2

FUNDAMENTAL CONCEPTS

> When we are self-indulgent and uncritical, when we confuse hopes and facts, we slide into pseudoscience and superstition.
>
> —Carl Sagan (1996, p. 27)

This chapter prepares readers for learning about alternatives to statistical tests through survey of fundamental concepts about research designs, variables, and estimation. Also reviewed are characteristics of statistical tests in general and those of three of the most widely used tests in comparative studies, the t test and F test for means and the chi-square (χ^2) test for two-way contingency tables. We will see in the next chapter that there are many misunderstandings about statistical tests, so readers should pay close attention to the discussions that follow. Exercises with answers for this chapter are available on this book's Web site.

TERMS AND IDEAS ABOUT COMPARATIVE STUDIES

Essential ideas about study design and the nature of independent or dependent variables are reviewed in this section. It is hoped that this presentation will build a common vocabulary for later chapters.

Independent Samples Designs and Correlated Designs

An independent variable (factor) has at least two levels. In an *independent samples (between-subjects) design*, each level is studied with an unrelated sample (group), and every case is measured once on the dependent (outcome) variable. If cases are randomly assigned to samples, the factor is a *manipulated* or *experimental variable* and the design is a *randomized-groups* or *completely randomized design*. If cases are classified into groups based on an intrinsic characteristic such as gender, the factor is a *nonexperimental* or *individual-difference variable*. Studies in which all factors are individual-difference variables are referred to as *nonexperimental, correlational,* or *observational studies*.

The samples are related in a *dependent-samples* or *correlated design*. There are two kinds. In a *repeated-measures* or *within-subjects design*, each case is measured at every level of the factor, such as pretest and posttest. This means that the "samples" are actually identical across the levels of the factor. R. Rosenthal, Rosnow, and Rubin (2000) distinguished between *intrinsically* and *nonintrinsically repeated-measures designs*. The logic of the former requires multiple assessments of each case, such as when maturational change is studied in individuals. The rationale of a nonintrinsically repeated-measures design does not require multiple testing of each case because the same factor could theoretically be studied with independent samples. For instance, the effect of caffeine versus no caffeine on athletic performance could be studied with unrelated groups in a completely randomized design or with just one group in a repeated-measures design. In the second kind of correlated design, a *matched-groups design*, a separate group corresponds to each level of the factor, just as in between-subjects designs. The difference is that each case in a matched-groups design is explicitly paired with a case in every other sample on at least one matching variable, which controls for this variable.

Compared to designs with independent samples, correlated designs may reduce error variance and increase statistical power. For these reasons, a correlated design may be chosen over an independent samples design even though the research question does not require dependent samples. These advantages have potential costs, though. Repeated-measures designs may require controls for order effects, and matched-groups designs are subject to regression effects if cases come from the extremes of their respective populations. See Ellis (1999) for a clear discussion of these and other design issues when studying dependent samples.

Balanced and Unbalanced Designs

An independent samples design is *balanced* if the number of cases in each group (n) is the same. If any two groups are of different size, the design is *unbalanced*. With no missing data, correlated designs are inherently balanced. Although there is no general statistical requirement for balanced designs, there are some potential drawbacks to unbalanced designs. One is loss of statistical power even if the total number of cases is the same for a balanced versus an unbalanced design. Suppose that $n_1 = n_2 = 50$ for a balanced two-group design. R. Rosenthal et al. (pp. 30–32) showed that the relative loss of power for an unbalanced design where $n_1 = 70$ and $n_2 = 30$ is equivalent to losing 16 cases (16% of the sample size) from the balanced design. The relative power loss increases as the group size disparity increases.

A critical issue concerns the reason why the group sizes are unequal. For example, an unbalanced design may arise because of randomly missing data from a design intended as balanced, such as when equipment fails and scores are not recorded. A handful of missing observations is probably of no great concern, such as if $n_1 = 100$ and $n_2 = 97$ as a result of three randomly missing scores. A more serious problem occurs when unbalanced designs are a result of nonrandomly missing data, such as when higher proportions of participants drop out of the study under one condition than another. Nonrandomly missing observations in this instance may cause a bias: Study participants who withdrew may differ systematically from those who remain, and the results may not generalize to the intended population. Unfortunately, there is no simple statistical "fix" for bias because of nonrandomly missing data. About all that can be done is to understand the nature of the data loss and how it affects the results; see West (2001) for more information.

Sometimes unbalanced designs are intentional—that is, based on a specific sampling plan. Standardization samples of contemporary ability tests are often stratified by demographic or other variables to match recent census data about the population of the United States. Because sizes of groups based on demographic variables such as gender or age are not usually equal in the population, samples so stratified may be unbalanced. Unequal group sizes in this case is actually an asset because it helps to ensure the representativeness of the sample in terms of relative group sizes. There are also times when groups with relatively low population base rates are intentionally oversampled. This is a common practice in research with special populations. Suppose that the base rate of clinical depression in the general population is 5%. In a particular study, a group of $n_1 = 50$ depressed patients is compared with $n_2 = 50$ control cases. This design is balanced, which maximizes the power of the group contrast. However, the base rate of depression in the sample is 10 times higher than in the population. Because sample base rates affect statistical tests and some types of effect size estimates, the results may not generalize if the population base rates are very different.

Schultz and Grimes (2002) made the point that equal group sizes are not always an asset even in randomized trials. Specifically, they show that forcing equal group sizes through restricted forms of random assignment, such as permuted-blocks randomization, may introduce bias compared to simple randomization, which does not guarantee equal group sizes. Thus, whether unequal group size is a problem depends on the research context.

Multiple Independent or Dependent Variables

Studies with just one independent variable are called *single-factor* or *one-way designs*. However, many behaviors studied by social scientists are affected by more than one variable. One of the goals of a *multifactor design* is to model this complexity by including two or more factors in the design. The terms *higher order*, *factorial*, or *blocking design*, among others, describe various kinds of multifactor designs. Blocking designs involve partitioning the total sample into groups based on an individual-difference variable (e.g., age) believed to affect outcome. If cases within each block are randomly assigned to levels of a manipulated factor, the resulting two-way design is a *randomized-blocks design*. Effect size estimation in single-factor designs is covered in chapters 4 through 6, and chapter 7 deals with this topic for multifactor designs.

Regardless of the number of factors, comparative studies with just one dependent variable are *univariate designs*. Many common statistical tests such as the t and F tests for means are generally univariate tests. *Multivariate designs* have at least two dependent variables, which allows measurement of outcome in more than one area. This book deals only with univariate designs. Because entire volumes are devoted to the basics of multivariate methods (e.g., Grimm & Yarnold, 1995, 2000), it is beyond the scope of this book to deal with them in detail. Also, multivariate analyses often wind up as a series of univariate analyses conducted with individual outcomes. This book's Web site has a supplemental chapter about multivariate effect size estimation in designs with independent samples and fixed factors.

Fixed-Effects and Random-Effects Factors

This distinction affects how the results are to be generalized and how effect size magnitude should be estimated. It is introduced by example: Suppose that the independent variable is dosage of a drug. There are theoretically an infinite number of dosages. If, say, five different dosages are randomly selected for study, the drug factor is a *random-effects factor*. Selecting dosages at random may give a representative sample from all possible levels. If so, the results of the study may generalize to the whole population of dosages. However, if the particular dosages for study are selected by some other means, the drug factor is probably a *fixed-effects factor*. For instance, a researcher may intentionally select five different dosages that form an equal-interval scale, such as 0 (control), 3, 6, 9, and 12 $\text{mg} \cdot \text{kg}^{-1}$. Because these dosages are not randomly selected, the results may not generalize to other dosages not included in the original study, such as 15 mg · kg^{-1}.

Qualitative factors are usually treated as fixed factors. This is especially true for individual-difference variables such as gender where all possible levels may be included in the study. Quantitative variables can be analyzed as either fixed or random factors. A *control factor* is a special kind of random factor that is not of interest in itself but is included for the sake of generality (Keppel, 1991). Suppose that participants are required to learn a list of words. If only a single word list is used, it is possible that the results are specific to the particular words on that list. Using several different lists matched on characteristics such as relative word frequency and treating word list as a random factor may enhance generalizability. Repeated-measures factors that involve trials or measurement at specific times, such as three and six months after treatment, are usually considered fixed. If there are many repeated measures and only some are randomly selected for analysis, the repeated-measures factor is considered random.

Designs with random factors may require special considerations in statistical testing and effect size estimation. Thus, it may be better to consider a factor as fixed instead of random if in doubt. Chapters 6 and 7 deal with designs in which there is at least one random factor. Please note that the *subjects factor* is almost always seen as random because its levels—the individual cases—are usually different from study to study.

Covariate Analyses

Both correlated and blocking designs may reduce error variance compared to independent samples and one-way designs, respectively. Another way is covariate analysis. A *covariate* is a variable that predicts outcome but is ideally unrelated to the independent variable. The variance explained by the covariate is removed, which reduces error variance. Suppose a basic math skills pretest is given to students before they are randomly assigned to different instructional conditions for introductory statistics. Outcome is measured with a common final examination. It is likely that the pretest will covary with exam scores. In an analysis of covariance (ANCOVA), the effect of the pretest is statistically removed from the outcome variable. With enough reduction in error variance, the power of the test of instructional condition may be increased. Because ANCOVA is a statistical method, it can be incorporated into any of the designs mentioned earlier. However, ANCOVA is usually appropriate only for randomly assigned groups, and it is critical to meet the statistical assumptions of this method. These points are elaborated in chapter 6 when effect size estimation in covariate analyses is discussed.

SAMPLING AND ESTIMATION

Basic issues in sampling and estimation are reviewed next, including types of samples, statistics as estimators of population parameters, and interval estimation (i.e., the construction of confidence intervals based on sample statistics).

Types of Samples

One of the hallmarks of behavioral research is the distinction between populations and samples. It is rare that whole populations are studied. If the population is large, vast resources may be needed to study it. For example, the budget for the 2000 census of the population of the United States was about $4.5 billion, and almost a million temporary workers were hired for the endeavor (U.S. Census Bureau, 2002). It may be practically impossible to study even much smaller populations. For example, the base rate of autism is about 4 in 10,000 children (.04%). If autistic children are dispersed over a large geographic area or live in remote regions, studying all of them may be impracticable.

Behavioral scientists must usually make do with small subsets of populations or samples. There are three general kinds of samples: random, systematic, and ad hoc. *Random samples* are selected by a chance-based method that gives all observations an equal probability of appearing in the sample, which may yield a representative sample. Observations in *systematic samples* are selected using some orderly sampling plan that *may* yield a representative sample, but this is not guaranteed. Suppose that an alphabetical list of every household is available for some area. A random number between 10 and 20 is generated and turns out to be 17. Every 17th household from the list is contacted for an interview, which yields a 6% (1/17) sample in that area.

Most samples in social science research are neither random nor systematic but rather *ad hoc samples*, also called *samples of convenience*, *locally available samples*, or *accidental samples*. All of these terms imply the study of samples that happen to be available. A group of undergraduate students in a particular class who volunteer as research participants is an example of a convenience sample. There are two problems with such samples. First, they are probably not representative. For instance, it is known that volunteers differ systematically from nonvolunteers. Second, distributional theories that underlie statistical tests generally assume random sampling. If the data are from ad hoc samples, there is a conceptual mismatch with the test's distributional theory. This is a criticism of statistical tests among others considered in the next chapter.

Despite the potential problems of ad hoc samples, it is often difficult or impossible to collect random or even systematic samples. True random sampling requires a list of all observations in the population, but such lists rarely exist. Also, the notion of random or systematic sampling does not apply to animal research: Samples in this area are almost never randomly selected from known populations of animals. Perhaps the best way to mitigate the influence of bias in ad hoc samples is to follow what is now a fairly standard practice: Measure a posteriori a variety of sample characteristics and report them along with the rest of the results, which allows readers of the study to compare its sample with those of other studies in the same area. Another option is to compare the sample demographic profile with that of the population (if such a profile exists) to show that the sample is not obviously unrepresentative.

Sample Statistics as Estimators

Values of *population parameters*, such as means (μ), variances (σ^2), or correlations (ρ), are usually unknown. They are instead estimated with *sample statistics*, such as M (means), s^2 (variances), or r (correlations). These statistics are subject to *sampling error*, which refers to the difference between an estimator and the corresponding population value. These differences arise because the values of statistics from random samples tend to vary around that of the population parameter. Some of these statistics will be too high and others too low (i.e., they over- or underestimate the population parameter), and only a relatively small number will exactly equal the population value. This variability among estimators from different samples is a statistical phenomenon akin to background (natural) radiation: It's always there, sometimes more or less, fluctuating randomly from sample to sample. The amount of sampling error is generally affected by the variability of population observations, how the samples are selected, and their size. If the population is heterogenous (e.g., σ^2 is large), values of sample statistics may also be quite variable. Obviously, values of estimators from biased samples may differ substantially from that of the corresponding parameter. Given reasonably representative sampling and constant variability among population observations, sampling error varies inversely with sample size. This implies that statistics in larger samples tend to be closer on average to the population parameter than in smaller samples. This property describes the *law of large numbers*, and it says that one is more likely to get more accurate estimates from larger samples than smaller samples.

Sample statistics are either biased or unbiased estimators of the corresponding population parameter. The sample mean is an *unbiased estimator* because its average (expected) value across all possible random samples equals the population mean. The sample variance—also called a *mean square*—is an unbiased estimator of population variance if computed as the ratio of the sum of squares over the degrees of freedom, or

$$s^2 = \frac{SS}{df} = \frac{\sum_{i=1}^{N} (X_i - M)^2}{N - 1} \tag{2.1}$$

where X is an individual score. In contrast, a sample variance derived as $S^2 = SS/N$ is a *negatively biased estimator* because its values are on average less than σ^2. All references to sample variances that follow assume Equation 2.1 unless otherwise indicated. Expected values of statistics that are *positively biased estimators* generally exceed that of the corresponding parameter.

There are ways to correct some statistics for bias. For example, although s^2 is an unbiased estimator of σ^2, the sample standard deviation s is a negatively biased estimator of σ. However, multiplication of s by the correction factor in parentheses that follows

$$\hat{\sigma} = \left(1 + \frac{1}{4(N - 1)}\right)s \tag{2.2}$$

yields the statistic $\hat{\sigma}$, which is a numerical approximation to the unbiased estimator of σ. Because the value of the correction factor in Equation 2.2 is larger than 1.00, $\hat{\sigma} > s$. There is also greater correction for negative bias in smaller samples than in larger samples. If $N = 5$, for instance, the unbiased estimate of σ is

$$\hat{\sigma} = \{1 + 1/[4\ (5 - 1)]\}\ s = (1.0625)s$$

but for $N = 50$, the unbiased estimate is

$$\hat{\sigma} = \{1 + 1/[4\ (50 - 1)]\}\ s = (1.0051)s$$

which shows relatively less adjustment for bias in the larger sample. In even larger samples, the value of the correction factor in the previous equation is essentially 1.00; that is, there is practically no adjustment for bias. This is another instance of the law of large numbers: Averages of even-biased statistics from large samples tend to closely estimate the corresponding parameter.

Point and Interval Estimation

Sample statistics are used for two types of estimation. *Point estimation* is when the value of a sample statistic (e.g., M) is taken as the sole estimate of a parameter (e.g., μ). Because of sampling error, however, it is quite unlikely that the two will be equal. *Interval estimation* recognizes this reality by constructing a *confidence interval* about a point estimate. A confidence interval reflects the amount of sampling error associated with that estimate within a specified level of uncertainty. A confidence interval can also be seen as a range of plausible values for the corresponding parameter. In graphical displays, confidence intervals may be represented as error bars around a single point. Carl Sagan (1996) called error bars "a quiet but insistent reminder that no knowledge is complete or perfect" (pp. 27–28). Wider reporting of confidence intervals is also part of suggested reform of statistical practice in the social sciences (see chapter 1).

We need a more precise definition of a confidence interval. The following is based on Steiger and Fouladi (1997, pp. 229–230):

1. A $1 - \alpha$ confidence interval for (on) a parameter is a pair of statistics yielding an interval that, over repeated samples, includes the parameter with probability $1 - \alpha$. (The symbol α is the level of statistical significance.)
2. A $100\ (1 - \alpha)\%$ confidence interval for a parameter is a pair of statistics yielding an interval that, over repeated samples, includes the parameter $100\ (1 - \alpha)\%$ of the time.

The value of $1 - \alpha$ is selected by the researcher to reflect the degree of statistical uncertainty. The lower bound of a confidence interval is the *lower confidence limit,* and the upper bound is the *upper confidence limit.* Because the most common levels of statistical significance in NHST are $\alpha = .05$ or $\alpha = .01$, one usually sees in the literature either 95% or 99% confidence intervals. However, it is possible to construct confidence intervals that correspond to other levels of statistical significance. For example, error bars around points that represent means in graphs are sometimes each one standard error wide, which corresponds roughly to $\alpha = .32$ and a 68% confidence level.

In traditional confidence intervals—those based on central test statistics (defined next)—the sample statistic is usually exactly between the lower and upper bounds. That is, the width of the interval is symmetrical around the estimator. The phrase "a confidence interval about" an estimator is sometimes used to describe a symmetrical confidence interval. However, this phrase neglects to mention the population parameter that the interval is intended to approximate. It is also the case that the estimator does not always fall at the very center of other kinds of confidence intervals, such as those based on noncentral test statistics (also defined next).

The traditional way to construct a confidence interval is by adding and subtracting from a statistic the product of its standard error and the two-tailed critical value at the α level of statistical significance in a relevant central test distribution, such as t. A *standard error* is the standard deviation of the sampling distribution of an estimator. The square of the standard error is the *conditional variance*, the variance of the sampling distribution. A *sampling distribution* is a probability distribution based on random samples all of size N. In general, standard errors vary directly with variability among population observations and inversely with sample size. The latter explains part of the law of large numbers: Distributions of statistics from larger samples are generally narrower than distributions of the same statistic from smaller samples. A *central test distribution* assumes that the null hypothesis is true. Central test distributions are used in null hypothesis significance testing (NHST) to determine the critical values of test statistics. Tables of critical values for distributions such as t, F, and χ^2 found in many introductory statistics textbooks are based on central test distributions.

Standard errors of statistics with simple sampling distributions can be estimated with formulas that have appeared in statistical textbooks for some time. By a "simple" distribution it is meant that (a) the statistic estimates only a single population parameter, and (b) both the shape and variance of its sampling distribution are constant regardless of the value of the parameter. Distributions of means and mean differences are simple as just defined, and traditional confidence intervals for them are discussed next.

Confidence Intervals for μ

The standard error in a distribution of random means is

$$\sigma_M = \sqrt{\frac{\sigma^2}{N}} \tag{2.3}$$

Because the population variance σ^2 is not generally known, this standard error is usually estimated as

$$s_M = \sqrt{\frac{s^2}{N}} = \frac{s}{\sqrt{N}} \tag{2.4}$$

This estimate is subject to sampling error because the variance s^2 is a sample statistic. The relevant test statistic for means when σ is unknown is central t, so the general form of a confidence interval for μ based on a single observed mean is

$$M \pm s_M \left[t_{2\text{-tail},\,\alpha}\,(N - 1)\right] \tag{2.5}$$

where the term in brackets is the positive two-tailed critical value in a central t distribution with $N - 1$ degrees of freedom at the α level of statistical significance. Suppose we find in a sample of 25 cases that $M = 100.00$ and $s = 9.00$. The standard error is

$$s_M = 9.00/25^{1/2} = 1.80$$

and $t_{2\text{-tail},\ .05}\ (24) = 2.064$. The 95% confidence interval for μ is thus

$$100.00 \pm 1.80\ (2.064)$$

or 100.00 ± 3.72, which defines the interval 96.28–103.72. The 99% confidence interval for μ is constructed the same way except $t_{2\text{-tail},\ .01}\ (24) = 2.797$:

$$100.00 \pm 1.80\ (2.797)$$

or 100 ± 5.03, which defines the interval 94.97–105.03. The 99% confidence interval is wider than the 95% confidence interval based on the same statistic because a greater margin of error is allowed.

Let us consider now the correct interpretation of the 95% confidence interval for μ derived earlier, 96.28–103.72:

1. This interval defines a range of outcomes that should be considered equivalent to the observed result ($M = 100.00$) given the amount of expected sampling error at the 95% confidence level.
2. It also provides a reasonable estimate of the population mean. That is, μ could be as low as 96.28 or it could be as high as 104.72, again at the 95% confidence level.
3. Of course, there is no guarantee that μ is actually included in the confidence interval. We could construct the 95% confidence interval around the mean in another sample, but the center or endpoints of this new interval will probably be different compared with the original. This is because confidence intervals are subject to sampling error, too.
4. However, if 95% confidence intervals are constructed around the means of all random samples drawn from the same population, then 95/100 of them will include μ.

The last point gives a more precise definition of what we mean by "95% confidence level" or "95% confident" from a *frequentist* or *long-run relative-frequency* view of probability as the likelihood of an outcome over repeatable events under constant conditions except for random error. This view also assumes that probability is a property of nature that is independent of what the researcher believes. In contrast, a *subjectivist* or *subjective degree-of-belief* view defines probability as a personal belief the researcher has about nature that is independent of nature's true state. The same view also does not distinguish between repeatable and unrepeatable (unique) events (Oakes, 1986; Reichardt & Gollob, 1997). Although researchers in their daily lives probably take a subjective view of probabilities, it is the frequentist definition that generally underlies sampling theory.

A researcher is probably more interested in knowing the probability that a *specific* 95% confidence interval contains μ than in knowing that 95/100 of all such intervals do. From a frequentist view, this probability for the unique confidence interval of our example, 96.28–103.72, is either 0 or 1.0. That is, this interval either contains μ or it does not. Thus, it is generally *incorrect* from this perspective to say that the interval 96.28–103.72 has a probability of .95 of including μ. Reichardt and Gollob (1997) noted that this kind of *specific probability inference* and the related *specific confidence inference* that one is 95% confident that the interval includes μ is permitted only in a very particular circumstance, which is that every possible value of μ is considered equally likely before the study is conducted. In Bayesian estimation, which is based on a subjectivist view of probability, the same circumstance is described by the *principle of indifference*, which says that in the total absence of information about the parameter, equal probabilities are assumed for all possible values. However, rarely do we have absolutely *no* information about likely or even plausible values for the population mean. In contrast, percentages associated with Bayesian confidence intervals *are* interpreted as probabilities that the parameter lies within the interval. This is what most researchers really want to know but generally cannot get from a traditional confidence interval. The fundamentals of Bayesian estimation are considered in chapter 9.

There is a kind of compromise language for describing traditional confidence intervals that "splits the difference" between frequentist and subjectivist views of probability. Applied to our example, it goes like this: The unique interval 96.28–103.72 estimates μ, with 95% confidence. This statement may not be incorrect from a frequentist perspective because it is not quite a specific confidence inference. It also gives a nod toward the subjectivist view because it associates a degree of belief with a specific interval. Like other compromises, however, it may not please purists who hold one view of probability or the other.

The issues raised about the proper interpretation of percentages associated with unique confidence intervals foreshadow similar difficulties in interpreting probabilities (p values) from statistical tests. Part of the problem is a clash between the long-run relative-frequency view of probability generally assumed by these tests and a subjective view of probability held by perhaps most researchers who use them. Another is the gap between what researchers really want to know and what a p value from a statistical test actually tells them.

Confidence Intervals for $\mu_1 - \mu_2$

The standard error in a distribution of differences (contrasts) between pairs of means from independent samples selected from different populations is

$$\sigma_{M_1 - M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{2.6}$$

where σ_1^2 and σ_2^2 are the population variances and n_1 and n_2 are the sizes of each sample (group). If we assume homogeneity of population variance (i.e., $\sigma_1^2 = \sigma_2^2$), the expression for the standard error reduces to

$$\sigma_{M_1 - M_2} = \sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{2.7}$$

where σ^2 is the common population variance. This variance is usually unknown, so the standard error is estimated by

$$s_{M_1 - M_2} = \sqrt{s_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{2.8}$$

where s_P^2 is the pooled within-groups variance, which is the average of the two group variances weighted by the degrees of freedom. It's equation is

$$s_P^2 = \frac{SS_W}{df_W} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{2.9}$$

where SS_W and df_W are, respectively, the pooled within-groups sum of squares and degrees of freedom. The latter can also be expressed as $df_W = df_1 + df_2 = N - 2$. Only in balanced designs can s_P^2 also be calculated as the average of the two group variances, or $(s_1^2 + s_2^2)/2$.

The general form of a confidence interval for $\mu_1 - \mu_2$ based on the difference between independent means is

$$(M_1 - M_2) \pm s_{M_1 - M_2}\,[t_{2\text{-tail},\,\alpha}\,(N - 2)] \tag{2.10}$$

where $M_1 - M_2$ is the observed mean contrast and $N - 2$ is the pooled within-groups degrees of freedom (df_W) of the positive two-tailed critical value of t at the α level of statistical significance. Suppose in a balanced two-group design where $n = 5$ we observe

$$M_1 - M_2 = 2.00,\ s_1^2 = 7.50,\ s_2^2 = 5.00$$

which implies $s_P^2 = (7.50 + 5.00)/2 = 6.25$. The standard error for the contrast is

$$s_{M_1 - M_2} = [6.25\ (1/5 + 1/5)]^{1/2} = 1.58$$

and $t_{2\text{-tail},\ .05}$ (8) = 2.306. The 95% confidence interval for the mean contrast is

$$2.00 \pm 1.58\ (2.306) \qquad \text{or} \qquad 2.00 \pm 3.65 \tag{2.11}$$

which defines the interval −1.65-5.65. Based on these results we can say that $\mu_1 - \mu_2$ could be as low as −1.65 or as high as 5.65, with 95% confidence. Please note that this interval includes zero as a reasonable estimate of $\mu_1 - \mu_2$. This fact is subject to misinterpretation. For example, it may be incorrectly concluded that $\mu_1 = \mu_2$ because zero falls between the lower and upper bounds of the confidence interval. However, zero is only one value within a range of estimates of $\mu_1 - \mu_2$, so in this sense it has no special status in interval estimation for this example. Besides, the confidence interval itself is subject to sampling error, so zero may not be included within the 95% confidence interval for $\mu_1 - \mu_2$ in a replication. It is the range of overlap between the two confidence intervals (if any) that would be of greater interest than whether zero is included in one interval or the other. These issues are elaborated next.

Now let us consider confidence intervals for contrasts between dependent means. Below we use the symbol M_D to refer the average *difference score* when two dependent samples are compared. A difference score is computed as $D = X_1 - X_2$ for each of the n cases in a repeated-measures design or for each of the n pairs of cases in a matched-groups design. (Difference scores are also called *gain scores* or *change scores*.) If $D = 0$, there is no difference; any other value indicates a higher score in one condition than in the other. The average of all the difference scores equals the dependent mean contrast, or $M_D = M_1 - M_2$. The standard error of M_D is

$$\sigma_{M_D} = \sqrt{\frac{\sigma_D^2}{n}} = \frac{\sigma_D}{\sqrt{n}} \tag{2.12}$$

where the σ_D^2 and σ_D are, respectively, the population variance and standard deviation of the difference scores. The variance σ_D^2 takes account of the population correlation of the scores between the conditions, which is designated in Equation 2.13 as ρ_{12}. Assuming homogeneity of variance, the variance of the difference scores is

$$\sigma_D^2 = 2\sigma^2(1 - \rho_{12}) \tag{2.13}$$

where σ^2 is the common population variance. When there is a stronger *subjects effect*—cases maintain their relative positions across the conditions—ρ_{12} approaches 1.00. This reduces the variance of the difference scores, which in turn reduces the standard error of the dependent mean contrast

(Equation 2.12). It is this subtraction of consistent individual differences from the standard error that makes confidence intervals based on dependent mean contrasts generally narrower than confidence intervals based on contrasts between unrelated means. It also explains the power advantage of the t test for dependent samples over the t test for independent samples, which is considered next. However, these advantages are realized only if $\rho_{12} > 0$. Otherwise, confidence intervals and statistical tests may be wider and less powerful (respectively) for dependent mean contrasts.

The population variance of the difference scores, σ_D^2, is usually unknown, but it is often estimated as

$$s_{M_D} = \frac{s_D^2}{n} = \frac{s_D}{\sqrt{n}} \tag{2.14}$$

where s_D^2 and s_D are, respectively, the sample variance and standard deviation of the difference scores. The former is calculated as

$$s_D^2 = s_1^2 + s_2^2 - 2\ cov_{12} \tag{2.15}$$

where cov_{12} is covariance of the observed scores across the conditions. It is the product of the cross-conditions correlation and the within-conditions standard deviations:

$$cov_{12} = r_{12}\ s_1\ s_2 \tag{2.16}$$

As r_{12} approaches 1.00, the variance s_D^2 gets smaller, which in turn decreases the estimated standard error of the dependent mean contrast.

The general form of a confidence interval for μ_D is

$$M_D \pm s_{M_D}\ [t_{\text{2-tail},\,\alpha}\ (n - 1)] \tag{2.17}$$

Suppose for a dependent samples design we observe the following data:

$$M_1 - M_2 = 2.00,\ s_1^2 = 7.50,\ s_2^2 = 5.00,\ r_{12} = .735$$

Given the above information,

$$\begin{aligned} s_D^2 &= 7.50 + 5.00 - 2\ (.735)\ (7.50^{1/2})\ (5.00^{1/2}) = 3.50 \\ s_{M_D} &= (3.50/5)^{1/2} = .837 \end{aligned}$$

The value of $t_{\text{2-tail},\ .05}\ (4)$ is 2.776, so the 95% confidence interval for μ_D is

$$2.00 \pm .837\ (2.776) \text{ or } 2.00 \pm 2.32 \tag{2.18}$$

which defines the interval –.32–4.32. Please note that the 95% confidence interval assuming a dependent-samples design is narrower than the 95% confidence interval based on the same means and variances for an independent-samples design, which is –1.65–5.65. (Compare Equations 2.11 and 2.18.) This result is expected because r_{12} is relatively high (.735) for the dependent-samples design (r_{12} is presumed to be zero when the samples are independent).

Confidence Intervals for Other Kinds of Statistics

Many statistics other than means have complex distributions. For example, distributions of sample proportions for a dichotomous variable are symmetrical only if the population proportion is $\pi = .50$; the same is true for the Pearson correlation r only if the population correlation is $\rho = 0$. Other statistics have complex distributions because they estimate more than one population parameter. This includes some widely used effect size indexes such as standardized mean differences, which for contrasts between independent means generally estimate $\delta = (\mu_1 - \mu_2)/\sigma$, the ratio of the population mean difference over the common population standard deviation. (Chapter 4 considers standardized mean differences in detail.)

Until recently, confidence intervals for statistics with complex distributions have been estimated with approximate methods. One such method involves *confidence interval transformation* (Steiger & Fouladi, 1997) in which the statistic is mathematically transformed into units that are normally distributed. The confidence interval is built by adding and subtracting from the transformed statistic the product of the standard error in the transformed metric and the appropriate positive two-tailed critical value of the normal deviate z. The lower and upper bounds of this interval are then transformed back into the metric of the original statistic, and the resulting interval may be asymmetrical around that statistic. The construction of confidence intervals for ρ based on the Fisher's Z transformation of r is an example of this approach, which is covered in many statistics textbooks (e.g., Glass & K. Hopkins, 1996, pp. 357–358). Other transformation-based methods for constructing confidence intervals for the population parameters estimated by effect size statistics are introduced in later chapters.

Another approximate method builds confidence intervals directly around the sample statistic and are thus symmetrical about it. The width of the interval on either side is a product of the two-tailed critical value of a central test statistic and an estimate of the *asymptotic standard error*, which estimates what the standard error of the statistic would be in a large sample (e.g., $N > 500$). However, if the researcher's sample is not large, the estimated standard error based on this approach may not be very accurate.

Another drawback to this method is that the distributions of some sample statistics, such as the multiple correlation R, are so complex that a computer is needed to derive the estimated standard error. Fortunately, there are increasing numbers of computer programs for calculating confidence intervals, some of which are mentioned later.

A more exact method for constructing confidence intervals for statistics with complex distributions is *noncentrality interval estimation* (Steiger & Fouladi, 1997). It also deals with situations that cannot be handled by approximate methods. This method is based on *noncentral test distributions* that do not assume that the null hypothesis is true. A bit of perspective is in order: Families of central distributions of t, F, and χ^2 are special cases of noncentral distributions of each test statistic just mentioned. Compared to central distributions, noncentral distributions have an additional parameter called the *noncentrality parameter*. This extra parameter basically indicates the degree of departure from the null hypothesis. For example, central t distributions are described by a single parameter, the degrees of freedom, but noncentral t distributions are described by both the degrees of freedom and a noncentrality parameter. If this parameter equals zero, the resulting distribution is the familiar and symmetrical central t distribution. As the value of the noncentrality parameter is increasingly positive, the noncentral t distributions described by it become increasingly positively skewed (e.g., Cumming & Finch, 2001, fig. 5). The same thing happens but in the opposite direction for negative values of the noncentrality parameter for t distributions.

Noncentrality interval estimation is impractical without relatively sophisticated computer programs for iterative estimation. Until just recently, such programs have not been widely available to applied researchers. A notable exception in a commercial software package for general statistical analyses is the Power Analysis module by J. Steiger in STATISTICA (StatSoft Inc., 2003), which can construct noncentral confidence intervals based on several different types of statistics (Steiger & Fouladi, 1997). This includes many of the standardized indexes of effect size introduced in later chapters. There are now also a few different stand-alone programs or scripts (macros) for noncentrality interval estimation, some available for free through the Internet. These programs or scripts are described in chapter 4, and the Web site for this book also has links to corresponding download pages.

Later chapters demonstrate the calculation of both approximate and more exact noncentral confidence intervals for standardized effect size indexes. The technique of bootstrapping, a method for statistical resampling, can also be used to construct confidence intervals. Chapter 9 reviews the rationale of bootstrapping.

LOGIC OF STATISTICAL SIGNIFICANCE TESTING

A brief history of NHST was given earlier. This section outlines the basic rationale and steps of NHST as it is often practiced today. The following review lays the groundwork for understanding limitations of NHST considered in the next chapter.

Contexts and Steps

There are two main contexts for NHST, *reject–support* (RS) and *accept–support* (AS). The former is the most common and concerns the case in which rejection of the null hypothesis supports the researcher's theory. The opposite is true in AS testing: It is the *failure* to reject the null hypothesis that supports what the researcher actually believes. Listed next are the main steps of NHST for both RS and AS testing. Each step is discussed in the sections that follow with emphasis on points that are not as well known as they should be.

1. Based on the research question, formulate the first of two statistical hypotheses, the null hypothesis H_0.
2. Formulate the second statistical hypothesis, the alternative hypothesis H_1.
3. Set the level of statistical significance α, which is the probability of a Type I error.
4. Collect the data and determine its probability p under H_0 with a statistical test. Reject H_0 if $p < \alpha$.

Null Hypotheses

The null hypothesis is a default explanation that may be rejected later given sufficient evidence. In RS testing, this default explanation is the opposite of the researcher's theory; in AS testing, the null hypothesis reflects the researcher's theory. In either RS or AS testing, the null hypothesis is usually a *point hypothesis* that specifies the numerical value of at least one population parameter. There are two different kinds of null hypotheses (J. Cohen, 1994). A *nil hypothesis* says that the value of a population parameter is zero or the difference between two or more parameters is zero. Examples of nil hypotheses are presented next:

$$H_0: \mu_D = 0 \qquad H_0: \mu_1 - \mu_2 = 0 \qquad H_0: \rho = 0$$

Nil hypotheses are usually statements of absence, whether of an effect, difference, or association. In contrast, a *non-nil hypothesis* asserts that a population parameter is not zero or that the difference between two or more parameters is not zero. It typically assumes a non-zero effect, difference, or association. Examples of non-nil hypotheses are given next:

$$H_0\text{: } \mu_D = 10.00 \qquad H_0\text{: } \mu_1 - \mu_2 = 5.00 \qquad H_0\text{: } \rho = .30$$

Nil hypotheses as default explanations are generally most appropriate when it is unknown whether effects or relations exist at all, such as in new research areas where most studies are exploratory. However, nil hypotheses are less suitable when it is known a priori that an effect is probably not zero. This is more likely in established research areas. For instance, it is known that women and men differ in certain personality characteristics (e.g., Feingold, 1994). Specification of H_0: $\mu_1 - \mu_2 = 0$ (i.e., H_0: $\mu_1 = \mu_2$) when testing gender differences in these characteristics may set the bar too low because this nil hypothesis is probably false. Accordingly, rejecting it is not an impressive scientific achievement. There are also situations where specification of a nil hypothesis is clearly indefensible. One example is using a nil hypothesis to test a reliability coefficient for statistical significance. For example, Abelson (1997a) noted that declaring a reliability coefficient to be nonzero based on such a test is the "ultimate in stupefyingly vacuous information" (p. 121). This is because what is really important to know is whether a reliability coefficient is acceptably high for a specific purpose, such as $r_{XX} > .90$ when a test is used for individual assessments that determine access to treatment resources.

Nil hypotheses are tested much more often in the social sciences than non-nil hypotheses. This is true even in established research areas where a nil hypothesis is often a "straw man" argument. There are at least three reasons for this puzzling situation: Many researchers are unaware of the possibility to specify non-nil hypotheses. Statistical software programs usually test only nil hypotheses. This means that tests of non-nil hypotheses must be computed by hand. Unfortunately, this is generally feasible only for relatively simple non-nil hypotheses, such as H_0: $\mu_1 - \mu_2 = 5.00$, which can be evaluated without difficulty with the t test.

Alternative Hypotheses

This second statistical hypothesis complements H_0. In RS testing, the alternative hypothesis H_1 represents the researcher's theory; in AS testing, it does not. Unlike the null hypothesis, the alternative hypothesis is typically a *range hypothesis* that specifies a range of values for the population parameter(s). The two kinds of alternative hypotheses are directional (one-tailed, one-sided) and nondirectional (two-tailed, two-sided). A *nondirectional alternative hypothesis* predicts any result not specified in H_0, but a *directional* *alternative hypothesis* specifies a range of values on only one side of the point prediction in H_0. For example, given H_0: $\mu_1 = \mu_2$, there is only one possible nondirectional alternative hypothesis, H_1: $\mu_1 \neq \mu_2$, but two possible directional alternatives, H_1: $\mu_1 > \mu_2$ or H_1: $\mu_1 < \mu_2$.

The choice between a nondirectional or directional H_1 is supposed to be made before the data are collected. If there are a priori reasons to expect a directional effect, the appropriate directional H_1 should be specified; otherwise, a nondirectional H_1 may be a safer bet. The choice between a directional or nondirectional H_1 affects the results of statistical tests as follows: It is easier to reject H_0 when a directional H_1 is specified *and* the data are in the same direction. If H_0 is actually false, there is also greater statistical power compared to a nondirectional H_1. However, if a directional H_1 is specified but the data indicate an effect in the opposite direction, then one is supposed to fail to reject H_0 even if the results are very inconsistent with it. In practice, however, these conventions are not always followed. For example, it is sometimes not possible to reject H_0 for a nondirectional H_1 but it is possible for a directional H_1. A researcher who initially specified a nondirectional H_1 may "switch" to a directional alternative hypothesis to reject the null hypothesis. It also happens that researchers "switch" from one directional H_1 to another depending on the data, again to reject H_0. Some would consider changing H_0 or H_1 based on sample results a kind of statistical "sin" that is to be avoided. Like admonitions against other kinds of sins, they are not always followed.

Level of Type I Error

Alpha (α) is the probability of making a Type I error; more specifically, it is the conditional prior probability of rejecting H_0 when it is actually true (Pollard, 1993). Alpha is a prior probability because it is specified before the data are collected, and it is a conditional probability because H_0 is assumed true. In other words,

$$\alpha = p\ (\text{Reject } H_0 \mid H_0 \text{ true}) \tag{2.19}$$

where the symbol "|" means *assuming* or *given*. Alpha can also be understood as the probability of getting a result from a random sample that leads to the incorrect decision to reject the null hypothesis. All these descriptions of α are also long-run, relative-frequency statements about the probability of a Type I error.

Conventional levels of α in the social sciences are either .05 or .01. When other levels are specified, they tend to be even lower, such as α = .001. It is rare for researchers to specify α levels higher than .05. The main reason is editorial policy: Manuscripts may be rejected for publication if

$\alpha > .05$. This policy would make more sense if the context for NHST were always RS testing where a Type I error is akin to a false positive because the evidence is incorrectly taken as supporting the researcher's theory. As noted by Steiger and Fouladi (1997), the value of α should be as low as possible from the perspective journal editors and reviewers, who may wish to guard against incorrect claims. In AS testing, however, they should worry less about Type I error and more about Type II error because false claims in this context arise from *not* rejecting H_0. Insisting on low values of α in this case may facilitate publication of erroneous claims.

It is important to realize that α sets the risk of a Type I error for a single hypothesis test. However, rarely is just one hypothesis tested. When multiple statistical tests are conducted, there is also an *experimentwise (familywise) probability* of Type I error, designated below as α_{EW}. It is the likelihood of making one or more Type I errors across a set of tests. If each individual test is conducted at the same level of α, then

$$\alpha_{EW} = 1 - (1 - \alpha)^c \tag{2.20}$$

where c is the number of tests, each conducted at the α level of statistical significance. In this equation, the term $(1 - \alpha)$ is the probability of *not* making a Type I error for any individual test; $(1 - \alpha)^c$ is the probability of making *no* Type I errors across all tests; and the whole expression is the likelihood of committing at least one Type I error among the tests. We need to understand a couple of things about this equation. If only one hypothesis is tested, then $\alpha_{EW} = \alpha$. If there are multiple tests, this equation is accurate only if the hypotheses or outcome variables are perfectly uncorrelated. If not, the estimated rate of experimentwise Type I error given by this equation will be too low. The result generated by the equation is the probability of one or more Type I errors, but it does not indicate how many Type I errors may have been committed (it could be 1, or 2, or 3 . . .) or on which hypothesis tests they occurred. Suppose that 20 statistical significance tests are conducted each at $\alpha = .05$ level in the same sample. The experimentwise Type I error rate is

$$\alpha_{EW} = 1 - (1 - .05)^{20} = .64$$

In other words, the risk of a making a Type I error across the whole set of 20 tests is 64%, given the assumptions just stated.

There are two basic ways to control experimentwise Type I error: Reduce the number of tests or lower α for each one. The former can be realized by honing one's questions down to the most substantively meaningful (prioritize the hypotheses). This also means that "fishing expeditions" where essentially every effect is tested are to be avoided. Another way to reduce the number of hypotheses is to use multivariate methods, which can test hypotheses across several variables at once. There is a relatively simple method to set α for individual tests called the *Bonferroni correction:* Divide a target value of α_{EW} by the number of tests, and set the corrected level of statistical significance α^* for each individual test to the value of this ratio. Suppose a researcher wishes to limit the experimentwise risk of Type I error to 10%. If a total of 20 tests are planned, then $\alpha^* = .10/20 = .005$ for each individual test. Other methods are considered in chapter 6. However, readers should know that not all methodologists believe that controlling experimentwise Type I error is a generally desirable goal in the social sciences. This opinion stems from the apparently low general statistical power of social science research, an issue discussed later.

Like the choice between a directional and nondirectional H_1, the decision about α is supposed to be made before the data are collected. For example, if $\alpha = .01$ but the estimated probability of the data under H_0 is .03, one is supposed to fail to reject H_0. However, the temptation to increase α (or α^*) from .01 to .05 to reject H_0 may be strong in this case. Increasing α based on the data is another form of statistical sin that occurs in the real world.

Statistical Tests

The most widely used test statistics in the social sciences are probably the t, F, and χ^2 statistics, but there are many others. Although different in their applications, assumptions, and distributions, all such tests do basically the same thing: A result is summarized with a sample statistic. The difference between the statistic and the value of the corresponding population parameter(s) specified in the null hypothesis is compared against an estimate of sampling error. A computer program for general statistical analyses will convert test statistics to probabilities based on the appropriate theoretical central test distribution. These probabilities are often printed in program output under the column heading p, which is the same abbreviation used in many journal articles. One should not forget that p actually stands for the conditional probability

$$p\ (\text{Data} \mid H_0 \text{ true})$$

which should be understood as the likelihood of the sample result or one even more extreme assuming the null hypothesis is true. Both p and α are derived in the same sampling distribution and are properly interpreted as long-run, relative-frequency probabilities. Unlike α, however, p is *not* the conditional prior probability of a Type I error because it is computed for a particular sample result. To differentiate the two probabilities, Gigerenzer

(1993) referred to p as the *exact level of significance*. If this exact significance level is less than the conditional prior probability of a Type I error ($p < \alpha$), then H_0 is rejected and the result is considered statistically significant at that level of α. If $\alpha = .05$ and $p = .032$, for example, then H_0 is rejected, the result is taken as statistically significant at the .05 level, and its exact level of statistical significance is .032.

It can be shown that many test statistics can be expressed as the product

$$\text{Test statistic} = f\,(N) \times \text{ES index} \qquad (2.21)$$

where $f\,(N)$ is a function of sample size for the particular test statistic and ES Index is an effect size index that expresses the degree of discrepancy between the data and H_0 in a standardized metric (R. Rosenthal, 1994). Various standardized effect size indexes are introduced in later chapters, but for now consider two implications of this equation:

1. Holding sample size constant, the absolute values of test statistics generally increase with no upper bound and their p values approach zero as the effect size increases.
2. Holding constant a non-zero effect size, increasing the sample size causes the same change in test statistics and p values.

These implications explain how it is possible for even trivial effects to be statistically significant in large samples. They also explain how even large effects may not be statistically significant in small samples. In other words, statistical significance does *not* imply that an effect is large, important, or even interesting. By the same token, one cannot conclude that the absence of statistical significance indicates a small or unimportant effect.

That p values from statistical tests (a) are both conditional and long-run, relative-frequency probabilities and (b) measure sample size as well as effect size makes them apparently difficult to correctly interpret. Evidence that p values are in fact widely misunderstood in the behavioral sciences like psychology is considered in the next chapter.

POWER

Power is the conditional prior probability of making the correct decision to reject H_0 when it is actually false, or

$$\text{Power} = p\ (\text{Reject } H_0 \mid H_0 \text{ false}) \qquad (2.22)$$

A Type II error, on the other hand, occurs when the sample result leads to the failure to reject H_0 when it is actually false. The probability of a Type II error is usually represented by β, and it is also a conditional prior probability:

$$\beta = p \text{ (Fail to reject } H_0 \mid H_0 \text{ false)} \tag{2.23}$$

Because power and β are complementary, or

$$\text{Power} + \beta = 1.00 \tag{2.24}$$

whatever increases power decreases the probability of a Type II error and vice versa. Summarized next are factors that affect the power of statistical tests:

1. The lower the level of α, the lower is power. Thus, reducing the chance of a Type I error increases the likelihood of a Type II error. However, there are other ways to increase power besides increasing α.
2. Power is greater in larger samples. This fact is demonstrated below for the test statistics t, F, and χ^2.
3. Power is greater when H_1 is directional and the population effect is in the same direction. If the two disagree, however, power is essentially zero.
4. Study design: (a) Correlated designs are generally more powerful than independent-samples designs. Blocking designs and covariate analyses may also increase power. (b) Balanced designs are generally more powerful than unbalanced designs. (c) Tests for effects of fixed factors are usually more powerful than tests for effects of random factors.
5. Power declines as the reliability of the scores in a particular sample is lower. With lower score reliability comes higher error variance, which makes it more difficult to detect a real effect.
6. Parametric tests such as t and F are generally more powerful than nonparametric tests. See Siegel and Castellan (1988) for information about nonparametric tests.
7. In general, the larger an effect in the population for a given design, the easier it is to detect in samples. However, the magnitude of the real effect that can theoretically be observed is somewhat under the control of the researcher. For example, a longer, more intense intervention may potentially have a larger effect than a shorter, less intense intervention.

A power analysis gives the probability of rejecting H_0. There are two kinds. An *a priori power analysis* is conducted before the data are collected. It involves (a) specification of the study's planned characteristics, such as the level of α and the sample size; and (b) estimation of the expected magnitude of the population effect. The latter may be based on theory, results of previous empirical studies, or an educated guess. If the researcher is uncertain about the population effect size, power can be calculated for a range of estimates. A variation is to specify a desired level of power and then estimate the minimum sample size needed to obtain it. A *post hoc power analysis* is conducted after the data are collected. The observed effect size is treated as the population effect size, and the probability of rejecting the null hypothesis given the study's characteristics is estimated. However, a post hoc analysis that shows low power is more like an autopsy than a diagnostic procedure. That is, it is better to think about power before it is too late.

In the past, researchers conducted power analyses by consulting special tables presented in sources such as J. Cohen (1988). Now there are several computer programs for power analysis on personal computers. Some of these programs, such as the Power Analysis module of STATISTICA, use noncentral test distributions, which are generally necessary for correct power estimates. Power analysis programs assume the user knows something about the effect size indexes described in later chapters.

Estimated power levels no higher than .50 are problematic. If power is only .50, the probability of rejecting H_0 when it is false is no greater than guessing the outcome of a coin toss. In fact, tossing a coin instead of actually conducting a study would be just as likely to give the correct decision and would save time and money, too (F. Schmidt & Hunter, 1997). Unfortunately, the results of several reviews indicate that the typical power of social science research may be no greater than about .50 (e.g., Sedlmeier & Gigerenzer, 1989). When power estimates are broken down by whether estimated effect sizes are small, medium, or large—specific definitions of these adjectives are given in chapter 4—their values are about .20, .50, and .80, respectively. Power for the study of large effects, .80, is certainly better than for the others listed but still results in a Type II error rate of 20%. Increasing sample sizes would address the problem of low power, but the number of additional cases necessary to reach even a power level of .80 when studying effects of small or medium magnitude may be so great as to be practically impossible (F. Schmidt, 1996). This is a critical limitation of NHST in the social sciences.

The concept of power does not stand by itself without NHST. As statistical tests play a smaller role in our analyses, the relevance of power will also decline. If statistical tests are not used at all, the whole idea of power is meaningless. Besides, it is a stronger scientific result to observe the same effect at $p < .10$ across two different smaller samples than to find $p < .05$ in one larger sample.

OVERVIEW OF SPECIFIC STATISTICAL TESTS

Reviewed next are essential characteristics of three widely used test statistics in the behavioral sciences, the *t* and *F* statistics for means and the χ^2 statistic for two-way contingency tables. The *F* statistic is part of a family of techniques known as the analysis of variance (ANOVA), but note that the *F* statistic is not synonymous with ANOVA. That is, we can conduct an ANOVA without computing the *F* statistic, but not vice versa. It is reviewed only for designs with a single fixed factor, but its basic logic generalizes to other kinds of designs with continuous outcome measures (chapters 6 and 7). The *t* statistic is discussed only for designs with two conditions, such as treatment versus control. It is important for readers to know that the *t*, *F*, and χ^2 statistics are not reviewed for their own sakes. This is because they are subject to the general drawbacks of all NHST methods that are considered in the next chapter. These shortcomings are so serious that it is recommended that the continued use of statistical tests as the primary inference tool in the behavioral sciences is not acceptable. However, familiarity with the sample descriptive statistics that contribute to the *t*, *F*, and χ^2 statistics gives one a head start toward understanding effect size estimation. It is also possible in many cases to compute effect size indexes from these test statistics. Please also note that the sampling distributions of the *t*, *F*, and χ^2 tests described are, respectively, the central *t*, central *F*, and central χ^2 distributions in which the null hypothesis is assumed to be true. In later chapters readers will learn more about noncentral *t* and noncentral *F* distributions in which the null hypothesis is assumed to be false. These distributions are required for calculating exact confidence intervals based on certain kinds of standardized indexes of effect size.

t TESTS FOR MEANS

The *t* tests reviewed compare means from either two independent or dependent samples. Both are actually special cases of the *F* test for means. Specifically, $t^2 = F$ when both statistics are computed for the same mean contrast for a nil hypothesis. The statistical assumptions of the *t* tests for independent versus dependent samples are the same as those of the corresponding *F* tests and are discussed later.

The general form of the *t* statistic for a contrast between independent means is

$$t(N-2) = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{M_1 - M_2}} \qquad (2.25)$$

where $N - 2$ is the pooled within-groups degrees of freedom (df_W), $M_1 - M_2$ and $s_{M_1 - M_2}$ are, respectively, the observed mean contrast and its standard error (Equation 2.8), and $\mu_1 - \mu_2$ is the population mean difference specified in the null hypothesis. If the latter is predicted to be zero, a nil hypothesis is tested; otherwise a non-nil hypothesis is tested.

The *t* statistic for a dependent mean contrast has the same overall form:

$$t(n - 1) = \frac{M_D - \mu_D}{s_{M_D}} \tag{2.26}$$

where the degrees of freedom equal the group size (n) minus 1, M_D and s_{M_D} are, respectively, the observed average difference score and its standard error (Equation 2.14), and μ_D is the population average difference score specified in the null hypothesis. For a nil hypothesis, $\mu_D = 0$, and this term drops out of the equation.

Assuming a nil hypothesis, both forms of the *t* statistic defined earlier express a mean contrast as the proportion of its standard error. If $t = 1.50$, for example, the first mean is 1½ standard errors higher than the second, but note that the sign of *t* is arbitrary because it depends on the direction of subtraction between the two means. It is important to realize that the standard error metric of the *t* test is affected by sample size, which is demonstrated now.

This is explained in Tables 2.1 and 2.2, described next. Table 2.1 presents the means and variances of two groups where $M_1 - M_2 = 2.00$. Table 2.2 reports the results of the independent samples *t* test for the data in Table 2.1 at three different group sizes, $n = 5$, 15, and 30, for a nondirectional H_1. (Readers are encouraged to reproduce these results.) Please note in Table 2.2 that the value of the pooled within-groups variance for these data, $s_P^2 = 6.25$, is unaffected by group size. This is not true for the standard error of $M_1 - M_2$, which, as expected, gets smaller as n increases. This causes the value of *t* to go up and its probability to go down for the larger group sizes. Consequently, the *t* statistic for $n = 5$ does not indicate a statistically significant contrast at the .05 level, but it does for the two larger group sizes. Results for the latter indicate less expected sampling error, but not a

TABLE 2.1
Means and Variances for Two Independent Samples

	Group	
	1	2
M	13.00	11.00
s^2	7.50	5.00

TABLE 2.2
Results of the Independent Samples *t* Test for the Data in Table 2.1 at Three Different Group Sizes

Statistic	Group size (n) 5	15	30
$s_{M_1 - M_2}$	1.58	.913	.645
t	1.26	2.19	3.10
df	8	28	58
p	.243	.037	.003
$t_{\text{2-tail, .05}}$	2.306	2.048	2.002
95% CI for $\mu_1 - \mu_2$	−1.64–5.64	.13–3.87	.71–3.29

Note. For all analyses, $M_1 - M_2 = 2.00$ and $s_P^2 = 6.25$. CI = confidence interval.

different or more substantial mean contrast. This reduction in sampling error is also evident in the 95% confidence intervals about the observed mean difference: Their widths decrease as n gets larger.

The standard error metric of the t test is also affected by whether the means are independent or dependent. This is demonstrated next. Table 2.3 presents raw scores and descriptive statistics for a small data set where the observed mean difference is 2.00. Reported in Table 2.4 are the results of two different t tests and 95% confidence intervals for the data in Table 2.3. The first analysis assumes $n = 5$ cases in each of two unrelated samples, but second analysis assumes $n = 5$ pairs of scores across two dependent samples. Only the second analysis takes account of the positive cross-conditions correlation for these data, $r_{12} = .735$. Observe in Table 2.4 the narrower 95% confidence interval, the higher value of t, and its lower p value in the dependent samples analysis relative to the independent samples analysis of the same data.

TABLE 2.3
Raw Scores and Descriptive Statistics for Two Samples

	Condition 1	2
	9	8
	12	12
	13	11
	15	10
	16	14
M	13.00	11.00
s^2	7.50	5.00

Note. In a dependent-samples analysis, $r_{12} = .735$ and $= s_D^2 = 3.50$.

TABLE 2.4
Results of the Independent Samples *t* Test and the Dependent Samples *t* Test for the Data in Table 2.3

Analysis	Standard error	95% CI for $\mu_1 - \mu_2$	*t*	*df*
Independent samples	1.58	−1.64–5.64	1.26[a]	8
Dependent samples	.837	.32–4.32	2.38[b]	4

Note. CI = confidence interval. For both analyses, $M_1 - M_2 = 2.00$.
[a] $p = .243$. [b] $p = .076$.

The results reported in Tables 2.2 and 2.4 show a special correspondence between 95% confidence intervals based on mean contrasts and results of the *t* test conducted with the same data at the .05 level for a nil hypothesis and a nondirectional alternative hypothesis: The confidence interval includes zero if H_0 is not rejected, but it does not include zero if H_0 is rejected. This relation is not surprising because the same basic information that goes into a confidence interval goes into a statistical test. However, much of this information is hidden if all that is reported is a test statistic and its *p* value. A mathematically sophisticated reader may be able to construct a confidence interval from the test statistic by solving for the standard error, but simply reporting the confidence interval makes this information accessible to all.

An important point should be made: Thompson (2002b) and others noted that it is erroneous to equate confidence intervals and statistical tests because of the special correspondence between them, mentioned previously. This is because the most informative use of confidence intervals compares them across different studies, not whether a particular interval includes zero. This most informative use concerns replication, something that results of statistical tests in a single study cannot address. The same author also makes the point that statistical tests cannot be conducted without a null hypothesis, but no hypothesis is required for a confidence interval. These ideas are elaborated in the next chapter.

F TESTS FOR MEANS

The *t* statistic compares only two means. Such contrasts are *focused comparisons*, and they address specific questions, such as whether treatment is superior to control. All focused comparisons are single-*df*, directional effects. The *F* statistic can also analyze focused comparisons—recall that $t^2 = F$ for a mean contrast. The *F* statistic, but not *t*, can also be used in *omnibus comparisons*, which simultaneously compare at least three means for equality. Suppose that factor A has $a = 3$ levels. The omnibus effect of

A has two degrees of freedom ($df_A = 2$), and the overall F test of this effect evaluates the following nil and nondirectional alternative hypotheses:

$$H_0\colon \mu_1 = \mu_2 = \mu_3 \qquad \text{and} \qquad H_1\colon \mu_1 \neq \mu_2 \neq \mu_3 \text{ (i.e., not } H_0\text{)}$$

Rejecting H_0 in favor of H_1 for the previous example says only that the differences among the observed means M_1, M_2, and M_3 are unlikely assuming equal population means. This result alone is often not very informative. That is, a researcher may be more interested in a series of focused comparisons, such as the contrast of the first level of A with the second (e.g., $H_0\colon \mu_1 = \mu_2$, $H_1\colon \mu_1 > \mu_2$), which break down the omnibus effect into specific directional effects. Accordingly, it is common practice to either follow an omnibus comparison with focused comparisons or forego the omnibus comparison and analyze only focused comparisons.

The logic of F as a test statistic for the omnibus comparison in a design with a single fixed factor A with two or more levels is explained next. There are separate sections about omnibus F statistics for designs with independent samples and for designs with dependent samples. The presentations for correlated designs are more technical. However, readers interested in methods for independent samples can skip the sections about correlated designs without difficulty.

Independent Samples *F* Test

The general form of the F statistic for the omnibus effect in a single-factor design with independent samples is

$$F\ (df_A,\ df_W) = \frac{MS_A}{MS_W} \tag{2.27}$$

where df_A and df_W are, respectively, the degrees of freedom for the numerator and denominator of F. The former equals the number of levels of factor A minus one, or $df_A = (a - 1)$, and the latter is the total within-groups degrees of freedom, or

$$df_W = \sum_{i=1}^{a} df_i = \sum_{i=1}^{a} (n_i - 1) = N - a \tag{2.28}$$

The numerator of F is the between-conditions (groups) mean square. Its equation is

$$MS_A = \frac{SS_A}{df_A} = \frac{\sum_{i=1}^{a} n_i\,(M_i - M_T)^2}{a - 1} \tag{2.29}$$

where SS_A is the between-conditions sum of squares, n_i and M_i are, respectively, the size and mean of the ith condition, and M_T is the mean for the total data set. The latter is the *grand mean*, the average of all N scores. The value of M_T can also be computed as the weighted average of the condition means or only in a balanced design as the arithmetic average of the condition means. Please note in this equation that group size contributes only to the numerator of the between-conditions variance, SS_A. The implication of this fact is demonstrated next.

The numerator of F, MS_A, reflects group size and sources of variability that give rise to unequal group means, such as a systematic effect of factor A or sampling error. It is the *error term* (denominator) of F, the pooled within-conditions mean square MS_W, that measures only unexplained variance. This is because cases within the same condition are treated alike, such as when patients in a treatment group are all given the same dosage of the same drug. Because drug is a constant for these patients, it cannot account for individual differences among them. The equation for the error term is

$$MS_W = \frac{SS_W}{df_W} = \frac{\sum_{i=1}^{a} df_i\,(s_i^2)}{\sum_{i=1}^{a} df_i} \tag{2.30}$$

where df_i and s_i^2 are, respectively, the degrees of freedom (i.e., $n_i - 1$) and variance of the ith group. When there are only two groups, $MS_W = s_P^2$ (Equation 2.9), and only in a balanced design can MS_W also be computed as the arithmetic average of the individual within-groups variances. Please note in this equation that group size contributes to both the numerator and denominator of MS_W, which effectively cancels out its effect on the error term of F.

The total sums of squares, SS_T, reflects the amount of variability in the total data set. It is the sum of squared deviations of the individual scores from the grand mean; it also equals $SS_A + SS_W$. We will see in later chapters that SS_T is important for effect size estimation with descriptive measures of association in essentially any comparative study where ANOVA is used.

Presented in Table 2.5 are the means and variances of three independent samples, and reported in Table 2.6 are results of the one-way F test for these data at three different group sizes, $n = 5$, 15, and 30. Observe that across all three ANOVA source tables in Table 2.6, the value of the error term is constant, $MS_W = 5.50$. The dependence of MS_A and F on group size is obvious: Both increase along with the group size, which also progressively lowers the probability values for F from $p = .429$ for $n = 5$ to $p = .006$ for $n = 30$. This change in p values occurs even though the group means and variances are constant across all analyses.

TABLE 2.5
Means and Variances for Three Independent Samples

	Group		
	1	2	3
M	13.00	11.00	12.00
s^2	7.50	5.00	4.00

TABLE 2.6
Results of the Independent Samples F Test at Three Different Group Sizes for the Data in Table 2.5

Source	SS	df	MS	F
		$n = 5$		
Between (A)	10.00	2	5.00	$.91^a$
Within (error)	66.00	12	5.50	
Total	76.00	14		
		$n = 15$		
Between (A)	30.00	2	15.00	2.73^b
Within (error)	231.00	42	5.50	
Total	261.00	44		
		$n = 30$		
Between (A)	60.00	2	30.00	5.45^c
Within (error)	478.50	87	5.50	
Total	538.50	89		

[a] $p = .429$. [b] $p = .077$. [c] $p = .006$.

Weighted- Versus Unweighted-Means Analysis

The standard F statistic described earlier is used in a *weighted-means analysis*. This is because the squared deviation of each condition mean from the grand mean is weighted by group size when MS_A, the numerator of F, is computed (Equation 2.29). If the design is unbalanced, the means from the bigger groups get a larger weight. This is not a problem if unequal study group sizes reflect unequal population group sizes. An *unweighted-means analysis* may be preferred if unequal group sizes are a result more of sampling artifacts. All means are given the same weight in this method. This is accomplished by (a) computing the grand mean as the arithmetic average instead of the weighted average of the group means and (b) substituting an average group size for the actual group sizes in the equation for MS_A. This average group size is the harmonic mean:

$$n_h = \frac{a}{\sum_{i=1}^{a} 1/n_i} \tag{2.31}$$

where n_i is the actual size of the ith group. Note that a weighted-means analysis and an unweighted-means analysis generate the same value of F in a balanced design.

Assumptions of the Independent Samples *F* Test

It is stated in many introductory statistics textbooks that the assumptions of the F test in designs with independent samples and fixed factors include independence of the observations, normal population distributions, and equal population variances. The latter is the assumption of *homogeneity of variance*, and it is necessary whenever error terms include variances averaged across different conditions (e.g., Equation 2.30). However, there are related requirements that may not be explicitly stated in introductory textbooks. These include the requirements that all levels of each fixed factor are included in the experiment and that treatments are additive and have no affect on the shapes or variances of population treatment distributions (Winer, Brown, & Michels, 1991). If a treatment is studied that is expected to affect both the average level and variability of cases, the latter requirement may be violated. Altogether these requirements are more restrictive than many researchers realize.

The p values from the F test are computed under these assumptions. If these assumptions are violated, then the p values of these tests—and decisions based on them, namely whether to reject H_0—may not be accurate. If observed p values wind up being too low because of violation of assumptions, there is a *positive bias* because H_0 is rejected more often than it should. In the RS context for statistical tests, this implies that the researcher's hypothesis is supported more often than it should be. It can also happen that observed p values can be too high because of violation of assumptions, which may reduce power.

There is a relatively large literature about the consequences of violating the assumptions of statistical tests in general and the F test in particular in fixed-effects ANOVA. It is beyond the scope of this section to review this literature in detail, so only an overview is presented; readers are referred to Winer et al. (1991, pp. 100–110) and a review article by Glass, Peckham, and Sanders (1972) for more information. The independence assumption is critical because nonindependence can seriously affect both p values and power regardless of whether the group sizes are equal or unequal. This requirement should generally be seen as an essential property of the research design. It was believed that the normality assumption is generally unimportant in that it can be violated with relative impunity with little effect on p values or power. It was also believed that the F test is relatively insensitive to variance heterogeneity. However, recent work by Wilcox (1987, 1998) and others indicate that (a) even relatively small departures from normality can sometimes distort the results of the standard t or F tests; (b) there can be serious positive bias in these tests when the ratio of the largest over the smallest within-groups variance is 9 or greater; and (c) the degree of inaccuracy in p values may be greater when the group sizes are small and unequal or when heterogeneity is associated with one outlier group than when it is spread across all the groups (Keppel, 1991). There are versions of both the t and F tests for independent samples that do not assume normality or homogeneity of variance (e.g., Winer et al., 1991, pp. 66–69; Wilcox, 1987), but they are not used nearly as often as the standard t and F tests based on these assumptions.

Reviewed in the next chapter is evidence that the assumptions of t, F, and other statistical tests seem to be infrequently met or evaluated in applied behavioral research. This is another serious shortcoming of the use of statistical tests in the behavioral sciences.

Dependent Samples *F* Test

The between-conditions variance, MS_A, and the pooled within-conditions variance, MS_W, are computed the same way regardless of whether the samples are independent or dependent (Equations 2.29–2.30). However, the latter no longer reflects just unexplained variance when the samples are dependent, so it is not the error term for the omnibus F statistic in a correlated design. This is because of the subjects effect, which in a one-way design is manifested in positive covariances between every pair of conditions. When the independent variable has three or more levels, the average covariance across all pairs of conditions, M_{cov}, estimates the subjects effect for the whole design. Removing this effect from the pooled within-conditions variance (literally, $MS_W - M_{cov}$) gives the error term for F in a one-way design with dependent samples. This error term reflects inconsistent performance across the conditions. This inconsistency may be a result of random variation or to a nonadditive effect, which means that the independent variable does not have the same relative impact on all cases. In other words, there is some characteristic of participants that moderates the effect of factor A, either amplifying or diminishing it. For example, a drug may be more effective for older patients than younger patients. This moderator effect is also known as a *person × treatment interaction*.

In an *additive model*, which assumes no true person × treatment interaction, the error term of the dependent samples F statistic is presumed to reflect only random error. In some sources, this error term is designated as MS_{res}, where the subscript refers to residual variance. However, an additive model is probably unrealistic for many within-subjects factors in the behavioral sciences. A *nonadditive model* assumes a true person × treatment interaction, and the error term in this model may be called $MS_{A \times S}$, where the subscript reflects this assumption. This notation is used later. Unfortunately, it is not possible to separately estimate variability because of random error versus a true person × treatment interaction when each case is measured once in each condition, which is typical in one-way within-subjects designs. This implies that $MS_{res} = MS_{A \times S}$ in the same data set, so the distinction between them is more conceptual than practical. Cases in which the assumption of additive versus nonadditive models makes a difference in effect size estimation are considered in chapter 6.

We can now define the general form of the omnibus F test for single-factor designs with dependent samples assuming a nonadditive model:

$$F\ (df_A,\ df_{A \times S}) = \frac{MS_A}{MS_{A \times S}} \tag{2.32}$$

where $df_{A \times S} = (a - 1)\ (n - 1)$ and $MS_{A \times S} = MS_W - M_{cov}$. The latter can also be expressed as

$$MS_{A \times S} = \frac{SS_{A \times S}}{df_{A \times S}} = \frac{SS_W - SS_S}{df_W - df_S} \tag{2.33}$$

where SS_S is the sum of squares for the subjects effect with $df_S = n - 1$ degrees of freedom. Equation 2.33 shows the removal of the subjects effect from the pooled within-conditions variability in a correlated design, which is the same basic subtraction that generates the error term of the dependent samples t statistic (Equation 2.15). Equation 2.33 also shows the decomposition of the total within-conditions sums of squares into two parts, one because of the subjects effect and the other associated with the error term, or $SS_W = SS_S + SS_{A \times S}$.

The F test for dependent samples has the same potential power advantage over the F test for independent samples as the t test for dependent samples has over its independent samples counterpart. This is demonstrated with the data for three samples presented in Table 2.7. The results of two different F tests conducted with these data are reported in Table 2.8. The first analysis assumes $n = 5$ cases in each of three independent samples, and the second analysis assumes $n = 5$ triads of scores across three dependent samples. Only the second analysis takes account of the positive correlations between each pair of conditions, which range from .730 to .839 (Table 2.7).

TABLE 2.7
Raw Scores and Descriptive Statistics for Three Samples

	Condition		
	1	2	3
	9	8	10
	12	12	11
	13	11	13
	15	10	11
	16	14	15
M	13.00	11.00	12.00
s^2	7.50	5.00	4.00

Note. In a dependent samples analysis, $r_{12} = .735$, $r_{13} = .730$, and $r_{23} = .839$.

Observe the higher F and lower p values for the dependent samples analysis even though the group means and variances are constant.

Assumptions of the Dependent Samples F Test

The same assumptions of the independent samples F test—independence, normality, and homogeneity of variance—apply to the dependent samples F test. However, there are additional assumptions that concern the correlations between multiple measures obtained from the same cases (or sets of matched cases). When a within-subjects factor has at least three levels, these assumptions are relatively complicated and quite difficult to meet in practice. An additional requirement is that of *sphericity* or *circularity*, which assumes that the population variances of the difference scores between every pair of conditions are equal. This assumption is critical

TABLE 2.8
Results of the Independent Samples F Test and the Dependent Samples F Test for the Data in Table 2.7

Source	*SS*	*df*	*MS*	*F*
	Independent samples analysis			
Between (*A*)	10.00	2	5.00	.91[a]
Within (error)	66.00	12	5.50	
Total	76.00	14		
	Dependent samples analysis			
Between (*A*)	10.00	2	5.00	3.53[b]
Within	66.00	12	5.50	
Subjects (*S*)	54.67	4	13.67	
$A \times S$ (error)	11.33	8	1.42	
Total	76.00	14		

[a] $p = .429$. [b] $p = .080$.

because even relatively minor violations of this assumption may lead to rejecting H_0 too often (i.e., resulting in a positive bias). There are statistical tests intended to detect departure from sphericity, but they have been criticized for restrictive assumptions of their own, such as normality. Some methodologists suggest that the sphericity requirement may not be tenable in most behavioral studies and that researchers should direct their efforts to controlling bias (Keppel, 1991). There are basically five options for dealing with the sphericity assumption that are briefly summarized next; see H. Keselman, Algina, and Kowalchuk (2001), Max and Onghena (1999), or Winer et al. (1991, pp. 239–273) for more information:

1. Assume maximal violation of sphericity, compute F in the usual way, but compare it against a higher critical value. This critical F has only 1 and $n - 1$ degrees of freedom; the standard critical F for comparing a dependent means has $a - 1$ and $(a - 1)$ $(n - 1)$ degrees of freedom. This method has been called the *Geisser–Greenhouse conservative test* or the *Geisser–Greenhouse correction*.
2. Estimate the degree of departure from sphericity with a statistic called estimated epsilon, $\hat{\varepsilon}$. This statistic ranges from $1/(a - 1)$, which indicates maximal departure to 1.00, which in turn indicates no violation of sphericity. The degrees of freedom for the critical value for F are then taken as $\hat{\varepsilon}$ $(a - 1)$ and $\hat{\varepsilon}$ $(a - 1)(n - 1)$, which makes the test more conservative for <1.00. There are somewhat different forms of $\hat{\varepsilon}$ that may be called the Box correction, the Geisser–Greenhouse epsilon, or the Huynh–Feldt epsilon.
3. Conduct focused comparisons between pairs of condition means instead of the omnibus comparison This implies that each contrast has its own specialized error term (i.e., it is not $MS_{A \times S}$ for the whole design). Because these unique error terms are based on data from only two conditions, the sphericity requirement does not apply.
4. Analyze data from all levels of the factor with multivariate analysis of variance (MANOVA), which also does not assume sphericity. In this approach, difference scores between adjacent levels of factor A are analyzed as multiple, correlated outcome variables (e.g., Stevens, 1992, chap. 13).
5. Use the statistical resampling method of bootstrapping to generate an empirical F test for repeated measures data. (Bootstrapping as an alternative to traditional statistical tests is discussed in chapter 9, this volume). In a recent Monte Carlo analysis, Berkovits, Hancock, and Nevitt (2000) found that

this method is relatively robust against violation of the sphericity assumption.

All of these options are concerned in large part with the estimation of accurate p values in correlated designs. Considering the limitations of p values outlined in the next chapter, perhaps an even better choice is to move away from traditional statistical tests to model-fitting techniques suitable for repeated-measures data, such as structural equation modeling or hierarchical linear modeling, among others. This point is elaborated later.

Analysis of Variance as Multiple Regression

All forms of ANOVA are nothing more than special cases of multiple regression (MR), which itself is just an extension of bivariate regression that analyzes one or more predictors of a continuous dependent variable. These predictors can be either continuous or categorical variables. Categorical predictors are represented in regression equations with special codes that each correspond to a single *df* contrast (i.e., a focused comparison) between the levels of that predictor. It is also possible in MR to estimate interaction effects between continuous or categorical predictors. In theory, one needs only a software program for MR to conduct any kind of ANOVA. The advantage of doing so is that the output of regression programs usually includes correlations, partial correlations, or standardized regression coefficients (beta weights), all of which are standardized measures of effect size. In contrast, software programs for ANOVA may print only source tables, and the F and p values in these tables measure both effect size and sample size (e.g., Table 2.6). The disadvantage of using MR programs instead of ANOVA programs is that the coding required for some kinds of designs, especially ones with repeated-measures factors, can become complicated. In contrast, ANOVA programs are typically easier to use because no special coding of the factors is required by the user. Fortunately, there are some straightforward ways to extract information about effect size from ANOVA source tables that are demonstrated in later chapters.

Entire books are written about the relation between ANOVA and MR (e.g., Keppel & Zedeck, 1989), so it is not possible to deal with this issue in substantive detail. However, readers should be aware of this alternative to using ANOVA to analyze means and estimate effect size.

χ^2 TEST OF ASSOCIATION

Whether there is a statistical relation between two categorical variables is the question addressed by the χ^2 test of association. A two-way contingency

TABLE 2.9
Results of the Chi-Square Test of Association for the Same Proportions at Different Group Sizes

		Outcome			
		Observed Frequencies			
Group	n	Recovered	Not recovered	Recovery rate	χ^2 (1)
		$n = 40$			
Treatment	40	24	16	.60	3.20[a]
Control	40	16	24	.40	
Total	80	40	40		
		$n = 80$			
Treatment	80	48	32	.60	6.40[b]
Control	80	32	48	.40	
Total	160	80	80		

[a] $p = .074$. [b] $p = .011$.

table summarizes the data analyzed by this test. Presented in the top half part of Table 2.9 is a 2 × 2 cross-tabulation that shows the frequencies of treatment and control cases (n = 40 each) that either recovered or did not recover. A total of 24 cases in the treatment group recovered, or 60%. Among control cases, 16 cases recovered, or 40%. The recovery rate among treated cases is thus 20% higher than among untreated cases.

The χ^2 test of association for two-way contingency tables takes the form

$$\chi^2 (r - 1, c - 1) = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{o_{ij}} - f_{e_{ij}})^2}{f_{e_{ij}}} \qquad (2.34)$$

where the degrees of freedom are the product of the number of rows (r) minus one and the number of columns (c) minus one, $f_{o_{ij}}$ is the observed frequency for the cell in the ith row and jth column, and $f_{e_{ij}}$ is the expected frequency for the same cell under the nil hypothesis that the two variables are unrelated. There is a quick way to derive by hand the value of f_e for any cell: Divide the product of the row and column totals for that cell by the total number of cases, N. It is that simple. The statistical assumptions of the χ^2 test of association include independence of the observations, classification of each observation into one and only one category (i.e., contingency table cell), and a sample size large enough so that the minimum expected value across the cells is about 5 for tables with more than a single degree of freedom or about 10 for tables with a single degree of freedom.

For the 2 × 2 cross-tabulation in the top half of Table 2.9, the expected frequency for each cell is $f_e = (40 \times 40)/80 = 20$. This shows a pattern where outcome is unrelated to treatment status because the expected recovery rate is the same for both groups, 50% (20/40). After application of this equation, the results are χ^2 (1) = 3.20, p = .074, so the nil hypothesis that group membership and recovery status are unrelated is not rejected at the .05 level. The effect of increasing the group size but keeping all else constant on the χ^2 test is demonstrated in the bottom part of Table 2.9. Reported there are results of the χ^2 test for the same proportions but a larger group size, n = 80. The null hypothesis is now rejected at the .05 level—χ^2 (1) = 6.40, p = .011—even though the improvement in recovery rate for treated versus control cases is unchanged, 20%.

Other common applications of the χ^2 test not described include a goodness-of-fit test for categorical variables and a test for correlated proportions, among others. All tests just mentioned are also sensitive to sample size.

STATISTICAL TESTS AND REPLICATION

Statistical tests provide a framework for making a dichotomous decision—reject or fail to reject H_0—about sample results in the face of uncertainty. This uncertainty is sampling error, which is estimated in some way by essentially all statistical tests. Of course, any decision based on a statistical test may not be correct (e.g., a Type I or Type II error). In any science, though, it is replication that is the ultimate arbiter: No matter how intriguing a result from a single study, it must be replicated before it can be taken seriously. Replication also is the ultimate way to deal with the problem of sampling error. Indeed, statistical tests are unnecessary with sufficient replication.

There is a much stronger tradition of replication in the natural sciences than in the social sciences. It is also true that statistical tests are infrequently used in the natural sciences. Whether this association is causal is a matter of debate. Some authors argue that the widespread use of statistical tests in the social sciences works against the development of a stronger appreciation for replication (e.g., Kmetz, 2000; F. Schmidt & Hunter, 1997). There are probably other factors that contribute to the difference in emphasis on replication across the social and natural sciences (Kupfersmid, 1988), but the possibility that statistical tests is one of them warrants careful consideration. Replication and meta-analysis as a method for research synthesis are considered in chapter 8.

CONCLUSION

Outlined in this chapter is a basic vocabulary for comparative studies and the logic of interval estimation for statistics with simple distributions, such as means, versus those with complex distributions, such as some effect size indexes. Confidence intervals for the former are constructed with central test statistics, but the latter may require noncentral test statistics. Special software tools are also typically needed for noncentral confidence intervals. A confidence interval based on a statistic sets a reasonable lower and upper bounds for the corresponding population parameter, but there is no guarantee that the value of the parameter is included in a particular confidence interval. The essential logic of statistical tests in general and characteristics of the t and F tests for means and the χ^2 test for two-way contingency tables in particular was also reviewed. Any statistical test measures both effect size and sample size. This is why neither the values of test statistics or their associated probabilities say much useful about effect size. Additional limitations of statistical tests are considered in the next chapter.

RECOMMENDED READINGS

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532–574.

Reichardt, C. S., & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical tests, and vice versa. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Mahwah, NJ: Erlbaum.

Smithson, M. J. (2000). *Statistics with confidence: An introduction for psychologists.* Thousand Oaks, CA: Sage.