Chapter 9

Noncentrality Interval Estimation and the Evaluation of Statistical Models

James H. Steiger
University of British Columbia

Rachel T. Fouladi
University of Texas at Austin

Noncentrality-based confidence interval estimates provide a superior alternative to significance testing for assessing model fit in most standard areas of behavioral statistics, from the t test through multiple regression and analysis of variance to the analysis of covariance structures. These confidence intervals provide all the information inherent in a significance test, and more, and deal with situations more traditional interval estimates cannot handle. For example, in the analysis of variance, noncentrality interval estimation allows computation of exact confidence intervals for (a) standardized measures of effect size and (b) statistical power. In multiple regression, one can compute an exact confidence interval on the squared multiple correlation. Because of computational complexities, noncentrality-based confidence intervals seldom have been computed, except in the analysis of covariance structures. Most of the reasons for not using these interval estimates are no longer relevant in the microcomputer age. In this chapter, we review some of the standard techniques, and provide computational examples.

Behavioral statistics has been, and continues to be, dominated by the significance testing tradition. Nearly every major textbook in behavioral statistics spends far more time and energy on the theory and mechanics of significance testing than on any other topic. Periodically, some of the more authoritative writers in our field have questioned this. The list of names includes many (Cohen, Meehl, Guttman, Rozeboom, to name just a few) who have imposing reputations for technical expertise, but also share a common reputation for *perspective*, manifested in an ability to sort out what is important and what is right.

Some important early contributions to the literature on hypothesis testing and interval estimation were reviewed recently by Cohen (1994). Most of the authors, including Cohen, concentrate on the fundamental logical problems and

limitations of significance testing, and make broad suggestions for improving the status of practice. A key suggestion that has surfaced repeatedly in these writings is that, as an analytic tool, the *confidence interval* is superior to the significance test. Schmidt and Hunter (1997) echo this view in chapter 3 of this volume, while providing a succinct critique of many of the arguments often used to defend significance testing. Because we agree fundamentally with many of the opinions of Rozeboom (1960), Meehl (1978), Guttman (1977), Cohen (1994), and Schmidt and Hunter (1997), we see no need to review all of their arguments here; but to keep the account relatively self-contained, we review a few key advantages of confidence intervals. Our fundamental contribution, however, is to suggest a significant change in the statistical methodology routinely employed in the most common situations in behavioral statistics. We suggest improved techniques that we think have real merit, and that, if given wide use, offer such substantial advantages that they will surely accelerate the ascendancy of interval estimation. Some of these implementations (such as the noncentrality-based techniques in structural modeling) are relatively new but already quite popular. Some are old, but hardly ever discussed in textbooks. All offer substantial advantages over the significance tests, and in some cases over other interval estimates currently in use. All are computer intensive, and require *very* careful software implementation. This latter fact explains why they have seldom been employed, but why their time finally may have arrived.

We begin by reviewing a situation familiar to us all—the simple two-group experiment based on two independent samples. We review the standard interval estimation procedures, then discuss an alternative *standardized* confidence interval discussed by Hedges and Olkin (1985), but not computationally practical at the time their book was written. We show why this interval estimation approach is superior, not only to the standard *t* test, but also to the standard confidence interval on mean differences discussed in textbooks. We then extend this idea through the analysis of variance, to the analysis of covariance structures, to multiple regression and beyond. So our scope is quite broad. Almost all the significance testing procedures currently recommended in major behavioral statistics books could be replaced with the superior confidence interval approaches we discuss here.

USE AND ABUSE OF SIGNIFICANCE TESTING LOGIC

In this section, we argue, as do numerous colleagues, that significance tests, though almost always reported in the analysis of social science data, are seldom to be preferred, and often simply inappropriate. We begin by returning briefly to first principles. Suppose we are performing a simple two-group experiment in

TABLE 9.1
2×2 Table for Statistical Decisions

		State of the World	
		H_0	H_1
Decision	H_0	Correct Acceptance	Type II Error β
	H_1	Type I Error α	Correct Rejection

which an experimental group is compared to a control group. The theoretical question of interest is frequently phrased as, "Has the experimental treatment made any difference?"

In this case, the statistical null and alternative hypotheses are

$$H_0: \mu_1 = \mu_2 \qquad H_1: \mu_1 \neq \mu_2.$$

We test this hypothesis, in practice, by taking two samples, often (but not necessarily) of equal size, and computing a two (independent) sample Student's *t* statistic. If the statistic's absolute value is sufficiently large, we reject H_0. Otherwise, loosely speaking, we "accept" (or, perhaps more appropriately, "fail to reject") H_0.

Back in our undergraduate statistics course, we were taught that, in the significance testing approach, four things can happen, two of them bad. We all memorized a little 2×2 table that summarized the possibilities and attached statistical jargon to them (See Table 9.1).

Most of us were steeped in the grand tradition of Educational and Psychological Statistics, i.e., that α, the Type I error rate, must be kept at or below .05, and that, if at all possible, β, the Type II error rate, must be kept low as well.

The conventions are, of course, much more rigid with respect to α than with respect to β. Seldom, if ever, is α allowed to stray above the magical .05 mark. Let's review where that tradition came from.

In the context of significance testing, we can define two basic kinds of situations, reject-support (RS) and accept-support (AS). In RS testing, *the null hypothesis is the opposite of what the researcher actually believes*, and rejecting it supports the researcher's theory. In a two group RS experiment, the experimenter believes the treatment has an effect, and seeks to confirm it through a significance test that rejects the null hypothesis.

In the RS situation, a Type I error represents, in a sense, a "false positive" for the researcher's theory. From society's standpoint, such false positives are particularly undesirable. They result in much wasted effort, especially when the false positive is interesting from a theoretical or political standpoint (or both), and as a result stimulates a substantial amount of research. Such follow-up research will usually not replicate the (incorrect) original work, and much confusion and frustration will result.

In RS testing, a Type II error is a tragedy from the researcher's standpoint, because a theory that is true is, by mistake, not confirmed. So, for example, if a drug designed to improve a medical condition is found (incorrectly) not to produce an improvement relative to a control group, a worthwhile therapy will be lost, at least temporarily, and an experimenter's worthwhile idea will be discounted.

As a consequence, in RS testing, society, in the person of journal editors and reviewers, insists on keeping α low. The statistically well-informed researcher makes it a top priority to keep β low as well. Ultimately, of course, everyone benefits if *both* error probabilities are kept low, but unfortunately there is often, in practice, a trade-off between the two types of error.

The RS situation is by far the more common one, and the conventions relevant to it have come to dominate popular views on statistical testing. As a result, the prevailing views on error rates are that relaxing α beyond a certain level is unthinkable, and that it is up to the researcher to make sure statistical power is adequate. One might argue how appropriate these views are in the context of RS testing, but they are not altogether unreasonable.

In AS testing, the common view on error rates we described above is clearly inappropriate. In AS testing, H_0 *is what the researcher actually believes*, so accepting it supports the researcher's theory. In this case, a Type I error is a false negative for the researcher's theory, and a Type II error constitutes a false positive. Consequently, acting in a way that might be construed as highly *virtuous* in the RS situation, for example, maintaining a very low Type I error rate like .001, is actually "stacking the deck" in favor of the researcher's theory in AS testing.

In both AS and RS situations, it is easy to find examples where significance testing seems strained and unrealistic. Consider first the RS situation. In some such situations, it is simply not possible to have very large samples. An example that comes to mind is social or clinical psychological field research. Researchers in these fields sometimes spend several days interviewing a single subject. A year's research may only yield valid data from 50 subjects. Correlational tests, in particular, have very low power when samples are that small. In such a case, it probably makes sense to relax α beyond .05, if it means that reasonable power can be achieved.

On the other hand, it is possible, in an important sense, to have power that is too high. For example, one might be testing the hypothesis that $\mu_1 = \mu_2$ with sample sizes of a million in each group. In this case, even with trivial differences between groups, the null hypothesis would virtually always be rejected.

The situation becomes even more unnatural in AS testing. Here, if n is too high, the researcher almost inevitably decides against the theory, even when it turns out, in an important sense, to be an excellent approximation to the data. It seems paradoxical indeed that in this context experimental precision seems to work against the researcher.

To summarize, in RS research:

1. The researcher wants to reject H_0.
2. Society wants to control Type I error.
3. The researcher must be very concerned about Type II error.
4. High sample size works for the researcher.
5. If there is "too much power," trivial effects become "highly significant."

In AS research:

1. The researcher wants to accept H_0.
2. "Society" should be worrying about controlling Type II error, although it sometimes gets confused and retains the conventions applicable to RS testing.
3. The researcher must be very careful to control Type I error.
4. High sample size works against the researcher.
5. If there is "too much power," the researcher's theory can be "rejected" by a significance test even though it fits the data almost perfectly.

Strictly speaking, the outcome of a significance test is the dichotomous decision whether or not to reject the null hypothesis. This dichotomy is inherently dissatisfying to psychologists and educators, who frequently use the null hypothesis as a statement of no effect, and are more interested in knowing how big an effect is than whether it is (precisely) zero. This has led to behavior like putting one, two, or three asterisks next to results in tables, or listing p levels next to results, when, in fact, such numbers, across (or sometimes even within!) studies need not be monotonically related to the best estimates of strength of experimental effects, and hence can be extremely misleading. Some writers (e.g., Guttman, 1977) view asterisk-placing behavior as inconsistent with the foundations of significance testing logic.

Probability levels can deceive about the "strength" of a result, especially when presented without supporting information. For example, if, in an ANOVA

table, one effect had a p level of .019, and the other a p level of .048, *it might be an error* to conclude that the statistical evidence supported the view that the first effect was stronger than the second. A meaningful interpretation would require additional information. To see why, suppose someone reports a p level of .001. This *could* be representative of a trivial population effect combined with a huge sample size, or a powerful population effect combined with a moderate sample size, or a huge population effect with a small sample. Similarly a p level of .075 *could* represent a powerful effect operating with a small sample, or a tiny effect with a huge sample. Clearly then, we need to be careful when comparing p levels.

In AS testing, which occurs frequently in the context of model fitting in factor analysis or "causal modeling," significance testing logic is basically inappropriate. Rejection of an "almost true" null hypothesis in such situations frequently has been followed by vague statements that the rejection shouldn't be taken too seriously. Failure to reject a null hypothesis usually results in a demand for cumbersome power calculations by a vigilant journal editor. Such problems can be avoided by using confidence intervals.

THE VALUE OF INTERVAL ESTIMATES

Much psychological research is exploratory. The fundamental questions we are usually asking are "What is our best guess for the size of the population effect?" and "How precisely have we determined the population effect size from our sample data?" Significance testing fails to answer these questions directly. Many a researcher, faced with an "overwhelming rejection" of a null hypothesis, cannot resist the temptation to report that it was "significant *well beyond* the .001 level." Yet we have seen previously (and demonstrate conclusively with numerical examples in a subsequent section) that a p level following a significance test can be a poor vehicle for conveying what we have learned about the strength of population effects.

Confidence interval estimation provides a convenient alternative to significance testing in most situations. Consider the 2-tailed hypothesis of no difference between means. Recall first that the significance test rejects at the α significance level if and only if the $1 - \alpha$ confidence interval for the mean difference excludes the value zero. Thus the significance test can be performed with the confidence interval. Most undergraduate texts in behavioral statistics show how to compute such a confidence interval. The interval is exact under the assumptions of the standard t test. However, the confidence interval contains information about experimental precision that is not available from the result of a significance test. Assuming we are reasonably confident about the metric of the
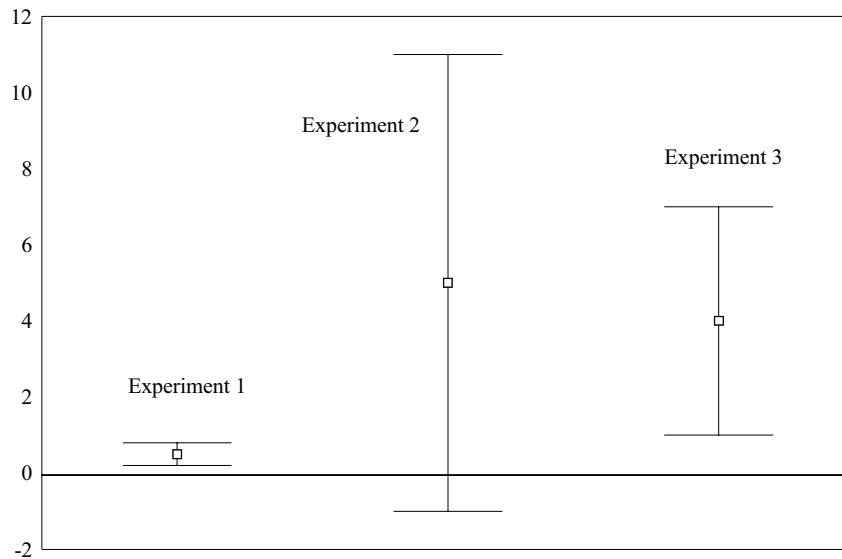
FIGURE 9.1 Confidence intervals reflecting different degrees of precision
of measurement.

data, it is much more informative to state a confidence interval on $\mu_1 - \mu_2$ than
it is to give the p level for the t test of the hypothesis that $\mu_1 - \mu_2 = 0$. In sum-
mary, we might say that, in general, a confidence interval conveys more infor-
mation, in a more naturally usable form, than a significance test. This is seen
most clearly when confidence intervals from several studies are graphed along-
side one another, as in Figure 9.1.

Figure 9.1 shows confidence intervals for the difference between means for 3
experiments, all performed in the same domain, using measures with approxi-
mately the same variability. Experiments 1 and 3 yield a confidence interval that
fails to include zero. For these experiments, the null hypothesis was rejected.
The second experiment yields a confidence interval that includes zero, so the
null hypothesis of no difference is not rejected. A significance testing approach
would yield the impression that the second experiment did not agree with the
first and the third.

The confidence intervals suggest a different interpretation, however. The first
experiment had a very large sample size, and very high precision of measure-
ment, reflected in a very narrow confidence interval. In this experiment, a small
effect was found, and determined with such high precision that the null hypothe-
sis of no difference could be rejected at a stringent significance level.

The second experiment clearly lacked precision, and this is reflected in the very wide confidence interval. Evidently, the sample size was too small. It may well be that the actual effect in conditions assessed in the second experiment was larger than that in the first experiment, but the experimental precision was simply inadequate to detect it.

The third experiment found an effect that was statistically significant, and perhaps substantially higher than the first experiment, although this is partly masked by the lower level of precision, reflected in a confidence interval that, though narrower than Experiment 2, is substantially wider than Experiment 1.

Suppose the 3 experiments involved testing groups for differences in IQ. In the final analysis, we may have had *too much power* in Experiment 1, as we are declaring "highly significant" a rather miniscule effect substantially less than a single IQ point. We had far too little power in Experiment 2. Experiment 3 seems about right.

Many of the arguments we have made on behalf of confidence intervals have been made by other authors as cogently as we have made them here. Yet, confidence intervals are seldom reported in the literature. Most important, as we demonstrate in the succeeding sections, there are several extremely useful confidence intervals that virtually *never* are reported. In what follows, we discuss *why* the intervals are seldom reported, *how* they can be computed, and *where* software performing all these techniques may be obtained.

REASONS WHY INTERVAL ESTIMATES ARE SELDOM REPORTED

In spite of the obvious advantages of interval estimates, they are seldom employed in published articles in psychology. On those infrequent occasions when interval estimates are reported, they are often not the optimal ones. There are several reasons for this status quo:

1. *Tradition.* Traditional approaches to psychological statistics emphasize significance testing much more than interval estimation.

2. *Pragmatism.* In RS situations, interval estimates are sometimes embarrassing. When they are narrow but close to zero, they suggest that a "highly significant" result may be statistically significant but trivial. When they are wide, they betray a lack of experimental precision.

3. *Ignorance.* Many people are simply unaware of some of the very valuable interval estimation procedures that are available. For example, the vast majority of psychologists are simply not aware that it is possible to compute a confidence interval on the squared multiple correlation coefficient.

The procedure is not discussed in standard texts, and it is not implemented in major statistical packages.

4. *Lack of availability.* Some of the most desirable interval estimation procedures are computer intensive, and are not implemented in major statistical packages like SAS, SPSS, STATISTICA, and so on. This makes it unlikely that anyone will try the procedure.

CONFIDENCE LIMITS, CONFIDENCE INTERVALS, AND THE INVERSION APPROACH TO INTERVAL ESTIMATION

In this section, we review the basic definition of a confidence interval, and the simple approach used to generate the simple confidence intervals found in most textbooks. Then we describe the less conventional, more computer-intensive approach which allows much more interesting and useful intervals to be derived. Here the discussion becomes somewhat more technical, and we employ notations that are common in mathematical statistics texts, but that the typical reader with a basic background in introductory applied statistics texts may find slightly intimidating. We try to strike a balance that provides sufficient, but not extraneous, detail. To begin, suppose we have a sample on n independent observations from some population. The "observations" can be individual numbers (e.g., measuring the heights of n people) or lists of numbers (measuring the height, weight, and age of n people). Suppose we use the letter X to stand for the data. A "statistic" is any function of the numbers in X. We can refer to statistics generically by using the standard mathematical notation for functions. So, for example, if we wish to discuss "statistics calculated on X" in very general terms, we could use a notation like $A(X)$. One can calculate numerous different statistics on the same data. For example, the sample mean of the heights would be one function, the correlation between height and weight another.

A common problem in statistics is to try to put limits on the value of an unknown parameter on the basis of fallible data. (We use the term *parameter* in the broad sense to refer to some numerical characteristic of a statistical population, as opposed to the strict sense, i.e., a formal argument of a probability distribution function.) For example, a politician might wish to estimate, on the basis of a modest opinion poll, the maximum level of support he or she is likely to receive in an upcoming election. *Confidence limits* and *confidence intervals* are techniques that frequently are employed to construct such limits.

An *upper confidence limit* (or *upper confidence bound*) is a statistic that, over repeated samples of size n, exceeds an unknown parameter θ a certain proportion of the time. For example, function $B(X)$ is a $1 - \alpha$ upper confidence limit

for θ if, over repeated samples, the probability that $B(X)$ is greater than or equal to θ is equal to $1 - \alpha$, or, in mathematical notation,

$$\Pr(B(X) \geq \theta) = 1 - \alpha. \qquad (9.1)$$

The basic reason behind the use of an upper confidence limit is to arrive at a number that one is quite confident exceeds the parameter. Note, one can seldom if ever be absolutely sure a statistic is greater than the unknown parameter, because the data may, through bad luck, be extremely unrepresentative of the population. Think of α in the preceding expression as an error rate. Suppose, for example, it is .05, and so $1 - \alpha = .95$. Then the preceding equation says that, if one takes a sample of data X and computes the statistic $B(X)$, it will, in the long run, be above the parameter value with probability .95, or 95% of the time. If $B(X)$ is used as a "statistical upper bound" for the parameter, it will be wrong about 5% of the time. It is common, after computing $B(X)$, to say that one is "95% confident that θ is below $B(X)$," or that one is "95% confident that θ does not exceed $B(X)$." To see why this might be useful, consider the opinion poll discussed earlier. Suppose the 95% upper confidence limit on the proportion of people intending to vote for the candidate is .65. The pollster could report back to the politician that "we are 95% confident your current support level is no greater than 65%." As another example, suppose an item is manufactured, and the parameter θ of interest is the failure rate for the item. The goal is to be reasonably certain that the failure rate is below a certain value.

In this case, one would frequently perform "reliability testing," by taking a sample X and computing an upper confidence limit for θ, the proportion of items that are defective. Suppose the upper limit is .001, or .1%. Then you might say "I am 95% confident that the defect rate is less than or equal to .1%."

Similar situations exist when one is establishing lower boundaries for a parameter, in which case *lower confidence limits* are computed.

A *lower confidence limit* (or *lower confidence bound*) is a statistic that is less than the unknown parameter a certain proportion of the time. A function $A(X)$ of the observed data X is a $1 - \alpha$ lower confidence limit for θ if, over repeated samples,

$$\Pr(A(X) \leq \theta) = 1 - \alpha. \qquad (9.2)$$

For example, a pollster might report to a politician that "I am 95% confident your support level is no worse than 47%." The problem with confidence limits is that they provide, by themselves, no indication of precision of measurement. In general, the less authoritative your database, the further you have to move a

confidence limit up (in the case of an upper limit) or down (in the case of a lower limit) in order to bound the parameter reliably. Returning to the political opinion poll, if the sample is moderately large, the pollster might report a lower limit of 44%, whereas in the same situation if the sample is quite small, the lower limit might have to be reported as 35% in order to gain the same degree of confidence. So, in order to be 95% confident, the political pollster would have to report a "minimum level of support" that is unduly pessimistic. Consequently, upper and lower confidence limits usually are combined to yield a *confidence interval* $(A(X), B(X))$, whose endpoints surround the parameter θ a certain proportion of the time. $A(X)$ and $B(X)$ bound a $1 - \alpha$ confidence interval for θ if, over repeated samples,

$$\Pr\big(A(X) \le \theta \le B(X)\big) = 1 - \alpha. \tag{9.3}$$

In practice, one usually constructs the confidence interval by choosing $A(X)$ and $B(X)$ to be, respectively, lower and upper $1 - \alpha/2$ confidence limits so that the confidence interval is symmetric.

The advantage of a confidence interval is that the width of the interval provides a ready indication of precision of measurement. That is, if the sample estimate has low sampling variability and high precision of estimate, then even a narrow confidence interval will bracket the true parameter a high percentage of the time, over repeated samples. Thus, the outcome of the confidence interval calculation is a report of a parameter value, together with an indication of how precisely it has been determined. In many situations involving exploratory research, this outcome more accurately reflects what an experimenter is hoping to learn from the data than a significance test does. So, for example, if the pollster reports "I am 95% confident that your support level is between 46% and 54%," the politician realizes that the election is up for grabs and that the support level is roughly 50% give or take 4%. This is probably more useful to the politician than being told that "a test of the hypothesis that your support level is 50% was not rejected." The politician is not really interested in whether the support level is exactly 50%—there is something artificial about testing for the significance of such a hypothesis. Rather, the key interest is in the best estimate of the support level, and how precise that estimate is. The location of the confidence interval, and its width, provide such information. If the politician feels that level of precision is inadequate, the pollster can report (based on statistical theory) that halving the width of the confidence interval will require quadrupling the size of the opinion poll!

Most confidence intervals discussed in standard textbooks are derived by simple manipulation of a statement about interval probability of a sampling dis-

tribution. For example, confidence intervals are usually introduced in terms of a simple Z statistic for testing the hypothesis $\mu = a$ when the population distribution is normal and the population standard deviation σ is known. If n is the sample size, and $\overline{X}$ is the ordinary sample mean based on the n observations, then the sampling distribution of the sample mean is normal, with a mean of μ, and a standard deviation (usually called the "standard error of the mean") of $\sigma/\sqrt{n}$. In any normal distribution, the probability that a score will fall between standard score values of -1.96 and $+1.96$ is .95. To convert the sample mean to its standard score equivalent, one subtracts its mean (μ) and divides by its standard deviation ($\sigma/\sqrt{n}$) to construct a test statistic Z,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

(9.4)

The resulting test statistic, in the long run, will fall between the 2.5% and 97.5% points of the standard normal curve (values of -1.96 and $+1.96$) with probability .95. As an inequality, these facts can be stated

$$\Pr\left(-1.96 \leq Z \leq +1.96\right) = .95,$$

(9.5)

or

$$\Pr\left(-1.96 \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq +1.96\right) = .95.$$

(9.6)

The confidence interval for μ is derived by manipulating this interval algebraically. Because σ and n are both positive, we may multiply all three sections of the inequality by $\sigma/\sqrt{n}$ without altering its correctness. We then obtain

$$\Pr\left(-1.96\frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq +1.96\frac{\sigma}{\sqrt{n}}\right) = .95.$$

(9.7)

One way of interpreting this inequality statement is that 95% of the time, the distance between μ and $\overline{X}$ is less than $1.96\sigma/\sqrt{n}$. That is, 95% of the time μ is within a certain distance of $\overline{X}$. Of course, this also means that 95% of the time $\overline{X}$ is within the same distance of μ. (If you and I are walking down the street and 95% of the time you are within 3 feet of me, then 95% of the time I am within 3 feet of you.) What this means, in turn, is that if we take $\overline{X}$ and construct an interval by adding and subtracting $1.96\sigma/\sqrt{n}$ from it, that interval will

have μ within its endpoints 95% of the time in the long run. Such insight is not necessary to derive the confidence interval, however. One may simply continue manipulating the interval algebraically. First, subtract $\overline{X}$ from all three sections of the inequality. Then multiply all three sections by -1, and reverse the direction of the inequality. This leads to the following expression:

$$\Pr\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}} \right) = .95. \qquad (9.8)$$

The expression states that if one constructs an interval with endpoints $\overline{X} \pm 1.96\sigma/\sqrt{n}$, this interval will contain the true parameter (μ) 95% of the time in the long run.

A number of simple inequalities can be converted into confidence intervals in this way. Typically, one finds, in elementary to intermediate texts, confidence intervals for (a) a single mean, (b) the difference between two means, (c) a single contrast on means, (d) a single variance, (e) the ratio of two variances, (f) a single correlation, and (g) a single proportion. An element common to the preceding intervals is that an interval statement about the distribution of the null distribution of a test statistic can be manipulated easily to yield the desired confidence interval. Situations where (a) the distribution of the test statistic changes as a function of the parameter to be estimated, and (b) simple interval manipulation does not yield a convenient confidence interval, are generally not discussed. As an example, consider the sample squared multiple correlation, whose distribution changes as a function of the population squared multiple correlation. Confidence intervals for the squared multiple correlation are very informative, yet are not discussed in standard texts, because a single simple formula for the direct calculation of such an interval cannot be obtained in a manner analogous to the way we obtain a confidence interval for μ.

A general method for confidence interval construction is available that includes the method discussed earlier as a special case, but also allows confidence limits and confidence intervals to be constructed when the aforementioned method cannot be applied. This method combines two general principles, which we call the *confidence interval transformation* principle and the *inversion confidence interval* principle. The former is obvious, but seldom discussed formally. The latter is referred to by a variety of names in several classic references (Kendall & Stuart, 1979; Cox & Hinckley, 1974), yet does not seem to have found its way into the standard textbooks, primarily because its implementation involves some difficult computations. However, the method is easy to discuss *in principle*, and no longer impractical. Interestingly, when the two principles are combined, a number of very interesting confidence intervals result.

First we discuss the confidence interval transformation principle.

Proposition 1. Confidence Interval Transformation Principle. Let $f(\theta)$ be a monotonic, strictly increasing continuous function of θ. Let l_1 and l_2 be endpoints of a $1 - \alpha$ confidence interval on quantity θ. Then $f(l_1)$ and $f(l_2)$ are endpoints of a $1 - \alpha$ confidence interval on $f(\theta)$.

To prove the proposition, recall that a function is monotonic and strictly increasing if, when plotted in the plane, the graph "keeps going up" from left to right, that is, it never flattens out or goes down. A monotonic, strictly increasing function is *order preserving*. Because the plot never flattens out, if $x > y$, then $f(x) > f(y)$. This can be seen easily by examining Figure 9.2.

If l_1 and l_2 are endpoints of a valid .95 confidence interval on quantity θ, then 95% of the time in the long run, θ is between l_1 and l_2. If $f( )$ is a monotonic strictly increasing function, l_2 is greater than θ, and θ is greater than l_1, then it must also be the case that $f(l_2) > f(\theta)$, and $f(\theta) > f(l_1)$. Consequently, if l_1 and l_2 are endpoints of a $1 - \alpha$ confidence interval for parameter θ then $f(l_1)$ and $f(l_2)$ are endpoints of a valid $1 - \alpha$ confidence interval on $f(\theta)$.

Here are two elementary examples of the confidence interval transformation principle.

Example 1. A Confidence Interval for the Standard Deviation. Suppose you calculate a confidence interval for the population variance σ^2. Such a confidence interval is discussed in many elementary textbooks. You desire a confidence interval for σ. Confidence intervals for σ are seldom discussed in textbooks. However, one may be derived easily. Because σ takes on only nonnegative values, it is a monotonic increasing function of σ^2 over its domain. Hence, the confidence interval for σ is obtained by taking the square root of the endpoints for the corresponding confidence interval for σ^2.

Example 2. Inverting the Fisher Transform. Suppose one calculates a confidence interval for $z(\rho)$, the Fisher transform of ρ, the population correlation coefficient. Taking the *inverse* Fisher transform of the endpoints of this interval will give a confidence interval for ρ. This is, in fact, the method employed to calculate the standard (approximate) confidence interval for a correlation.

These examples show why the confidence interval transformation principle is very useful in practice. Frequently a statistical quantity we are very interested in (like ρ) is a simple function of a quantity (like $z(\rho)$) we are not so interested in, but for which we can easily obtain a confidence interval.
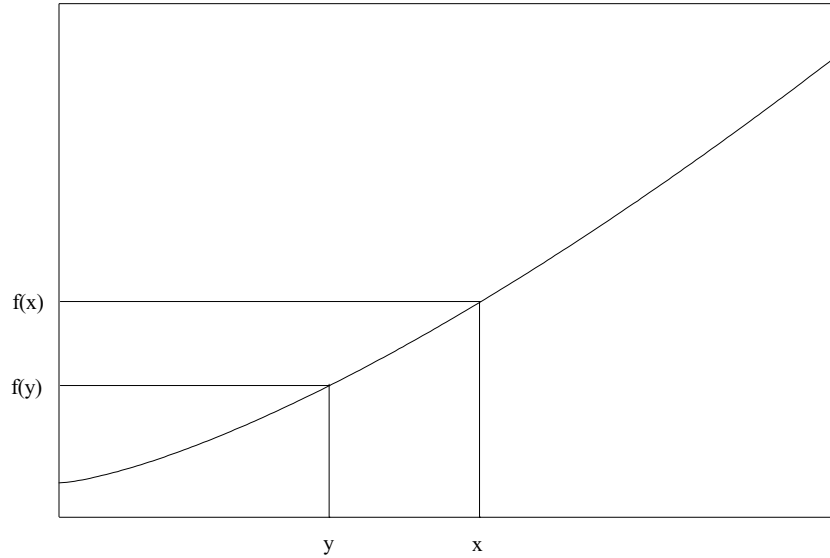
FIGURE 9.2   Order-preserving properties of a monotonic, strictly increasing function.

As an example, suppose we take two independent samples of size n_1 and n_2. We calculate sample means $\bar{x}_1$ and $\bar{x}_2$ and sample variances s_1^2 and s_2^2. Under the standard assumptions of normality and homogeneity of variance, the two-sample t statistic is used to decide whether an experimental and control group differ:

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)\left(\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right)}} . \tag{9.9}$$

The traditional approach is to compute the t statistic and perform a significance test. A better, but less frequently employed procedure is to report a confidence interval on the quantity $E = \mu_1 - \mu_2$ using the following standard formula for the endpoints (where $t*$ is the critical value from Student's t distribution):

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm t*_{\alpha/2,\, n_1+n_2-2} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right)} . \tag{9.10}$$

In practice, an even more useful quantity than E is the *standardized* effect size, defined as

$$E_s = \frac{\mu_1 - \mu_2}{\sigma}. \tag{9.11}$$

E_s is a standardized, or "metric-free" measure of effect size. If one two-group study reports its results in pounds, the other in kilograms, then the unstandardized effect E will not be in the same scale of measurement in the two studies, whereas E_s will be. Consequently, a confidence interval on E_s is more informative than one on E. This is especially true when different studies are compared, or when information is combined across studies.

In the context of meta-analysis, Hedges and Olkin (1985) discussed a variety of methods for estimating E_s, most of which are approximations. The exact method (which they discuss on page 91 of their book) involves the noncentrality interval estimation approach. This approach was considered impractical for general use at the time their book was written, so the authors provided nomographs only for some limited cases involving very small samples.

Before continuing, we digress briefly to recall some mathematical background on the key *noncentral* distributions for the less advanced reader. The normal, t, χ^2, and F distributions are statistical distributions covered in most introductory texts. These distributions can be related to the normal distribution in various ways: for example, squaring a random variable that has a standard normal distribution yields a random variable that has a χ^2 distribution with 1 degree of freedom. The t, χ^2, and F distributions are special cases of more general distributions called the *noncentral t, noncentral χ^2*, and *noncentral F*. Each of these noncentral distributions has an additional parameter, called the *noncentrality parameter.* For example, whereas the F distribution has two parameters (the "numerator" and "denominator" degrees of freedom), the *noncentral F* has these two plus a *noncentrality parameter.* When the *noncentral F* distribution has a noncentrality parameter of zero, it is identical to the F distribution, so it includes the F distribution as a special case. Similar facts hold for the t and χ^2 distributions. What makes the noncentrality parameter especially important is that it is related very closely to the truth or falsity of the typical null hypotheses that these distributions are used to test. So, for example, when the null hypothesis of no difference between two means is correct, the standard t statistic has a distribution that has a noncentrality parameter of zero, whereas if the null hypothesis is false, it has a noncentral t distribution. In general, the more false the null hypothesis, the larger the noncentrality parameter.

Suppose we take data from two independent samples, and calculate the two sample t-statistic shown in Equation 9.9. The statistic has a distribution which is noncentral t, with noncentrality parameter

$$\delta = E_s \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \, . \tag{9.12}$$

When the null hypothesis is true, δ is zero and is not a particularly interesting quantity. However, E_s is a statistical quantity of considerable interest, and may be obtained from δ by a simple monotonic transformation

$$E_s = \delta \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \, . \tag{9.13}$$

Hence, if we can obtain a confidence interval for δ, we also can obtain a confidence interval for E_s, using the confidence interval transformation principle.

We now describe how to obtain a confidence interval for δ. When we discuss continuous probability distributions, we often talk in terms of the cumulative distribution function, or CDF, and we use the notation $F(\ )$ to denote this function. The CDF evaluated at a point x is defined as the probability of obtaining a value *less than or equal to x*, hence the term *cumulative*. Many normal curve tables in the back of standard textbooks are CDF tables. For example, in the unit standard normal distribution, half of the cases fall at or below 0, so $F(0) = .50$. Ninety-five percent of the cases fall at or below 1.645, so $F(1.645) = .95$, and $F(-1.645) = .05$. Sometimes, when solving problems involving the normal curve table, one needs to "reverse" the table. For example, if I asked you what point in the normal curve has 95% of the cases at or below it, you would scan down the table until you found .95, move to the number in the column next to .95, and report back "1.645" as your answer. This process of reversing the roles of the two columns in the table is equivalent to inverting the CDF function. In mathematical notation, we say that the CDF function has an inverse, and that $F^{-1}(.95) = 1.645$. In a similar vein, $F^{-1}(.5) = 0$, and $F^{-1}(.05) = -1.645$.

Obtaining a confidence interval for δ is simple in principle, though not in practice. Consider the graph in Figure 9.3. This graph shows the .05 and .95 cumulative probability points for a noncentral distribution for fixed degrees of freedom, and varying values of the noncentrality parameter δ. These functions, labeled "5th percentile" and "95th percentile" in the graph, can be denoted more formally as $F^{-1}(.05, \delta)$ and $F^{-1}(.95, \delta)$, respectively, because they are the inverse of the CDF of the noncentral t, for fixed probability level, evaluated at δ.

To develop a confidence interval for δ, we need to find functions of the sample data that bracket δ a certain proportion of the time, and Figure 9.3 provides the key to obtaining such a function. Consider the upper curve in the graph. For any value of δ along the X axis, this curve plots the observed value t below which the noncentral t will occur 95% of the time. Now, suppose the *true value*
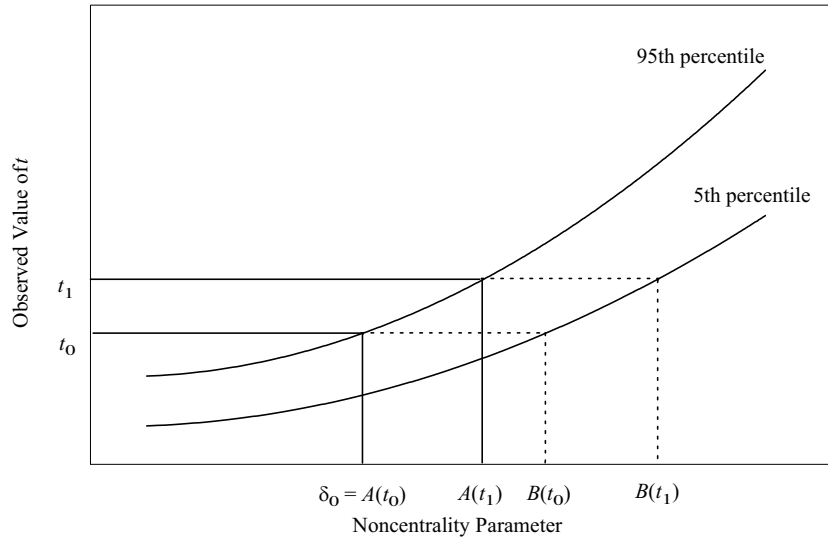
FIGURE 9.3   Noncentrality interval estimation with a confidence belt

of δ is δ_0. This means that 95% of the time the observed t value will be less than the value on the Y-axis marked as t_0.

Careful consideration of the upper curve reveals that its *inverse* (which exists, because the function is monotonic and strictly increasing in δ) can be used to construct a lower confidence limit, or a lower bound on a confidence interval. To see this, move along the Y-axis and, from any value t draw a horizontal line straight out until it intersects with the upper curve, then draw a point straight down until the X-axis is intersected. Call the value obtained this way $A(t)$, because it is a function of the t value chosen from the Y-axis. This value is the *inverse* of the function of the upper curve, evaluated at t. Note that each value of t corresponds to one and only one value of δ.

Imagine that the true noncentrality parameter is δ_0. Imagine further that, for each value of t that is observed, you compute the *inverse* of the upper curve function at t by drawing a line straight over to the upper curve, then straight down to the X-axis. With such a procedure, 95% of the time you will observe a value of t that is less than t_0, and so 95% of the time you will observe a value of $A(t)$ that is less than $A(t_0)$. But $A(t_0) = \delta_0$. Consequently, $A(t)$ produces a 95% lower confidence limit for δ, because it produces numbers that are below δ exactly 95% of the time.

If we call the inverse of the lower curve's function $B(t)$, a similar procedure provides an upper 95% confidence limit for δ. That is, draw a horizontal line from an observed t value on the Y-axis to the lower (5th percentile) curve, then

down (perpendicular) to the *X*-axis. By a similar logic to that discussed previously, the value obtained by this procedure will be above δ 95% of the time, and will therefore be a 95% upper confidence limit. Taken together, $A(t)$ and $B(t)$ provide a 90% confidence interval for δ.

This method works, in general, so long as the $\alpha/2$ and $1 - \alpha/2$ probability points are montonic and strictly increasing as a function of the unknown parameter with the other (known) parameters considered as fixed values. Note that, in practice, one does not have to generate the entire curve of values, because the endpoints of the confidence interval are simply those values of the unknown parameter for which the cumulative probabilities of the observed data are $1 - \alpha/2$ and $\alpha/2$. So if you have a computer routine that can solve for these two values directly, there is no need to plot this curve. (In practice, numerical analysis root-finding techniques like the method of *bisection* are substantially faster than the graphical approach shown here for demonstration purposes. Computer software can calculate the intervals in approximately one second for most practical examples.)

The following proposition expresses succinctly the result of our graphical investigation.

> **Proposition 2. Inversion Confidence Interval Principle.** Let v be the observed value of X, a random variable having a continuous (cumulative) probability distribution expressible in the form $F(v, \theta) = \Pr(X \leq v|\theta)$ for some numerical parameter θ. Let $F(v, \theta)$ be monotonic, and strictly decreasing in θ, for fixed values of v. Let l_1 and l_2 be chosen so that $\Pr(X \leq v \mid \theta = l_1) = 1 - \alpha/2$ and $\Pr(X \leq v \mid \theta = l_2) = \alpha/2$. Then l_1 is a lower $1 - \alpha/2$ confidence limit for θ, l_2 is an upper $1 - \alpha/2$ confidence limit for θ, and the interval with l_1 and l_2 as endpoints is a $1 - \alpha$ confidence interval for θ.

We call the method we have just described *noncentrality interval estimation*, because in practice one frequently estimates the noncentrality parameter en route to a more interesting statistical quantity. The following numerical example shows how the method is used to estimate standardized effect size in a two-group experiment.

Example 3. Estimating the Standardized Effect in Two-Group Experiments. In this example, we apply the noncentrality interval estimation approach to two hypothetical two-group experiments, each involving two independent samples of equal size. Experiment 2 was based on an extremely large sample size of 300 per group, whereas Experiment 1 had only 10 per group. The two-tailed *p* levels for the experiments were approximately the same, with Experiment 1 having the higher *p* level (.0181). Experiment 2 had a *p* level of .0167, and thus was "more
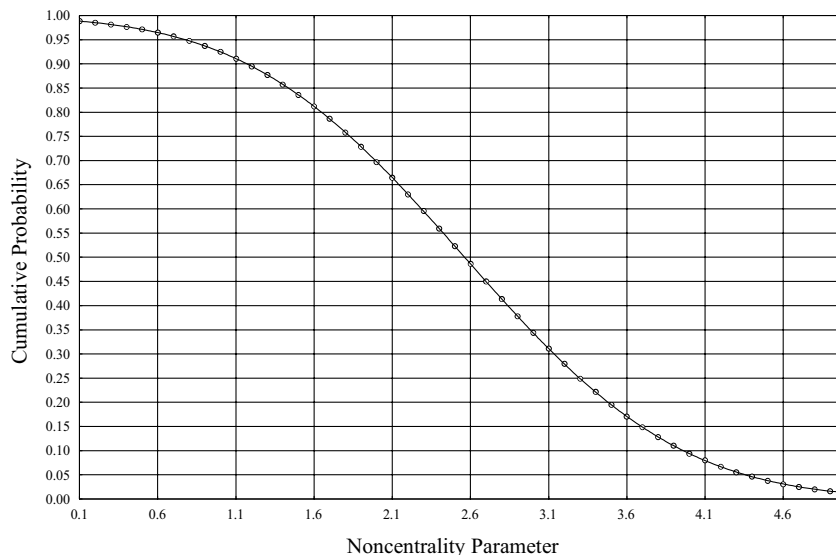
FIGURE 9.4   Noncentrality and the cumulative probability of an observed t statistic

significant." Summary statistics for the experiments, along with confidence in-
tervals for the standardized effect size, are shown in Table 9.2.

We now proceed to demonstrate how the confidence interval for E_s may be
calculated. First, we calculate a .95 confidence interval for δ. The t statistic in
Group 1 has an observed value of 2.60 with 18 degrees of freedom. The end-
points of the confidence interval for δ are those values of δ that generate the
unique noncentral $t_{18,\delta}$ distributions in which the observed value of 2.60 has cu-
mulative probability .975 and .025. If a good noncentral t distribution calculation
program is available, and its output can be plotted, these values may be approxi-
mated fairly closely by graphical analysis. Figure 9.4 shows a plot of the cumu-
lative probability of the value 2.60 as a function of δ for the family of noncentral
t distributions with 18 degrees of freedom.

TABLE 9.2
Comparison of confidence intervals for E_s in two experiments.

	Experiment 1	Experiment 2
n per group	10	300
Observed t statistic	2.60	2.40
p level (2-tailed)	.0181	.0167
95% Confidence Interval for E_s	(.1950, 2.1034)	(.0355, .3563)
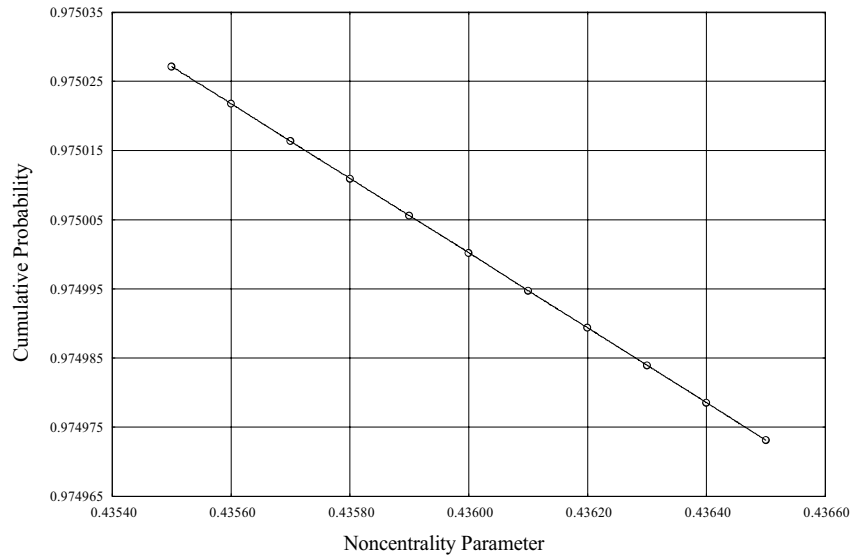
FIGURE 9.5   Calculating the lower confidence limit for the noncentrality parameter

Zooming in on a narrower region of Figure 9.4, we obtain the view shown in Figure 9.5, in which we can pinpoint, with a high degree of accuracy, the value of δ for which the cumulative probability is .975 at approximately .4360. This is the lower endpoint of the .95 confidence interval for δ. In a similar manner, we can determine the value for the upper endpoint as 4.7033.

We recall from Equation 9.13 that, using the confidence interval transformation principle, these endpoints may be transformed into a confidence interval for the standardized effect size E_s by multiplying them by

$$0.44721 = \sqrt{\frac{10+10}{10\times10}} .$$

Consequently, the .95 confidence interval for E_s has endpoints .1950 and 2.1034. In a similar manner, we can determine the confidence interval for Experiment 2 to be .0355 and .3563.

The confidence intervals for E_s demonstrate clearly that the experiments have rather different implications. In Experiment 2, a 95% confidence interval for E_s ranges from only .0355 to .3563 standard deviation units. In other words, the experiment determined, with a high degree of precision, that the effect is at best

moderate, and quite possibly less than a tenth of a standard deviation. On the other hand, the confidence interval in Experiment 1 demonstrates that the experimental effect has not been determined with precision. The confidence interval for E_s includes such disparate values as .195 (small effect) and 2.10 (very powerful effect). Clearly, the confidence intervals for E_s convey much more useful information than the p levels.

APPLICATIONS OF NONCENTRALITY INTERVAL ESTIMATION

The noncentrality interval estimation approach we illustrated in the preceding section can be applied in a number of common data-analytic situations. Here we examine several, showing how noncentrality interval estimation adds substantial new information to the analysis.

Standardized Effect Size for Planned Orthogonal Contrasts.

As a result of our preceding analysis of Experiments 1 and 2, we might wish to estimate the difference in standardized effect sizes with a confidence interval. We can do this by using a planned orthogonal contrast. Planned orthogonal contrasts are an extension of the two-sample t statistic that can be performed routinely on K independent samples to test hypotheses of the form

$$\Psi = \sum_{k=1}^{K} c_k \mu_k = 0. \tag{9.14}$$

The c_k are referred to as "linear weights," or "contrast weights," and determine the hypothesis being tested. For example, if $K = 2$, and the contrast weights are $+1$ and -1, then $\Psi = \mu_1 - \mu_2$ and the hypothesis being tested is that the two means are equal. Under the standard assumptions of normality and homogeneity of variance, such hypotheses may be tested with a t statistic of the form

$$t = \frac{\hat{\Psi}}{\sqrt{\hat{\sigma}_{\hat{\Psi}}^2}} = \frac{\sum_{k=1}^{K} c_k \overline{X}_{\bullet k}}{\sqrt{\left(\sum_{k=1}^{K} \frac{c_k^2}{n_k}\right) MS_{error}}}. \tag{9.15}$$

The t statistic has a noncentral t distribution with degrees of freedom equal to those for mean square error, and a noncentrality parameter given by

$$\delta = \frac{\Psi}{\sigma\sqrt{\sum_{k=1}^{K}\frac{c_k^2}{n_k}}} = \frac{\Psi/\sigma}{\sqrt{\sum_{k=1}^{K}\frac{c_k^2}{n_k}}} = \frac{\sum_{k=1}^{K}\frac{c_k\mu_k}{\sigma}}{\sqrt{\sum_{k=1}^{K}\frac{c_k^2}{n_k}}}. \tag{9.16}$$

Confidence intervals for Ψ may be constructed with endpoints

$$\hat{\Psi} \pm t^*\sqrt{\left(\sum_{k=1}^{K}\frac{c_k^2}{n_k}\right)MS_{error}}. \tag{9.17}$$

However, a confidence interval for the standardized contrast

$$\Psi_s = \sum_{k=1}^{K}\frac{c_k\mu_k}{\sigma}, \tag{9.18}$$

is generally more informative, because it expresses how false the null hypothesis is in standardized units of measurement. Because

$$\Psi_s = \frac{\Psi}{\sigma} = \delta\sqrt{\sum_{k=1}^{K}\frac{c_k^2}{n_k}}, \tag{9.19}$$

it is a trivial matter to convert a confidence interval for δ into one for Ψ_s, the standardized contrast.

Example 4. Contrasting the Mean Differences for Two Experiments. Suppose, for the sake of simplicity, that MS_{error} is equal to 100 in both experiments in Table 9.2. If we compare the mean differences for the two experiments, the contrast weights are

$$c_1 = 1, \; c_2 = -1, \; c_3 = -1, \; c_4 = 1.$$

The mean differences are 11.628 in Experiment 1 and 1.959 in Experiment 2. The t statistic comparing the two mean differences is 2.127. The noncentrality interval estimation technique may be applied to the data in Table 9.2 to obtain a .95 confidence interval estimate on the difference in standardized effects be-

tween Experiments 1 and 2. We find that this confidence interval has endpoints of .0739 and 1.859. Since the confidence interval does not include zero, the first ("less significant") experiment has a significantly higher standardized effect than the second ("more significant") experiment, although the size of the effect difference has not been established with much precision.

This confidence interval provides much more useful, and much more accurate information about the relative effects in the two experiments than a comparison of the *p* levels.

**Exact Confidence Intervals for Root Mean Square
Standardized Effect Size in the One-Way Fixed Effects
Analysis of Variance.**

The ideas we developed for a single contrast on means in the context of the *t*-statistic generalize readily to the case of several contrasts in the analysis of variance. In orthogonal analysis of variance designs with equal cell sizes, the *F* statistic has a noncentral *F* distribution that, in general, is a simple function of the *root mean square standardized effect*. Here we examine the simplest special case, the one-way fixed-effects analysis of variance with *n* observations per group. Consider the *F* statistic in a one-way, fixed-effects ANOVA with *n* observations per group, and *K* groups. Let α_k be the treatment effect associated with group *k*, and σ^2 be the error variance. The *F* statistic with $K - 1$ and $K(n - 1)$ degrees of freedom has noncentrality parameter

$$\delta = n \sum_{k=1}^{K} \left(\frac{\alpha_k}{\sigma} \right)^2 . \tag{9.20}$$

The quantity δ/n is thus the sum of squared standardized effects. Now, suppose we wish to "average" these standardized effects in order to obtain an overall measure of strength of effects in the design. One possibility is simply the arithmetic average of the *K* standardized effects, that is, $\delta/(nK)$. One problem with this measure is that it is the average *squared* effect, and so is not in the proper unit of measurement. A second problem is that because of the way effects are defined in the analysis of variance, there are only $K - 1$ mathematically independent effects, because there is one constraint imposed upon the effects for identifiability, that is, that the effects sum to zero. Because the definition of "effect" in the analysis of variance depends on a mathematical restriction that is arbitrary, there are in fact infinitely many ways we *could* choose to define an ANOVA effect. Often, the "effects" defined by the standard ANOVA restriction need not coincide with experimental effects the way we commonly think of them. Consider the very simple special case of a two-group experiment involv-

ing a treatment group and a placebo control. Suppose the population standard deviation is 1, and the control group has a population mean of 0, the experimental group has a population mean of 2. In this case, there is one standardized experimental effect, and it is 2 standard deviations in size. On the other hand, the analysis of variance defines two effects, and they are $\alpha_1 = -1$ and $\alpha_2 = +1$, respectively. So, if we average the sum of squared ANOVA effects, we come up with an average squared effect of 1. Clearly, this is misleading, an artifact of the way ANOVA effects are defined.

There is, in our opinion, no simple, universally acceptable solution to this problem. However, averaging with K appears to underestimate effect levels consistently. Consequently, we propose to average by the number of independent effects, that is, $K - 1$. With this stipulation, $\delta/[(K-1)n]$ is the average squared standardized effect, and the root mean square standardized effect is

$$RMSSE = \sqrt{\frac{\delta}{(K-1)n}} \; . \tag{9.21}$$

With this definition, we find that, in the above numerical example, the RMSSE is 1.41.

In order to obtain a confidence interval for *RMSSE*, we proceed as follows. First, we obtain a confidence interval estimate for δ by iteration, using the noncentrality interval estimation approach. Next, we directly transform the endpoints by dividing by $(K-1)n$. Finally, we take the square root. The result is an exact confidence interval for the root mean square standardized effect in the analysis of variance.

Example 5. Confidence Intervals on the RMSSE. Suppose a one-way fixed-effects ANOVA is performed on 4 groups, each with an *n* of 20. An overall *F* statistic of 5.00 is obtained, with a *p* level of .0032. The *F* test is thus "highly significant" and the null hypothesis is rejected at the .01 level. In this case, the noncentrality interval estimate provides a somewhat less awe-inspiring account of what has been found. Specifically, the 95% confidence interval for δ ranges from 1.866 to 32.5631, and the corresponding confidence interval for the root mean square standardized effect ranges from .1764 to .7367. Effects are almost certainly "there," but they are on the order of half a standard deviation.

Example 6. Confidence Intervals on Hays' η^2. Fleishman (1980) described the calculation of confidence intervals on the noncentrality parameter of the noncentral *F* distribution to obtain, in a manner equivalent to that employed in the previous two examples, confidence intervals on Hays' η^2, which is defined as

$$\eta^2 = \frac{\sigma_a^2}{\sigma_t^2}, \tag{9.22}$$

where σ_a^2 is the variance due to effects, and σ_t^2 is the total variance. Fleishman also discussed the "signal to noise ratio"

$$f^2 = \frac{\sigma_a^2}{\sigma_e^2}. \tag{9.23}$$

Fleishman (1980) defined the "effect variance" σ_a^2 in the fixed-effects case as

$$\sigma_a^2 = \frac{\sum_{k=1}^{K} \alpha_k^2}{K}, \tag{9.24}$$

and so f^2 relates to δ in a one-way fixed-effects ANOVA via the equation

$$f^2 = \frac{\delta}{nK}. \tag{9.25}$$

Fleishman (1980) cites an example given by Venables (1975) of a 5-group ANOVA with $n = 11$ per cell, and an observed F of 11.221. In this case, the .90 confidence interval for the noncentrality parameter δ has endpoints 19.380 and 71.549, whereas the confidence interval for f^2 ranges from .352 to 1.301.

Exact Confidence Intervals for RMSSE in Fixed-Effect Factorial ANOVA

The method of the preceding section may be generalized to completely randomized factorial designs in the analysis of variance. However, some modification is necessary, because the noncentrality parameter for factorial fixed-effects designs is a function of the number of cells in which an effect operates. Consider, for example, the two-way fixed-effects ANOVA, with J rows, K columns, and n observations per cell. Consider the F statistic for row effects. This statistic is essentially the one-way analysis of variance computed on the row means collapsed across columns. Consequently, the statistic has a noncentral F distribution with noncentrality parameter

$$\delta_a = nK \sum_{j=1}^{J} \left(\frac{\alpha_j}{\sigma_e} \right)^2. \tag{9.26}$$

Each row effect operates on K columns, and naturally the noncentrality parameter of the F distribution reflects that fact. Similarly, the F statistic for column effects is essentially a one-way analysis of variance performed on the column means collapsed across rows, and the noncentrality parameter for the overall F statistic for column effects is

$$\delta_b = nJ \sum_{k=1}^{K} \left(\frac{\beta_k}{\sigma_e} \right)^2 . \tag{9.27}$$

In general, the RMSSE for a particular effect is of the form

$$RMSSE_{effect} = \sqrt{\frac{\delta_{effect}}{n_{effect} df_{effect}}} , \tag{9.28}$$

where δ_{effect} is the noncentrality parameter for the F statistic for the effect, and n_{effect} is the total number of observations in the collected cell means used to compute the effect. For example, in a two-way ANOVA, the row effects are estimated by summing across the K columns to reduce the ANOVA, in effect, to a one-way ANOVA on cell means based on nK observations per "cell." For the AB interaction, however, n_{effect} is n, because interactions are computed on individual cells, not on rows or columns that are summed across. In a three-way ANOVA, with J rows, K columns, and H levels of the third factor, n_{effect} for the row effect is nKH.

Example 7. RMSSE in a Two-way ANOVA. Suppose a two-way 2×7 ANOVA is performed with $n = 4$ observations per cell, and the source table is as in Table 9.3. In this source table, all 3 effects are significant. There is a significant main effect for factors A and B, and a significant AB interaction. Notice that the p level for the A main effect (.0186) is about half that for the interaction (.0369).

One might be tempted to declare the A main effect to be "more significant" than the interaction. However, the 90% RMSSE confidence intervals for the main effects and interaction would seem to dispute that. The low ends for the confidence intervals are virtually identical. The upper end of the confidence interval for the AB interaction effect is substantially higher for the AB interaction than for either main effect.

This suggests that there is less power (and precision of estimate) for detecting interaction effects than for detecting main effects in this design. It is wise to remember this when making decisions about the "additivity" of models in the analysis of variance. The table also dramatizes that, because of the different power curves, different numbers of cells, and different constraints on effects, it is

TABLE 9.3
A Two-Way (2×7) Fixed Effects ANOVA

Source	SS	df	MS	F	p level	RMSSE Lower	RMSSE Upper
A	14.40	1	14.40	6.00	.0186	.136	.782
B	38.16	6	6.36	2.65	.0285	.135	.754
AB	36.00	6	6.00	2.50	.0369	.139	1.038
Error	100.80	42	2.40				

very risky to characterize one result as "more significant" than another on the basis of p levels in ANOVA.

Exact Confidence Intervals for the Squared Multiple Correlation

One very common statistical application that practically cries out for a confidence interval is multiple regression analysis. Publishing an observed R^2 together with the result of a hypothesis test that the population squared multiple correlation, P^2, is zero, conveys little of the available statistical information. A confidence interval on P^2 is much more informative. *Exact* confidence intervals on P^2 can be computed using the inversion interval estimation approach. Yet general purpose statistical packages do not calculate such a confidence interval, and numerous well-known textbooks on multivariate analysis at both the theoretical and applied levels (e.g., Anderson, 1984; Morrison, 1990) do not allude to the possibility of calculating such an interval. The result is that numerous multiple regression studies have published R^2 values (along with various "shrunken" estimators) with no indication of experimental precision.

Kramer (1963) and Lee (1972) described methods for calculating the cumulative distribution of the squared multiple correlation coefficient. Both authors included tables in their articles. For a given observed R^2, fixed sample size, and number of predictors, the distribution of R^2 can be expressed as a function of P^2. (See, for example, Lee (1972), p. 178.) Consequently, the inversion confidence interval principle can be employed. However, the tables of Kramer and Lee provide only the upper percentage points of the distribution. Consequently, only the lower confidence limit, or "statistical lower bound" can be determined, and this must be accomplished by tedious linear interpolation.

Steiger and Fouladi (1992) provided a computer program, R2, for calculating exact confidence intervals on P^2. The program iterates an exact confidence in-

terval and confidence limit, using the noncentrality interval estimation approach. Such intervals can be quite revealing.

Example 8. Confidence Intervals for the Squared Multiple Correlation. Suppose a criterion is predicted from 45 independent observations on 5 variables and the observed squared multiple correlation is .40. In this case a 95% confidence interval for P^2 ranges from .095 to .562! A 95% lower confidence limit is at .129. On the other hand the R^2 value is significant "beyond the .001 level," because the p level is .0009, and the shrunken estimator is .327. Clearly, it is far more impressive to state that "the R^2 value is significant at the .001 level" than it is to state that "we are 95% confident that P^2 is between .095 and .562." But we believe the latter statement conveys the quality and meaning of the statistical result more accurately than the former.

Some writers, like Lee (1972), prefer a lower confidence limit, or "statistical lower bound" on the squared multiple correlation to a confidence interval. The rationale, apparently, is that one is primarily interested in assuring that the percentage of variance "accounted for" in the regression equation exceeds some value. Although we understand the motivation behind this view, we hesitate to accept it. *The confidence interval, in fact, contains a lower bound, but also includes an upper bound, and, in the interval width, a measure of precision of estimation.* It seems to us that adoption of a lower confidence limit can lead to a false sense of security, and reduces that amount of information available in the model assessment process.

We believe that confidence intervals always should be reported with a multiple correlation. However, we add a note of caution. Strictly speaking, such confidence intervals (as well as the significance test) will not be accurate unless distributional assumptions have been met, and the independent variables in the regression equation specified a priori. In many cases, the final regression equation has been determined by some kind of exploratory stepwise approach, and no attempt has been made at cross-validation. It is important to reemphasize that estimates of P^2 and confidence intervals are biased by this specification search. For the interval estimation approach discussed here to be valid, a cross-validation sample should be used.

Asymptotic Confidence Intervals for Goodness of Fit in the Analysis of Covariance Structures

A key area where noncentrality interval estimation has been applied with excellent results is in the analysis of covariance structures, sometimes referred to as "causal modeling." In this area, the statistical inference is usually of the accept-

support (AS) variety. In this kind of situation, standard significance testing logic is badly strained.

Until approximately 1980, models were evaluated in the analysis of covariance structures by using the chi-square test of fit. The problem with the procedure is that it tests a hypothesis of perfect fit. Since this hypothesis is often false, the statistical decision rendered by the chi-square statistic often boiled down to a question of sample size. With small samples, poorly fitting models might be "accepted," while with large samples a model with excellent fit (in the practical sense) might be overwhelmingly rejected. The results were often embarrassing. Sometimes models which appeared to fit very well were rejected "beyond the .01 level." Awkward mental contortions were required to simultaneously praise the maximum likelihood chi-square statistic as a technical breakthrough, while ignoring its result.

Steiger and Lind (1980) demonstrated that performance of statistical tests in common factor analysis could be predicted from a noncentral chi-square approximation. The noncentrality parameter was n times the "population discrepancy function," which is the (maximum likelihood or generalized least squares) discrepancy function calculated on the population covariance matrix. Consequently, the population discrepancy function was an excellent candidate for a descriptive index of how badly a model fit in a particular population. Steiger and Lind suggested abandoning the tradition of hypothesis testing in favor of constructing a *confidence interval on the population discrepancy function* (or some particularly useful function of it). This approach offers two worthwhile pieces of information at the same time. It allows one, for a particular model and data set, to express (a) how bad fit is in the population, and (b) how precisely the *population* badness of fit has been determined from the *sample* data.

Steiger (1989, 1990b) implemented three noncentrality-based indices of fit in the computer program EzPATH, including the index originally proposed by Steiger and Lind (1980). All these indices can be computed with confidence intervals. One index, the RMSEA, divides the population fit function F^* by the degrees of freedom, then takes a square root to obtain a "Root Mean Square Error of Approximation," in a manner roughly analogous to the RMSSE we recommended for the fixed effects factorial ANOVA earlier in this article. Most current structural modeling programs (e.g., LISREL, EQS, SEPATH, CALIS, RAMONA) calculate the RMSEA, which Browne and Cudeck (1992) also recommend.

The other two noncentrality-based indices developed by Steiger were population analogs of the GFI and AGFI of Jöreskog and Sörbom (1984). Jöreskog and Sörbom recommended the finite sample equivalents of these as sample-based indices, but offered no population rationale for them. Steiger (1989) and Maiti and Mukherjee (1990) demonstrated that the sample-based GFI and AGFI could

be viewed as biased estimators of Steiger's (1989) equivalent population quantities, and that both of these indices, under fairly general conditions, could be written as a simple monotonic function of the population noncentrality parameter. For example, for structural models (based on p observed variables) that are invariant under a constant scaling factor, Steiger's Γ_1, the population equivalent of the GFI (i.e., the GFI calculated on the population covariance matrix) can be written

$$\Gamma_1 = \frac{p}{2F^* + p}. \tag{9.29}$$

This simple monotonic relationship implies that, via the confidence interval transformation principle, a confidence interval on the noncentrality parameter of a noncentral chi-square distribution can be converted easily into a confidence interval on Γ_1.

In the documentation for the structural equation modeling program SEPATH, Steiger (1995) extended the noncentrality-based indices to multiple samples, and gave a simplified formula for estimating the bias in the Jöreskog and Sörbom (1984) indices.

There are several advantages to the noncentrality-based approach. First, when the RMSEA and adjusted gamma are employed, the index is automatically corrected for model parsimony. For example, as models become more complex, fit tends to improve, all other things being equal, whereas degrees of freedom decrease. The RMSEA, calculated in the population (for single-sample models) with the equation

$$R^* = \sqrt{\frac{F^*}{df}}, \tag{9.30}$$

compensates for this by dividing by the degrees of freedom. Second, high sample size now works "for the experimenter" instead of against the experimenter, because larger sample sizes result in smaller confidence interval widths, reflecting greater precision of estimation. Third, the distinction between a "statistically significant" badness of fit and a "meaningful" badness of fit can now be made. The following example clarifies these advantages.

Example 9. Evaluating the fit of a circumplex model. A perfect, equally spaced circumplex correlation matrix (Guttman, 1954) has equal correlations on sub-diagonal strips. For example, a 6×6 correlation matrix would be of the form

TABLE 9.4
Correlation Pattern for a 6×6 Circumplex

1					
ρ_1	1				
ρ_2	ρ_1	1			
ρ_3	ρ_2	ρ_1	1		
ρ_2	ρ_3	ρ_2	ρ_1	1	
ρ_1	ρ_2	ρ_3	ρ_2	ρ_1	1

shown in Table 9.4. Guttman (1954) observed a correlation matrix that has been reprinted in a number of places, including Jöreskog (1978).

Suppose we were to test the null hypothesis that the Guttman (1954) correlation matrix is a perfect, equally spaced circumplex, using structural equation modeling software. The sample size ($n = 710$) is very large in this example. Hence, we would expect the precision of estimation to be very high. At the same time, we would have to keep in mind that the "accept-support" approach of the chi-square test commonly used in structural modeling would be of very limited usefulness in this situation. We recognize that a model with as many constraints as this one will almost certainly not fit perfectly in the population, and we have very high power to detect an imperfect fit.

The chi-square statistic yields, in this case, a value of 27.05 with 12 degrees of freedom. The probability level is .008, indicating that the null hypothesis of perfect fit must be rejected. However, a reasonable conclusion from confidence interval analysis is that, although it is highly probable that the data do not fit a circumplex *perfectly,* they do fit a circumplex well. The 90% confidence interval for the Steiger-Lind (1980) RMSEA index is between .021 and .064.

The corresponding confidence interval for the *adjusted* population *gamma* coefficient, the population equivalent of the Jöreskog-Sörbom (1984) AGFI, is between .972 and .997.

Both confidence intervals show excellent fit of the model was determined with high precision. A reasonable conclusion would seem to be that Guttman's data fit the model in Table 9.4 *very* well.

Statistical Bounds on Power.

Occasionally, in the aftermath of a failure to reject a statistical significance test, reviewers or authors speculate about the role of inadequate power in causing the failure to reject. Ironically, statistical inference about power itself is frequently

absent from such discussions. Often it need not be. In many situations power is, all other factors held constant, a monotonic, strictly increasing function of a noncentrality parameter. Consequently, we can use the confidence interval transformation principle to construct post hoc statistical upper bounds on power, *after* a significance test has been performed.

Taylor and Muller (1995) have discussed such an approach in the general context of the multivariate linear model. The procedure is, in principle, quite straightforward. For example, consider the F test in the 1-way fixed-effects ANOVA. Suppose we obtain a 90% confidence interval on the noncentrality parameter. Since power and the noncentrality parameter are monotonically functionally related for a given sample size and α, we may use the confidence interval transformation principle to obtain a confidence interval, *after seeing the data*, for power.

To avoid misunderstanding, we emphasize that (a) we do not favor the significance testing approach for exploratory social science research, and that (b) in situations where significance tests are to be performed, it is better to analyze power before gathering one's data. However, situations arise where data have been gathered, a significance test has been performed, and then someone raises a question about power.

In such situations, a confidence interval on power provides, in its upper and lower limits, a "best case" and "worst case" scenario, respectively, for power in the test just performed. The upper bound of the confidence interval for power, can be considered a 95% *statistical upper bound on power*. This is a number below which the true power occurs 95% of the time over repeated samples. If the 95% statistical upper bound on power is below a reasonable target value, say .90, it means that the most optimistic reading of the available evidence suggests that power was inadequate to detect the effects present in your data with a significance test. The lower end of a 90% confidence interval is a 95% *statistical lower bound on power*. If this end of the confidence interval exceeds some reasonable value, it can confirm that power was almost certainly adequate in the experiment just performed. Post-hoc statistical bounds on power combine information about the precision of estimate in a study with information about the actual effects in the study. As such, it relies more on available information and less on speculation than posterior power analyses based on hypothetical effect sizes.

Example 10. A Confidence Bound on Power in a One-Way ANOVA. Suppose a 1-way ANOVA is performed on two independent groups, with sample sizes of 15 in each group, and an F value of 2.0 is obtained. In this case, the two-tailed p level is .1683, and the null hypothesis is, of course, not rejected. The 90% confidence interval on the noncentrality parameter ranges from 0 to 9.459. When the noncentrality parameter is zero, "power," strictly speaking, does not exist (i.e.,

the null hypothesis is true), but the rejection probability is alpha (i.e., .05). So, in a sense, the lower bound of the confidence interval for power is .05. The upper bound of the confidence interval on power is the power corresponding to a non-centrality parameter of 9.459. This value (which may be calculated as .843) is a 95% *upper confidence limit* on power. The 90% confidence interval on the RMSSE ranges from 0 to .794. This suggests that (a) effect size cannot be determined with high precision in this design, and (b) even if the effect size is assumed to be the maximum statistically reasonable value, power is .843. This suggests the sample size in this study is too low to afford the precision of estimation deemed desirable in many areas of social science.

CONCLUSIONS AND NOTES ON APPLICATIONS

In this chapter, we have discussed confidence interval methods that offer a superior alternative to significance testing in situations where confidence intervals are seldom applied, or applied in a suboptimal manner. These confidence intervals provide all the information inherent in a significance test. They are no longer computationally impractical, and should augment or replace the corresponding significance test procedures. We have, in the body of the chapter, given examples of how the procedures may be used in many common statistical testing situations.

Many users will find that these techniques serve as a superior replacement for significance testing in common situations. Others will consider this view too radical, and will use them to augment the more traditional approaches.

Several times in this article, we emphasized the value of using the width of the confidence interval as in index of precision of estimate of a parameter. It should be remembered that the width of a confidence interval is generally a random variable, subject to sampling fluctuations of its own, and may be too unreliable at small sample sizes to be useful for some purposes.

In this regard, there are two additional issues that arise in implementing the inversion approach to interval estimation. The first issue arises in some common situations when the parameter space (i.e., the set of all possible parameter values) is bounded. For example, suppose one is constructing a confidence interval for the squared multiple correlation, or for the RMSEA index of fit in structural modeling. Neither of these parameters takes on negative values, so the parameter space is bounded on the left at zero. The inversion approach to interval estimation requires one to find a values of a parameter θ that imply sampling distributions in which the observed statistic is at the $\alpha/2$ and $1-\alpha/2$ quantiles. These values are the endpoints of the confidence interval. In some cases, however, the value of the observed statistic is so low that it is not possible to find a non-

negative value of θ that places it at the required percentage point. Standard procedure in this case is to arbitrarily set the confidence limit at zero, since the parameter cannot be less than zero. This maintains the correct coverage probability for the confidence interval, but the width of the confidence interval may be suspect as an index of precision of measurement when either or both ends of the confidence interval are at zero. In such cases, one might consider obtaining alternative indications of precision of measurement, such as an estimate of the standard error of the statistic. Often such estimates are readily available. A more proactive solution is to assure, in advance, that sample size is adequate to provide reasonable precision of estimation across a typical range of parameter values. For example, MacCallum, Browne, and Sugawara (1996) provide guidelines for appropriate sample size when using the RMSEA as an index of fit in structural equation modeling. Steiger and Fouladi (1992) provide a computer program, *R2*, for calculating appropriate sample size in multiple regression. These guidelines (developed in the context of power calculation within a hypothesis testing approach) should be given careful attention during the design of structural modeling and multiple regression studies. If they are followed, confidence intervals should seldom intrude on the boundaries of the parameter space.

There is a second issue that is probably of less concern in practice. When the *true* parameter is on the boundary of the parameter space, the coverage probability for the confidence interval may be *higher* than the nominal value. For example, suppose the population squared multiple correlation is zero. In such a situation, it is not possible to obtain a confidence interval that "misses" the true parameter on the low side, and so the confidence interval is *conservative*, i.e., the actual coverage probability is $1 - \alpha/2$, rather than $1 - \alpha$.

The fine details of programming computations were not discussed in this paper, but their importance should not be underestimated. Implementing the techniques is *much* more difficult than understanding them. Much of the development behind these methods is highly technical. In general, noncentral distribution routines present many more programming challenges than their central variants, and iterative routines used in the inversion approach must be programmed very cautiously to assure reliable performance.

Some of the methods discussed in this chapter already have been implemented in software whose availability is described on the website http://www.interchg.ubc.ca/steiger/homepage.htm. The program *R2*, available for computers running either the MSDOS or Windows operating system, computes confidence intervals, power, sample size required to achieve a given power, and other statistics on the squared multiple correlation. This program is available now. Other software will be announced as it becomes available. Interested readers may contact the senior author via email, at steiger@unixg.ubc.ca.

ACKNOWLEDGMENTS

REFERENCES

Anderson, T. W. (1984). *Introduction to multivariate statistical analysis* (2nd ed.). New York: Wiley.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models.* Beverly Hills, CA: Sage.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Cox, D. R., & Hinckley, D. V. (1974). *Theoretical statistics.* New York: Chapman & Hall.

Fleishman, A. E. (1980). Confidence intervals for correlation ratios. *Educational and Psychological Measurement, 40,* 659–670.

Guttman, L. B. (1954). A new approach to factor analysis: The radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences.* New York: Columbia University Press.

Guttman, L. B. (1977). What is not what in statistics. *The Statistician, 26*, 81–107.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika, 43,* 443–477.

Jöreskog, K. G., & Sörbom, D. (1984). *Lisrel VI. Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods.* Mooresville, IN: Scientific Software.

Kendall, M., & Stuart, A. (1979). *The advanced theory of statistics.* (Vol. 2)*.* New York: MacMillan.

Kramer, K. H. (1963). Tables for constructing confidence limits on the multiple correlation coefficient. *Journal of the American Statistical Association*, *58*, 1082–1085.

Lee, Y. S. (1972). Tables of upper percentage points of the multiple correlation coefficient. *Biometrika, 59*, 175–189.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130–149.

Maiti, S. S., & Mukherjee, B. N. (1990). A note on the distributional properties of the Jöreskog–Sörbom fit indices. *Psychometrika, 55,* 721–726.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Morrison, D. F. (1990). *Multivariate statistical methods.* (3rd Ed.). New York: McGraw-Hill.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57,* 416–428.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Steiger, J. H. (1989). *EzPATH: A Supplementary Module for SYSTAT and SYGRAPH.* Evanston, IL: SYSTAT Inc.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*, 173–180.

Steiger, J. H. (1995). *Structural equation modeling (SEPATH).* In *Statistica/W 5.0.* Tulsa, OK: StatSoft, Inc.

Steiger, J. H., & Fouladi, R. T. (1992). *R2:* A Computer Program for Interval Estimation, Power Calculation, and Hypothesis Testing for the Squared Multiple Correlation. *Behavior Research Methods, Instruments, and Computers, 4,* 581–582.

Steiger, J. H., & Lind, J. C. (1980). *Statistically based tests for the number of common factors.* Paper presented at the May annual meeting of the Psychometric Society, Iowa City, IA.

Taylor, D. J., & Muller, K. E. (1995). Computing confidence bounds for power and sample size of the general linear univariate model. *The American Statistician, 49,* 43–47.

Venables, W. (1975). Calculation of confidence intervals for non-centrality parameters. *Journal of the Royal Statistical Society, Series B, 37,* 406–412.