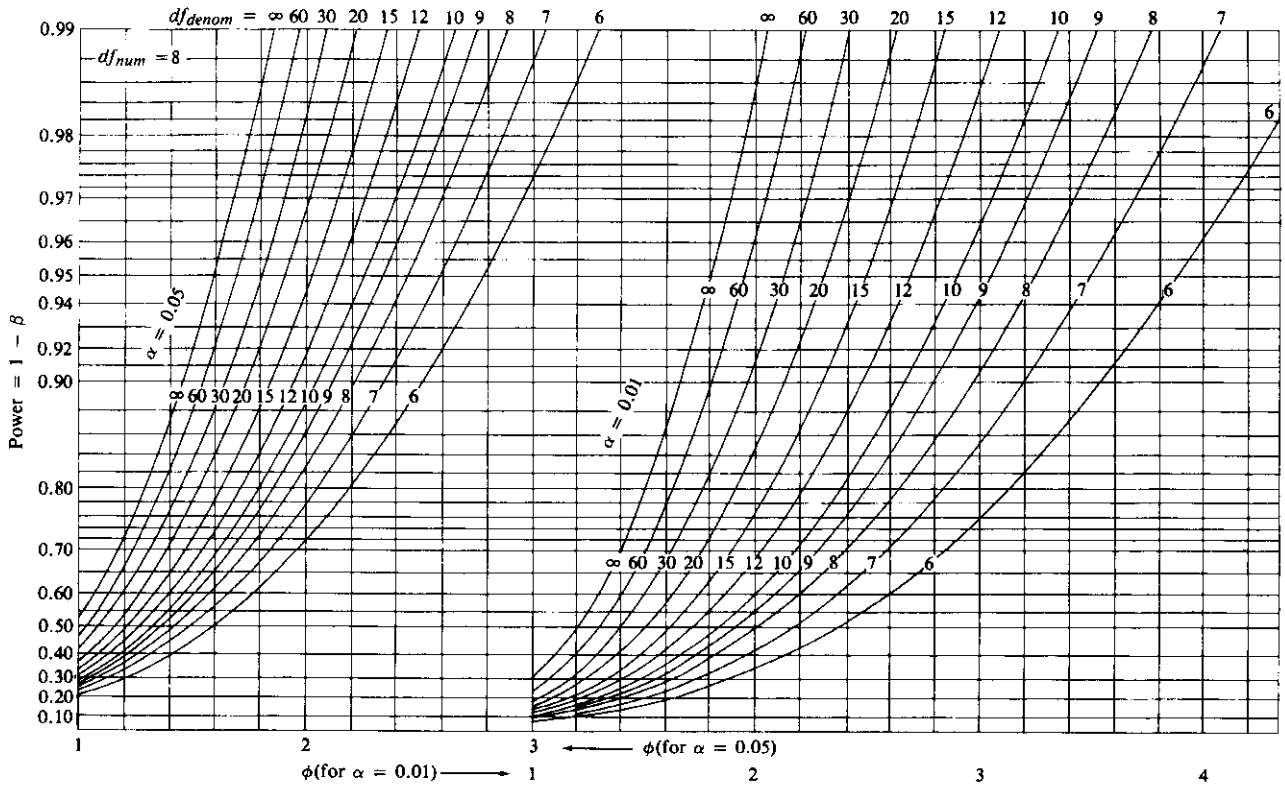

df_{denom} = ∞ 60 30 20 15 12 10 9 8 7 6
df_{num} = 3
Power = 1 − β
α = 0.05
α = 0.01
φ(for α = 0.01)
φ(for α = 0.05)

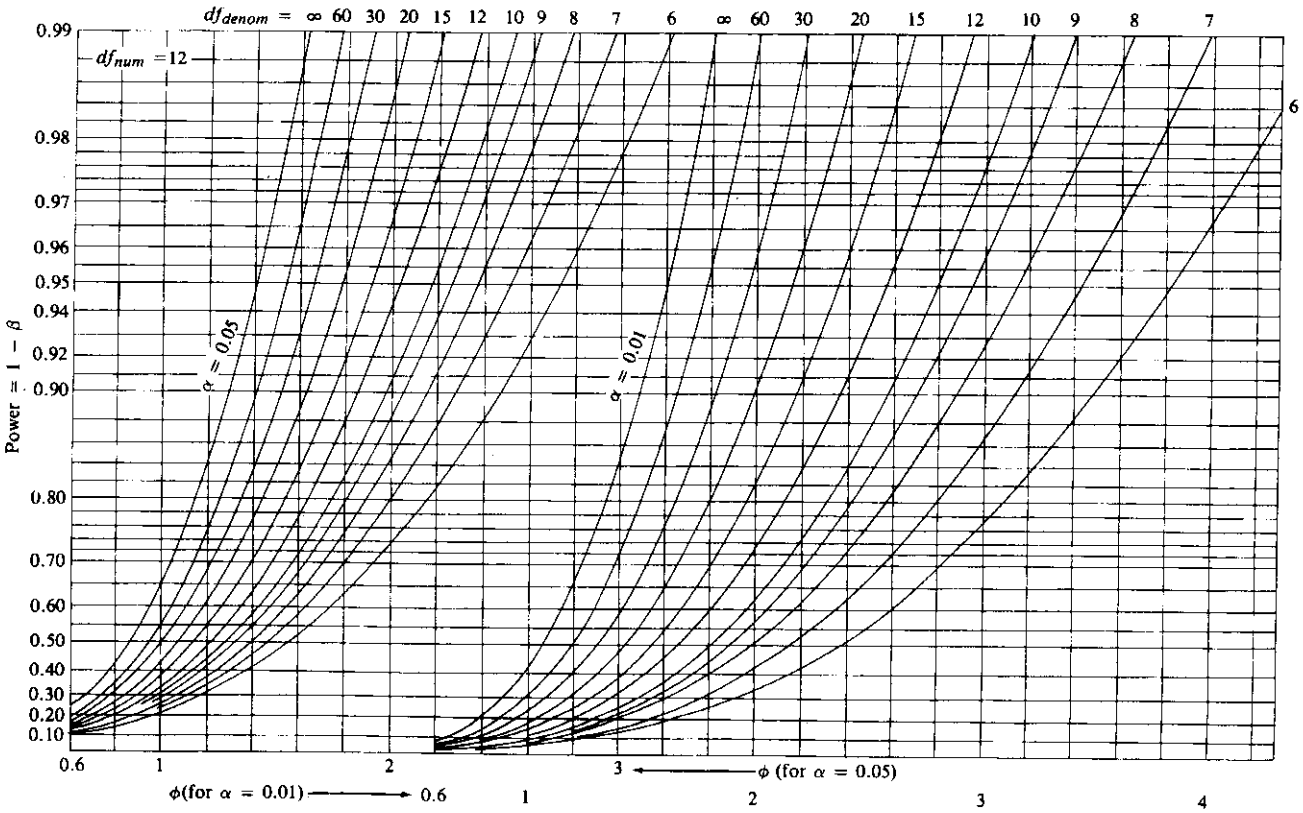

df_{denom} = ∞ 60 30 20 15 12 10 9 8 7 6
df_{num} = 4
Power = 1 − β
α = 0.05
α = 0.01
0.99
0.98
0.97
0.96
0.95
0.94
0.92
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
φ (for α = 0.01)
φ (for α = 0.05)
1
2
3
4
5


df_{denom} = ∞ 60 30 20 15 12 10 9 8 7 6
df_{num} = 5
Power = 1 − β
α = 0.05
α = 0.01
0.99
0.98
0.97
0.96
0.95
0.94
0.92
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
φ (for α = 0.01)
φ(for α = 0.05)
1
2
3
4


df_{denom} = ∞ 60 30 20 15 12 10 9 8 7 6
df_{num} = 6
Power = 1 − β
α = 0.05
α = 0.01
0.99
0.98
0.97
0.96
0.95
0.94
0.92
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
1 2 3
φ(for α = 0.05)
φ(for α = 0.01)
1 2 3 4


df_{denom} = ∞ 60 30 20 15 12 10 9 8 7 6
df_{num} = 7
Power = 1 − β
α = 0.05
α = 0.01
0.99
0.98
0.97
0.96
0.95
0.94
0.92
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
1 2 3
φ(for α = 0.05)
φ(for α = 0.01)
1 2 3 4


df_{denom} = ∞ 60 30 20 15 12 10 9 8 7 6
df_{num} = 8
Power = 1 − β
α = 0.05
α = 0.01
0.99
0.98
0.97
0.96
0.95
0.94
0.92
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
1
2
3
4
φ(for α = 0.01) ⟶
⟵ φ(for α = 0.05)


df_{denom} = ∞ 60 30 20 15 12 10 9 8 7 6
df_{num} = 12
Power = 1 − β
α = 0.05
α = 0.01
0.99
0.98
0.97
0.96
0.95
0.94
0.92
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.6
1
2
3
4
φ(for α = 0.01) ⟶
⟵ φ (for α = 0.05)

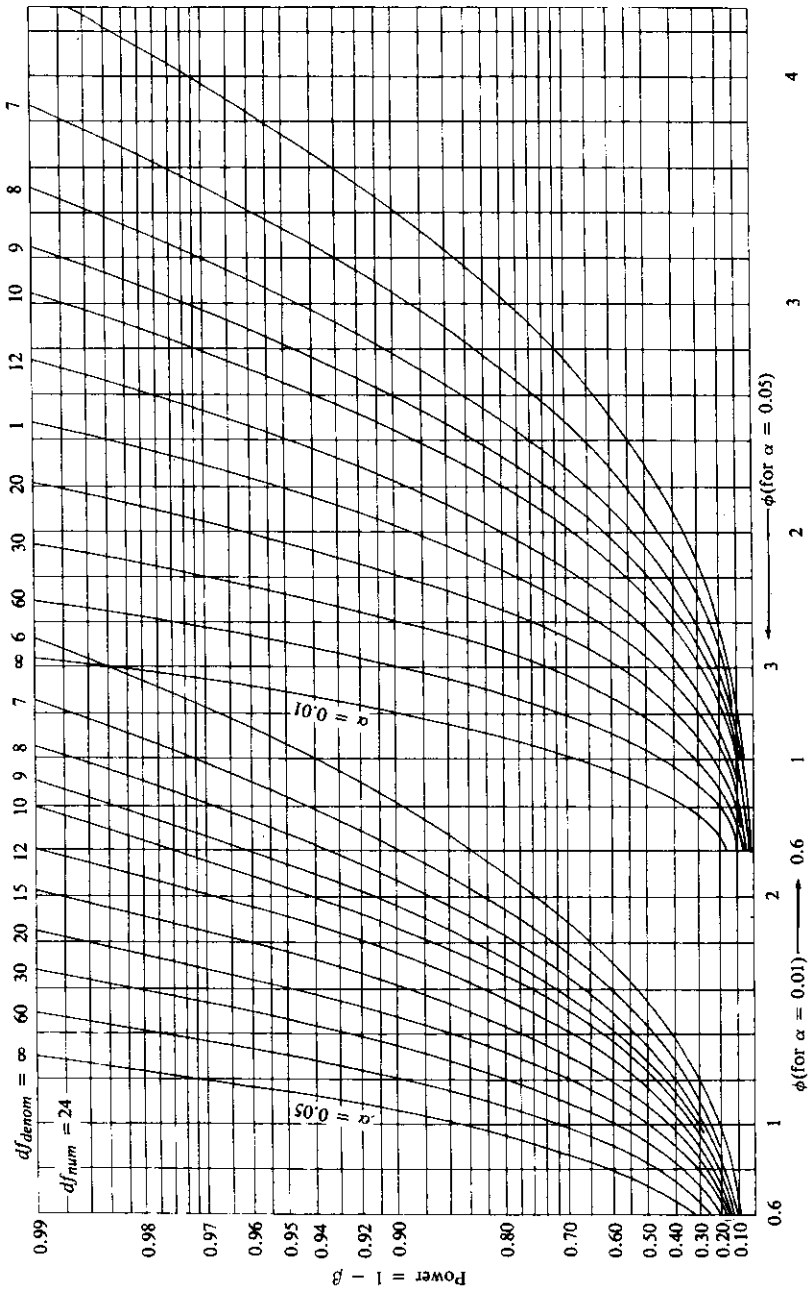


Table A-3 Selected Values from the *t* Distribution

df	α = .05	α = .01	*df*	α = .05	α = .01
1	12.71	63.66	18	2.10	2.88
2	4.30	9.92	19	2.09	2.86
3	3.18	5.84	20	2.09	2.84
4	2.78	4.60	21	2.08	2.83
5	2.57	4.03	22	2.07	2.82
6	2.45	3.71	23	2.07	2.81
7	2.36	3.50	24	2.06	2.80
8	2.31	3.36	25	2.06	2.79
9	2.26	3.25	26	2.06	2.78
10	2.23	3.17	27	2.05	2.77
11	2.20	3.11	28	2.05	2.76
12	2.18	3.06	29	2.04	2.76
13	2.16	3.01	30	2.04	2.75
14	2.14	2.98	40	2.02	2.70
15	2.13	2.95	60	2.00	2.66
16	2.12	2.92	120	1.98	2.62
17	2.11	2.90	∞	1.96	2.58

This table is abridged from Table 12 in E. S. Pearson and H. O. Hartley (Eds.), *Biometrika tables for statisticians* (3rd ed., Vol. 1), Cambridge University Press, New York, 1970, by permission of the *Biometrika* Trustees.