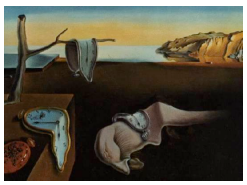

Journal of Statistical Software

May 2007, Volume 20, Issue 8. *http://www.jstatsoft.org/*

Confidence Intervals for Standardized Effect Sizes: Theory, Application, and Implementation

Ken Kelley
Indiana University

Abstract

The behavioral, educational, and social sciences are undergoing a paradigmatic shift in methodology, from disciplines that focus on the dichotomous outcome of null hypothesis significance tests to disciplines that report and interpret effect sizes and their corresponding confidence intervals. Due to the arbitrariness of many measurement instruments used in the behavioral, educational, and social sciences, some of the most widely reported effect sizes are standardized. Although forming confidence intervals for standardized effect sizes can be very beneficial, such confidence interval procedures are generally difficult to implement because they depend on noncentral t, F, and χ^2 distributions. At present, no main-stream statistical package provides exact confidence intervals for standardized effects without the use of specialized programming scripts. Methods for the Behavioral, Educational, and Social Sciences (**MBESS**) is an R package that has routines for calculating confidence intervals for noncentral t, F, and χ^2 distributions, which are then used in the calculation of exact confidence intervals for standardized effect sizes by using the confidence interval transformation and inversion principles. The present article discusses the way in which confidence intervals are formed for standardized effect sizes and illustrates how such confidence intervals can be easily formed using **MBESS** in R.

Keywords: standardized effect size, effect size, confidence intervals, noncentral distributions, null hypothesis significance test.

1. Introduction

In the behavioral, educational, and social sciences (BESS), units of measurement are many times arbitrary, in the sense that there is no necessary reason why the measurement instrument is based on a particular scaling. Many, but certainly not all, constructs dealt with in the BESS are not directly observable and the instruments used to measure such constructs do not generally have a natural scaling metric as do many measures, for example, in the physical

sciences. The scaling generally used for such instruments in the BESS tend to be on a scale that is thought to be reasonable, yet such scales are chosen by the developer of the instrument. For example, there might be two scales that measure the same dimension of "attitude" with the same psychometrically sound properties, yet one based on a scale such that $N(50, 100)$ and the other based on a scale such that $N(100, 225)$, where $N(\mu, \sigma^2)$ represents a normally distributed quantity with population mean μ and population variance σ^2. Even though scores on the two measures are not directly comparable, they can be transformed so that (a) the first measure is reported in terms of the second, (b) the second measure reported in terms of the first, or (c) both measures can be transformed so that they are in a common metric different from either of the two original measures (e.g., one where $N(0, 1)$). Of these transformation possibilities, the third approach is generally taken and thus effects are reported in standardized units where the values of the measurement scale can be regarded as a scale-free number that represents a pure measure of the magnitude of the effect.

In the BESS, a large debate has been underway for some time about the importance of reporting effect sizes and confidence intervals (e.g., Schmidt 1996; Meehl 1997; Thompson 2002; Cohen 1994; Kline 2004, and the references contained therein) rather than only the dichotomous reject or fail-to-reject decision from a null hypothesis significance test (see Krantz 1999, for a review of the tension that sometimes exists between statisticians and methodologists regarding this debate). On the surface, it seems there is no reason not to report effect sizes and their corresponding confidence intervals. However, effects sizes based on raw scores are not always helpful or generalizable due to the lack of natural scaling metrics and multiple scales existing for the same phenomenon in the BESS. A common methodological suggestion in the BESS is to report standardized effect sizes in order to facilitate the interpretation of results and for the cumulation of scientific knowledge across studies, which is the goal of meta-analysis (e.g., Hunter and Schmidt 2004; Glass, McGaw, and Smith 1981; Hedges and Olkin 1985). A standardized effect size is an effect size that describes the size of the effect but that does not depend on any particular measurement scale. A standardized effect size thus represents a pure number, in the sense that the magnitude of the effect is not wedded to a particular scale. The act of standardization is not, however, generally based on a set of known population parameters, but rather the standardization process is based on sample statistics. This is the case for two reasons: (a) many measures have not been normed in order to determine the population parameters of interest for the population in general or for a subpopulation of interest and (b) it is generally desirable to base the standardized effect size on the particular sample rather than mixing population parameters with sample statistics.

2. CI construction

Seminal work by Neyman (1935; 1937) laid the foundation for confidence interval formation. Recall that the general procedure for a confidence interval yields an upper and lower limit, such that the probability that the *fixed* parameter is contained within a *random* interval is $1-\alpha$, where α is the Type I error rate and $1-\alpha$ is the confidence level coverage. The general form of the confidence interval is given as

$$p[\theta_{\mathrm{L}}(\mathbf{X}) \leq \theta \leq \theta_{\mathrm{U}}(\mathbf{X})] = 1 - \alpha, \tag{1}$$

where θ is some parameter of interest, $\theta_{\mathrm{L}}(\mathbf{X})$ and $\theta_{\mathrm{U}}(\mathbf{X})$ are the lower and upper random confidence limits, respectively, which are based on the observed data, $\mathbf{X}$, and p denotes

probability. For notational ease $\theta_{\mathrm{L}}(\mathbf{X})$ and $\theta_{\mathrm{U}}(\mathbf{X})$ will be denoted θ_{L} and θ_{U}, respectively, with the understanding that the lower and upper confidence limits are random because they depend on the random data.

Two sections follow that discuss the ways in which θ_{L} and θ_{U} can be calculated for confidence intervals for different types of effect sizes commonly used in the BESS. The first approach is the standard approach that is generally given in applied texts and implemented in statistical software programs. The second approach, however, is more difficult conceptually and computationally than the first approach and is not generally discussed in applied texts nor implemented in statistical software without specialized programming scripts. The second approach is ultimately what is of interest to the present work, with the first approach given to provide a context for the second approach.

2.1. CIs for pivotal quantities

Suppose the population standard deviation is known for a normally distributed population with some unknown mean for a sample of size N, the following inequality holds with $1 - \alpha$ probability

$$p\left[z_{(\alpha/2)} \leq \frac{M - \mu}{\sigma_M} \leq z_{(1-\alpha/2)}\right] = 1 - \alpha \tag{2}$$

where M is the sample mean and σ_M is the population standard deviation of the sampling distribution of the mean, defined as $\sigma/\sqrt{N}$, and $z_{(\cdot)}$ represents the quantile from the standard normal distribution at the subscripted probability value. The inequality contained within the brackets can be manipulated by first multiplying the inequality by σ_M, which removes the value in the denominator of the inequality's center:

$$p\left[z_{(\alpha/2)}\sigma_M \leq M - \mu \leq z_{(1-\alpha/2)}\sigma_M\right] = 1 - \alpha. \tag{3}$$

The value of M from the center of the inequality can be removed by subtracting M from the center and both sides:

$$p\left[z_{(\alpha/2)}\sigma_M - M \leq -\mu \leq z_{(1-\alpha/2)}\sigma_M - M\right] = 1 - \alpha. \tag{4}$$

Multiplying the inequality by -1 to make $-\mu$ positive—requiring the inequalities to be reversed as is the case when inequalities are multiplied by a negative value—the resultant equation can be given as

$$p\left[-z_{(1-\alpha/2)}\sigma_M + M \leq \mu \leq -z_{(\alpha/2)}\sigma_M + M\right] = 1 - \alpha. \tag{5}$$

Further manipulation of the inequality yields

$$p\left[M - z_{(1-\alpha/2)}\sigma_M \leq \mu \leq M - z_{(\alpha/2)}\sigma_M\right] = 1 - \alpha. \tag{6}$$

Because $z_{(\alpha/2)}$ is always negative and the normal distribution is symmetric, $-1 \times z_{(\alpha/2)}$ is equivalent to $z_{(1-\alpha/2)}$ and the right hand side of the inequality thus reduces to $M + z_{(1-\alpha/2)}$, which implies that Equation 6 can be written as

$$p\left[M - z_{(1-\alpha/2)}\sigma_M \leq \mu \leq M + z_{(1-\alpha/2)}\sigma_M\right] = 1 - \alpha. \tag{7}$$

Although often optimal, in the sense that the confidence interval is as narrow as possible (Casella and Berger 2002), there is no reason to restrict the confidence intervals to those where the lower and upper rejection region are equal. The specified α can be conceptualized as consisting of two parts, α_{L} and α_{U}, where α_{L} is the rejection region for the lower limit and α_{U} the rejection region for the upper limit (i.e., the proportion of time that the parameter will be less than the lower confidence limit or greater than the upper confidence limit, respectively). Thus, more generally, the confidence interval formation given in Equation 7 can be written as

$$p\left[M + z_{(\alpha_{\mathrm{L}})}\sigma_M \leq \mu \leq M + z_{(1-\alpha_{\mathrm{U}})}\sigma_M\right] = 1 - (\alpha_L + \alpha_U), \tag{8}$$

or for convenience

$$p\left[M - z_{(1-\alpha_{\mathrm{L}})}\sigma_M \leq \mu \leq M + z_{(1-\alpha_{\mathrm{U}})}\sigma_M\right] = 1 - (\alpha_L + \alpha_U), \tag{9}$$

where $\alpha_{\mathrm{L}} + \alpha_{\mathrm{U}} = \alpha$. The confidence interval method given in Equation 9, or a closely related rewriting, is what is typically given in applied texts. Generally discussed is only one type of special case of nonequal rejection regions where $\alpha_{\mathrm{L}} \neq \alpha_{\mathrm{U}}$: for one sided confidence intervals where α_{L} or α_{U} is set to zero.

The equations above each assumed that σ was known. In almost all applications, σ is unknown and it is necessary to use a (central) t-distribution with the appropriate degrees of freedom instead of basing the confidence interval on critical values from the standard normal distribution. With unknown σ, the analog of Equation 2 would be

$$p\left[t_{(\alpha/2;\nu)} \leq \frac{M - \mu}{s_M} \leq t_{(1-\alpha/2;\nu)}\right] = 1 - \alpha, \tag{10}$$

where s_M is the estimated standard deviation of the sampling distribution of the mean defined as $s/\sqrt{N}$ with s being the square root of the unbiased estimate of the variance, ν are the degrees of freedom, which are $N-1$ in the context of single sample designs when σ is unknown, and $t_{(\alpha/2;\nu)}$ is the $\alpha/2$ quantile from a t-distribution with ν degrees of freedom. Through a set of manipulations and reductions analogous to Equations 2 through 8, a $(1-\alpha)100\%$ confidence interval can be obtained for μ when σ is unknown, which is given as

$$p\left[M + t_{(\alpha_{\mathrm{L}};\nu)}s_M \leq \mu \leq M + t_{(1-\alpha_{\mathrm{U}};\nu)}s_M\right] = 1 - (\alpha_L + \alpha_U), \tag{11}$$

or for convenience

$$p\left[M - t_{(1-\alpha_{\mathrm{L}};\nu)}s_M \leq \mu \leq M + t_{(1-\alpha_{\mathrm{U}};\nu)}s_M\right] = 1 - (\alpha_L + \alpha_U). \tag{12}$$

Notice that in Equation 2 the quantity in the center of the inequality is simply a z-test statistic, whereas in Equation 10 the quantity in the center of the inequality is simply a t-test statistic. The logic of transforming a probabilistic statement for a particular z-test or t-test statistic to a confidence interval for μ was possible in the manner done because the center of the inequality could be reduced to only the population parameter of interest (i.e., μ) and the interval did not depend on any unknown parameters. This procedure used to transform the probability statement to a confidence interval is known as *inverting the test statistic* (Casella and Berger 2002, Section 9.2.1; see also Kempthorne and Folks 1971, Section 13.3) because the $\alpha 100\%$ region of implausible parameter values (i.e., where $p < \alpha$ under the null hypothesis) is inverted to form the $(1-\alpha)100\%$ region of plausible parameter values (i.e.,

where $p > \alpha$ under the null hypothesis). Confidence intervals can be formed by inverting the test statistic when the quantity of interest is a *pivotal quantity*. A pivotal quantity, sometimes termed a pivot, is a quantity whose probability distribution does not depend on any unknown parameters (e.g., Casella and Berger 2002, Chapter 9; Stuart, Ord, and Arnold 1999, Chapter 19). When the test statistic cannot be transformed into a confidence interval by reducing the probabilistic statement concerning the test statistic into a probabilistic statement concerning the parameter, implying the quantity is not pivotal, a more general approaches to confidence interval formation is required. This more general approach to confidence interval formation is discussed in the next section.

2.2. CI formation for nonpivotal quantities

Although confidence intervals for commonly used pivotal quantities (e.g., the mean, mean difference, variance, regression coefficients, etc) are well known, many effects of interest in the BESS are not pivotal. In particular, standardized effect sizes (e.g., standardized mean differences, standardized regression coefficients, coefficients of variation, etc.) and effect sizes that are bounded (e.g., correlation coefficients, squared multiple correlation coefficients, proportions, etc.), which are considered standardized since they do not depend on the particular measurement scale are not generally pivotal quantities. Thus, confidence intervals for such effects cannot be obtained by inverting their corresponding test statistic, which was the method discussed in the previous section for a quantity that was pivotal. Many applied statistics texts either do not mention confidence intervals for such nonpivotal quantities or provide only approximate methods of confidence interval formation, sometimes without mentioning that the methods are approximations.

Confidence interval formation by inverting the test statistic requires that the effect of interest be a pivotal quantity. If not, another more general approach is required. This more general approach to confidence interval formation has been termed *pivoting the cumulative distribution function* (Casella and Berger 2002, Section 9.2.3; see also Kempthorne and Folks 1971, Section 13.4). When a quantity related to a test statistic is not pivotal, the sampling distribution of the estimate depends on an outside parameter that is almost certainly not known and implies that the test statistic cannot be inverted. The solution to such a problem is to find the value of the unknown parameter that leads to the observed cumulative probability of the test statistic being $1 - \alpha_{\mathrm{L}}$, which becomes the lower confidence limit of the parameter, and to find the value of the unknown parameter that leads to the observed cumulative probability of the test statistic having probability α_{U}, which becomes the upper confidence limit of the parameter. For effects of interest in the BESS, these unknown parameter values are generally noncentrality parameters (e.g., Steiger and Fouladi 1997; Cumming and Finch 2001; Smithson 2003; Steiger 2004).

For example, forming a confidence interval for μ when σ is unknown was shown in Equations 11. However, suppose what is of interest is forming a confidence interval for the population standardized mean,

$$\text{ᔑ} = \frac{\mu}{\sigma}, \tag{13}$$

which is estimated by replacing the parameters with their sample analogs,

$$m = \frac{M}{s}, \tag{14}$$

where m is the sample estimate of ם, the population standardized mean.[1] Because s (or s_M) cannot be pivoted, as s is necessary to standardize M, and the center of the quantity does not contain the population quantity, rewriting Equation 10 yields

$$p\left[t_{(\alpha/2;\nu)}/\sqrt{N} \leq \frac{M-\mu}{s} \leq t_{(1-\alpha/2;\nu)}/\sqrt{N}\right] = 1-\alpha, \tag{15}$$

where the $\sqrt{N}$ was removed from the center of the inequality by multiplying the inequality by $1/\sqrt{N}$. The lack of pivotibility of the quantity is due to s necessarily being involved in the center of the inequality since it is what standardizes the mean and the population value not being contained within the inequality. Furthermore, the population effect size is not contained within the center of the inequality. Because the test statistic cannot be pivoted, it is necessary to pivot the cumulative distribution function of the test statistic itself. Before pivoting the cumulative distribution function, a discussion of noncentral distributions is necessary.

The widely used (central) t-distribution is a special case of the more general noncentral t distribution, when the noncentrality parameter equals zero. Johnson, Kotz, and Balakrishnan (1995) provide a modern overview of noncentral t-distributions (see also Johnson and Welch, 1940). From Johnson *et al.* (1995), in the single sample situation when the distribution is normal and σ unknown, the population noncentrality parameter from a noncentral t-distribution can be written as

$$\lambda = \frac{\mu-\mu_0}{\sigma/\sqrt{N}} = \frac{\mu\sqrt{N}}{\sigma}, \tag{16}$$

where μ_0 is the specified value of the null hypothesis, which will be set to zero in the present context without loss of generality. Thus, there is a one-to-one relationship between the nonpivotable quantity in Equation 10 and the noncentrality parameter from a t-distribution. Although a known probability distribution does not literally exist for $(M-\mu)/s$, it is indirectly available via the noncentral t-distribution, which is denoted $t_{(\nu;\lambda)}$, where ν are the degrees of freedom and λ is the noncentrality parameter. The noncentrality parameter, which can be conceptualized as an index of the magnitude of difference between the null and alternative hypotheses, can be estimated as

$$\hat{\lambda} = \frac{M}{s}\sqrt{N}, \tag{17}$$

or, in terms of m,

$$\hat{\lambda} = m\sqrt{N}, \tag{18}$$

where Equation 17 and 18 are equivalent to the observed t-test statistic:

$$\hat{\lambda} = t = m\sqrt{N}. \tag{19}$$

Given the linkage between the test statistic, the standardized mean, and the noncentral t-distribution, it is helpful to discuss two important principles when forming confidence intervals for standardized effect sizes. Theses principles are the *confidence interval transformation principle* and the *inversion confidence interval principle*, both of which will be discussed momentarily. The names of these principles were coined and the concepts discussed in Steiger and Fouladi (1997) with a review given in Steiger (2004). Steiger and Fouladi (1997) did not

[1] The character ם, used to denote the population standardized mean, is the Phoenician letter mem, which was a precursor to the Greek letter "μ" and the Latin letter "m" (Powell 1996).

literally develop the theory behind these principles, rather their important work combined disparate statistical theory into a formal context for forming confidence intervals. When these principles are combined a set of powerful tools for confidence interval formation become available for standardized effect sizes that are related to a noncentrality parameter. Almost all of the effects commonly used in the BESS can be linked to a noncentrality parameter from t, F, or χ^2 distributions. Methods for forming confidence intervals for noncentrality parameters from t, F, and χ^2 distributions have been implemented in the *Methods for the Behavioral, Educational, and Social Sciences* (**MBESS**; Kelley 2007b,a) R package (R Development Core Team 2007).

The *confidence interval transformation principle* is beneficial for forming a confidence interval on a parameter that is monotonically related to another parameter, when the latter has a tractable method of obtaining the confidence interval whereas the former might not. Let $f(\theta)$ be a monotonic transformation of θ defined by the function $f(\cdot)$. The $(1-\alpha)100\%$ confidence limits for $f(\theta)$ are $f(\theta_{\mathrm{L}})$ and $f(\theta_{\mathrm{U}})$,

$$p\left[f(\theta_{\mathrm{L}}) \leq f(\theta) \leq f(\theta_{\mathrm{U}})\right] = 1 - \alpha. \tag{20}$$

Thus, for monotonically related parameters the confidence interval for the transformed population quantity is obtained by applying the same transformation to the limits of the confidence interval as was done to the population quantity itself (Steiger and Fouladi 1997; Steiger 2004).

The *inversion confidence interval principle* states that if $\hat{\theta}$ is an estimate of θ with a cumulative distribution that depends on ς, some necessary outside parameter, the probability of observing an estimate (i.e., $\hat{\theta}$) smaller than that obtained is given as $p(\hat{\theta}|\varsigma)$ (i.e., it is the cumulative probability). Calculation of a confidence interval for θ based on the inversion confidence interval principle involves finding θ_{L} such that $p(\hat{\theta}|\theta_{\mathrm{L}}) = 1 - \alpha_{\mathrm{L}}$ for the lower limit and θ_{U} such that $p(\hat{\theta}|\theta_{\mathrm{U}}) = \alpha_{\mathrm{U}}$ for the upper limit. The confidence interval for θ then has coverage of $1 - (\alpha_{\mathrm{L}} + \alpha_{\mathrm{U}})$ and is given as

$$p\left[\theta_{\mathrm{L}} \leq \theta \leq \theta_{\mathrm{U}}\right] = 1 - \alpha, \tag{21}$$

where θ is some parameter of interest with θ_{L} and θ_{U} being the lower and upper confidence interval limits, where θ_{L} will be greater than θ $\alpha_{\mathrm{L}}100\%$ of the time and $\theta_{\mathrm{U}}100\%$ will be less than θ $\alpha_{\mathrm{U}}100\%$ of the time. The confidence interval procedure is general and need not have equal rejection regions. For example, the most common confidence interval without equal rejection regions is obtained by setting α_{L} or α_{U} (whichever is appropriate for the specific situation) to zero, which is simple a one sided confidence interval.

Returning to the example of confidence interval formation for the standardized mean, it now becomes apparent that indeed, a λ_{L} value can be found such that $p(t|\lambda_{\mathrm{L}}) = 1 - \alpha_{\mathrm{L}}$ and a λ_{U} value can be found such that $p(t|\lambda_{\mathrm{U}}) = \alpha_{\mathrm{U}}$. Given the values of λ_{L} and λ_{U}, these noncentral values can be transformed into the metric of the standardized mean. Manipulation of Equation 18 shows that

$$m = \frac{t}{\sqrt{N}}. \tag{22}$$

λ_{L} and λ_{U} can be substituted for t in Equation 22, so that the lowest and highest plausible values of the standardized mean can be obtained given the specified values of α_{L} and α_{U}. Thus, the confidence interval for the standardized mean is given as

$$p\left[\frac{\lambda_{\mathrm{L}}}{\sqrt{N}} \leq \text{ᔑ} \leq \frac{\lambda_{\mathrm{U}}}{\sqrt{N}}\right] = 1 - \alpha. \tag{23}$$

This confidence interval is realized by first computing the confidence interval on the noncentrality parameter from a t-distribution and then transforming the limits of the confidence interval into the metric of the standardized mean.

For example, suppose $M = 50$, $s = 10$, and $N = 25$. The estimated noncentrality parameter (i.e., the estimated t-test statistic for the test of the null hypothesis that $\mu = 0$) from Equation 18 is $\hat{\lambda} = 25$. A 95% confidence interval for the population noncentrality parameter is

$$\text{CI}_{.95} = [17.68259 \leq \lambda \leq 32.6888]\,, \tag{24}$$

where $\text{CI}_{.95}$ represents the 95% confidence interval limits. Dividing the limits of the confidence interval by $\sqrt{N}$ will then transform the confidence limits for λ into confidence limits for ᵡ, which is given as

$$\text{CI}_{.95} = [3.536517 \leq \text{ᵡ} \leq 6.453777]\,. \tag{25}$$

Although there is no cumulative distribution function developed specifically for the standardized mean, by pivoting the cumulative distribution function for the noncentral t-distribution, limits of the confidence interval for ᵡ can be calculated. These limits are those that correspond to the theoretical cumulative probability distribution of the standardized mean for the specified level of confidence.

The present section has discussed a method of obtaining an exact confidence interval for a standardized mean that itself has no known probability distribution function. Although the logic and method is not overly complex, determining θ_{L} and θ_{U} has been a serious issue for some time. The difficulty in finding the necessary confidence limits, by finding noncentrality parameters that have the obtained statistic at the specified quantile, has led to tri-entry (probability, degrees of freedom, and noncentrality parameter) tables (see a review of such tables in Johnson *et al.* 1995, chapter 31) that will almost certainly yield approximate results in any applied situation. As will be discussed in the following section, **MBESS** has functions that return the exact values of θ_{L} and θ_{U} for the most general cases of noncentral t, F, and χ^2 distributions. Thus, any effect size (e.g., those that are standardized) that has a monotonic relation to the noncentrality parameter from one of these distributions can be formed with **MBESS**. Within **MBESS**, many commonly used standardized effect sizes have functions that compute their respective confidence intervals directly (instead of the user forming a confidence interval for the corresponding noncentrality parameter and then transforming to the scale of the effect size). Some of the most popular standardized effect sizes will be discussed, and examples using **MBESS** given, in the remainder of the article in the analysis of variance (ANOVA) and regression contexts.

CIs for the standardized mean in **MBESS**

The `ci.sm()` function from **MBESS** can be used to form confidence intervals for the population standardized mean (i.e., ᵡ). Although other optional specifications exist, a basic specification of the function `ci.sm()` would be of the form

`ci.sm(sm=`m`, N=`N`, conf.level=`$1-\alpha$`)`

where `sm` is the observed standardized mean, `N` is the sample size, and `conf.level` is the confidence level coverage.

3. CIs for standardized effect sizes in an ANOVA context

The comparison of means is a commonly used technique in the BESS, as research questions are many times related to issues involving means and mean differences. This section is organized with three subsections that deal with standardized effect sizes: one concerning the mean difference of two groups, one concerning omnibus effects when several group means are involved, and a section concerning targeted effects (i.e., comparisons) when several group means are involved.

3.1. The standardized mean difference for two independent groups

One of the most commonly used effect sizes in the BESS is the standardized mean difference. The population standardized mean difference is defined as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \tag{26}$$

and is generally estimated by

$$d = \frac{M_1 - M_2}{s}, \tag{27}$$

where μ_1 and μ_2 are the population means of group 1 and group 2, respectively, with M_1 and M_2 as their respective estimates, and s is the square root of the unbiased estimate of the within group variance, which is estimated as

$$s = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}} \tag{28}$$

with per group sample sizes of n_1 and n_2 for group 1 and group 2, respectively, and with s_1^2 and s_2^2 being the unbiased estimate of the variance for group 1 and group 2, respectively, assuming $\sigma_1^2 = \sigma_2^2$. The typical two-group t-test is defined as

$$t = \frac{M_1 - M_2}{s_{M_1 - M_2}}, \tag{29}$$

where $s_{M_1 - M_2}$ is the standard error of the mean difference and is given as

$$s_{M_1 - M_2} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \tag{30}$$

which has $N - 2$ degrees of freedom ($N = n_1 + n_2$).

Notice that the difference between d from Equation 27 and the two-group t-test statistic from Equation 29 is the quantity $\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ contained in the denominator of the t-test statistic (implicit in $s_{M_1 - M_2}$), which is multiplied by s to estimate the standard error of the mean difference. Because $\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ can be rewritten as $\sqrt{\frac{n_2 + n_1}{n_1 n_2}}$, multiplying the inverse of this quantity by d leads to an equivalent representation of the t-test statistic:

$$t = d\sqrt{\frac{n_1 n_2}{n_2 + n_1}}. \tag{31}$$

Given Equation 31, it can be seen that Equation 27 can be written as

$$d = t\sqrt{\frac{n_2 + n_1}{n_1 n_2}}. \tag{32}$$

The population noncentrality parameter for the two-group t-test is given as

$$\begin{aligned} \lambda &= \frac{\mu_1 - \mu_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} && (33) \\ &= \delta\sqrt{\frac{n_1 n_2}{n_2 + n_1}}, && (34) \end{aligned}$$

which is estimated, as in the single sample situation, with the observed t-test statistic. Thus,

$$\begin{aligned} \hat{\lambda} &= \frac{M_1 - M_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} && (35) \\ &= d\sqrt{\frac{n_1 n_2}{n_2 + n_1}} && (36) \\ &= t && (37) \end{aligned}$$

The analog of Equation 10 in the two group situation is

$$p\left[t_{(\alpha_\mathrm{L};\nu)} \leq \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s\sqrt{\frac{n_2+n_1}{n_1 n_2}}} \leq t_{(1-\alpha_\mathrm{U};\nu)}\right] = 1 - \alpha, \tag{38}$$

The test statistic can be pivoted such that the confidence interval for the (unstandardized) mean difference, which is discussed in many sources is given as

$$p\left[(M_1 - M_2) - t_{(1-\alpha_\mathrm{L};\nu)}s\sqrt{\frac{n_2 + n_1}{n_1 n_2}} \leq \mu_1 - \mu_2 \leq (M_1 - M_2) + t_{(1-\alpha_\mathrm{U};\nu)}s\sqrt{\frac{n_2 + n_1}{n_1 n_2}}\right] = 1-\alpha. \tag{39}$$

However, when what is of interest is the standardized mean difference, Equation 38 cannot be pivoted as was done in Equation 39. As discussed in the single sample situation, in order to form a confidence interval for δ, a confidence interval for λ is found and then the limits transformed into the scale of δ by using Equation 32. Thus, the value of λ_L is found such that $p(\hat{\lambda}|\lambda_\mathrm{L}) = 1 - \alpha_\mathrm{L}$ and the value λ_U is found such that λ_U $p(\hat{\lambda}|\lambda_\mathrm{U}) = \alpha_\mathrm{U}$, in exactly the same manner as they were found in the single sample situation, the difference being the noncentrality parameter is from the two-group context with corresponding degrees of freedom $N - 2$.

Given λ_L and λ_U, the confidence interval for λ is given as

$$p\left[\lambda_\mathrm{L} \leq \lambda \leq \lambda_\mathrm{U}\right] = 1 - \alpha, \tag{40}$$

which is generally of interest only in so much as its transformation allows for a confidence interval to be formed for δ, as is given by the following:

$$p\left[\lambda_\mathrm{L}\sqrt{\frac{n_2 + n_1}{n_1 n_2}} \leq \delta \leq \lambda_\mathrm{U}\sqrt{\frac{n_2 + n_1}{n_1 n_2}}\right] = 1 - \alpha. \tag{41}$$

Confidence intervals for the standardized mean difference were detailed in Steiger and Fouladi (1997) and various aspects of such confidence intervals have been discussed in Cumming and Finch (2001), Kelley (2005), Smithson (2003), Algina, Keselman, and Penfield (2006) and Steiger (2004).

CIs for the standardized mean difference in **MBESS**

The `ci.smd()` function from **MBESS** can be used to form confidence intervals for the population standardized mean difference (i.e., δ). Although other optional specifications exist, a basic specification of the function `ci.smd()` would be of the form

`ci.smd(smd=`d`, n.1=`n_1`, n.2=`n_2`, conf.level=`$1-\alpha$`)`

where `smd` is the observed standardized mean difference, `n.1` and `n.2` are the sample sizes for group 1 and group 2, respectively, and `conf.level` is the desired confidence level coverage.

3.2. Standardized effect sizes for omnibus effects in an ANOVA context

Examining the differences among several group means is commonly done in the BESS. For example, the depression level of mildly depressed individuals might be examined as a function of whether or not the individuals were randomly assigned to the control group, the counseling only group, the drug only group, or the combination of counseling and drug group. Both targeted effects, such as the mean difference between the counseling only and the drug only group might be of interest, and omnibus effects, such as the proportion of variance on the depression scores is accounted for by the grouping factor, have standardized effect sizes that have corresponding confidence intervals for their population values. Whereas a targeted effect of interest might be a follow-up comparison, an omnibus effect of interest might be the ratio of the group sums of squares to the total sums of squares. So as to have measures of effect that are not wedded to a particular measurement scale, this section reviews selected standardized effects sizes in the ANOVA context, shows their relation to the noncentrality parameter, and illustrates how such confidence intervals for population standardized effect sizes can be easily obtained with **MBESS**.

In a precursor of suggestions that were to be emphasized nearly a generation later in the BESS, Fleishman (1980) discussed omnibus effect sizes and their corresponding confidence intervals in an ANOVA context. Fleishman (1980) showed the relationship between certain ANOVA effects sizes and their corresponding noncentrality parameters from noncentral F-distributions. Given what was akin to the confidence interval inversion principle and the inversion confidence interval principle, confidence intervals for these effect sizes can be formed given confidence intervals for noncentral F-parameters. Such confidence intervals can easily be obtained in **MBESS**, as will be illustrated momentarily for selected effect sizes.

Let Λ_{p} be the noncentrality parameter for the pth factor of a noncentral F-distribution in a multi-factor (or single factor) fixed effects ANOVA, which is given as

$$\Lambda_{\mathrm{p}} = \frac{\sum_{j=1}^{J} n_{\mathrm{p}j}\tau_{\mathrm{p}j}^2}{\sigma_\epsilon^2}, \tag{42}$$

where $n_{\mathrm{p}j}$ is the sample size of the jth group of factor p ($j = 1, \ldots, J; N = \sum n_{\mathrm{p}j}$), $\tau_{\mathrm{p}j}$ is the effect associated with being in the jth level of factor p, which is defined as

$$\tau_{\mathrm{p}j} = \mu_{\mathrm{p}j} - \mu, \tag{43}$$

with $\mu_{\mathrm{p}j}$ being the population mean of the jth level of factor p and μ the overall population mean, and σ^2_ϵ is the mean square error (Fleishman 1980).[2] Alternatively, Equation 42 can be written as

$$\Lambda_{\mathrm{p}} = \frac{N\sigma^2_{\mathrm{p}}}{\sigma^2_\epsilon}, \tag{44}$$

where σ^2_{p} is the variance due to factor p:

$$\sigma^2_{\mathrm{p}} = \frac{\sum_{j=1}^{J} n_{\mathrm{p}j}\tau^2_{\mathrm{p}j}}{N}. \tag{45}$$

Notice that in a single factor design the equations reduce, specifically the p subscript in each of the equations is ignored, and J is simply the number of groups.

Two effect sizes closely related to Λ_{p}, and to one another, are the signal-to-noise ratio and the proportion of variance in the dependent variable that is accounted for by knowing the level of the factor (or group status in a single factor design; e.g., Fleishman 1980; Hays 1994; Cohen 1988). Formally the signal-to-noise ratio is defined as,

$$\phi^2_{\mathrm{p}} = \frac{\sigma^2_{\mathrm{p}}}{\sigma^2_\epsilon} \tag{46}$$

$$= \frac{\Lambda_{\mathrm{p}}}{N} \tag{47}$$

and the proportion of variance in Y accounted for by knowing the level of the factor (or group status in a single factor design) is defined as

$$\eta^2_{\mathrm{p}} = \frac{\sigma^2_{\mathrm{p}}}{\sigma^2_T} \tag{48}$$

$$= \frac{\sigma^2_{\mathrm{p}}}{\sigma^2_\epsilon + \sigma^2_{\mathrm{p}}} \tag{49}$$

$$= \frac{\Lambda_{\mathrm{p}}}{\Lambda_{\mathrm{p}} + N}, \tag{50}$$

where σ^2_T is the total variance of the dependent variable. There is also a close relation between ϕ^2_{p} and η^2_{p}:

$$\eta^2_{\mathrm{p}} = \frac{1}{1 + \frac{1}{\phi^2_{\mathrm{p}}}}. \tag{51}$$

$$= \frac{\phi^2_{\mathrm{p}}}{1 + \phi^2_{\mathrm{p}}}. \tag{52}$$

Just as in the situation of the noncentral t a Λ_{L} value can be found such that $p(F|\Lambda_{\mathrm{L}}) = 1-\alpha_{\mathrm{L}}$ and a Λ_{U} value can be found such that $p(F|\Lambda_{\mathrm{U}}) = \alpha_{\mathrm{U}}$, where F is the value of the F-test statistic for factor p from the factorial ANOVA procedure (or simply the F-test statistic from a single factor ANOVA). Given Λ_{L} and Λ_{U}, a confidence interval for Λ can be formed,

$$p\left[\Lambda_{\mathrm{L}} \leq \Lambda_{\mathrm{p}} \leq \Lambda_{\mathrm{U}}\right] = 1 - \alpha, \tag{53}$$

[2] Notice that the p representing the factor has not been italicized. This is the case to emphasize that these methods are for fixed effect ANOVA models.

which is generally of interest only in so much as its transformation allows for a confidence interval to be formed for ϕ^2_{p}, η^2_{p}, and/or possibly other effects from an ANOVA context. The confidence intervals of interest are transformations of Equation 53 by manipulating Equations 46 and 48 in order to transform the noncentrality parameter into the effect size of interest. Thus, a confidence interval for ϕ^2_{p} is given as

$$p\left[\frac{\Lambda_{\text{L}}}{N} \leq \phi^2_{\text{p}} \leq \frac{\Lambda_{\text{U}}}{N}\right] = 1 - \alpha \tag{54}$$

and a confidence interval for η^2_{p} is given as

$$p\left[\frac{\Lambda_{\text{L}}}{\Lambda_{\text{L}} + N} \leq \eta^2_{\text{p}} \leq \frac{\Lambda_{\text{U}}}{\Lambda_{\text{L}} + N}\right] = 1 - \alpha. \tag{55}$$

The square root of the signal-to-noise ratio also has a substantively appealing interpretation, as the standard deviation of the standardized means (e.g., p. 275 of Cohen 1988; Steiger 2004). Such a measure can easily be obtained due to the confidence interval transformation principal simply by taking the square root of the confidence limits for ϕ^2_{p}:

$$p\left[\sqrt{\frac{\Lambda_{\text{L}}}{N}} \leq \phi_{\text{p}} \leq \sqrt{\frac{\Lambda_{\text{U}}}{N}}\right] = 1 - \alpha. \tag{56}$$

CIs for standardized omnibus effects in ANOVA with **MBESS**

The `ci.snr()` function from **MBESS** can be used to form confidence intervals for the population signal-to-noise ratio (i.e., ϕ^2_{p}) for the pth fixed effects factor in an ANOVA setting. Although other optional specifications exist, a basic specification of the function `ci.snr()` would be of the form

`ci.snr(F.value=`F`, df.1, df.2, N=`N`, conf.level=`$1-\alpha$`),`

where `F.value` is the observed F value, `df.1` is the numerator degrees of freedom for the particular F-test statistic, `df.2` is the denominator degrees of freedom of the F-test statistic, N is the total sample size, and `conf.level` is the desired confidence level coverage.

The `ci.pvaf()` function from **MBESS** can be used to form confidence intervals for the population proportion of variance accounted for in the dependent variable by knowing group status (i.e., η^2_{p}) for the pth fixed effects factor in an ANOVA setting. Although other optional specifications exist, a basic specification of the function `ci.pvaf()` would be of the form

`ci.pvaf(F.value=`F`, df.1, df.2, N=`N`, conf.level=`$1-\alpha$`),`

where all of the input parameters are the same as for the `ci.snr` function.

The `ci.srsnr()` function from **MBESS** can be used to form confidence intervals for the square root of the signal-to-noise ratio for the pth fixed effects factor (i.e., ϕ_{p}) in an ANOVA setting. Although other optional specifications exist, a basic specification of the function `ci.srsnr()` would be of the form

`ci.srsnr(F.value=`F`, df.1, df.2, N=`N`, conf.level=`$1-\alpha$`),`

where all of the arguments are the same as for the `ci.snr()` and `ci.pvaf()` function.

3.3. Standardized effect sizes for targeted effects in an ANOVA context

The methods discussed in the previous section have been for omnibus measures of effect. However, targeted effects that have been standardized in an ANOVA context can also be easily implemented with the methods that have been discussed. Like many research design texts, Maxwell and Delaney (2004) discuss methods of forming comparisons among means to determine if a null hypothesis that a particular contrast (e.g., $\mu_1 - \mu_2 = 0$, $(\mu_1 + \mu_2)/2 - \mu_3 = 0$, $(\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2 = 0$, etc.) equals zero can be rejected. The test statistic can be given as

$$t = \frac{\hat{\Psi}}{\sqrt{s_\epsilon^2 \sum_{j=1}^{J} \left(\frac{c_j^2}{n_j}\right)}}, \tag{57}$$

where $\hat{\Psi}$ is given as

$$\hat{\Psi} = \sum_{j=1}^{J} c_j M_j \tag{58}$$

and s_ϵ^2 is the mean square error $\left(\text{note that } \sum_{j=1}^{J} c_j = 0, \text{ as is the case with all comparisons}\right)$. Notice that this is simply a t-test for a targeted contrast. The t-test in Equation 57 can be pivoted in order to form a confidence interval for Ψ.

As detailed in Steiger (2004), hypothesis tests for comparisons of the form Equation 57 can be standardized. As before, the process of standardization leads to a test statistic that cannot be pivoted, implying that the confidence interval for the standardized effect (ψ) is not directly available given the confidence limits for the unstandardized effect (Ψ). This issue is literally just an extension of the methods discussed in the context of two independent groups, where a confidence interval for δ was formed. The way in which $\hat{\Psi}$ is standardized only involves division by s_ϵ, the root mean square error, which yields a population standardized comparison given as

$$\psi = \frac{\Psi}{\sigma_\epsilon}. \tag{59}$$

The noncentrality parameter from a t-distribution in this context is thus given as

$$\lambda = \frac{\psi}{\sqrt{\sum_{j=1}^{J} \left(\frac{c_j^2}{n_j}\right)}}, \tag{60}$$

which is simply the t-test statistic if the population values replaced the sample estimates, as was done in Equation 57. Thus, when a confidence interval is found for λ, the limits can be transformed into a confidence interval for ψ (which is equal to δ when $J = 2$ because the ANOVA reduces to an independent groups t-test) by setting Equation 60 equal to ψ:

$$\psi = \lambda \sqrt{\sum_{j=1}^{J} \left(\frac{c_j^2}{n_j}\right)}. \tag{61}$$

Thus, a confidence interval for ψ is given as

$$p\left[\lambda_{\mathrm{L}} \sqrt{\sum_{j=1}^{J} \left(\frac{c_j^2}{n_j}\right)} \leq \psi \leq \lambda_{\mathrm{U}} \sqrt{\sum_{j=1}^{J} \left(\frac{c_j^2}{n_j}\right)}\right] = 1 - \alpha. \tag{62}$$

CIs for standardized targeted effects in ANOVA with **MBESS**

The `ci.sc()` function from **MBESS** can be used to form confidence intervals for the population standardize comparison (i.e., ψ) in an ANOVA setting. Although other optional specifications exist, a basic specification of the function `ci.sc()` would be of the form

`ci.sc(means=c(`$M_1, M_2, \ldots, M_J$`), error.variance=`s^2_ϵ`, c.weights=c(`$c_1, c_2, \ldots, c_J$`), n=c(`$n_1, n_2, \ldots, n_J$`), N=`N`, conf.level=`$1 - \alpha$`).`

where `means` is a vector of the J group means, `error.variance` is the mean square error, `c.weights` is a vector of the J contrast weights (that must sum to zero), `n` is the vector of the J levels of sample sizes (or group sample sizes in a single factor design), and N is the total sample size (which need not be specified in single factor designs).

4. CIs for standardized effect sizes in a regression context

Multiple and simple regression are very popular methods in the BESS, especially for observational research, as research questions many times involve how a set of regressor (predictor/independent/explanatory) variables influence or explain a criterion (predicted/outcome/dependent) variable. This section is organized in two sections: one concerning the omnibus effect (i.e., the squared multiple correlation coefficient) and another section concerning targeted effects (i.e., individual regression coefficients).

4.1. Standardized effect sizes for omnibus effects in multiple regression

In the special case of fixed regressor variables, where the values of the regressors are selected a priori as part of the research design, the population proportion of variance in Y that is accounted for by the K regressors is the squared multiple correlation coefficient, denoted P^2. Notice that P^2 from a regression context is equivalent to the η^2 statistic discussed in the ANOVA context. This comes as no surprise, as both ANOVA and multiple regression are special cases of the general linear model. The population proportion of variance accounted for (in the regression context) can be written as

$$\mathrm{P}^2 = \frac{\sigma^2_Y - \sigma^2_\epsilon}{\sigma^2_Y} \tag{63}$$

$$= \frac{\sigma^2_{Y \cdot \mathbf{X}}}{\sigma^2_Y} \tag{64}$$

$$= \frac{\boldsymbol{\sigma}'_{\mathrm{XY}} \boldsymbol{\Sigma}^{-1}_{\mathbf{X}} \boldsymbol{\sigma}_{\mathrm{XY}}}{\sigma^2_Y} \tag{65}$$

where σ^2_Y is the variance of Y, $\sigma^2_{Y \cdot \mathbf{X}}$ is the variance of Y as predicted from the K X variables (i.e., the variance of $\hat{Y}$), σ^2_ϵ is the error variance (i.e., the variance of $Y - \hat{Y}$), $\boldsymbol{\Sigma}_{\mathbf{X}}$ is the population covariance matrix of the K X variables, and $\sigma_{\mathbf{X}Y}$ is the vector of population covariances between the K X variables and Y. Of course, in addition to the omnibus effect of P^2, the targeted effects provided by the regression coefficients are also of interest. The vector of K population regression coefficients, excluding the intercept, are obtained as

$$\boldsymbol{\mathcal{B}} = \boldsymbol{\Sigma}^{-1}_{\mathbf{X}} \boldsymbol{\sigma}_{\mathbf{X}Y} \tag{66}$$

with the population intercept given as

$$\mathcal{B}_0 = \mu_Y - \boldsymbol{\mu}_{\mathbf{X}}'\boldsymbol{\mathcal{B}}. \tag{67}$$

The test of the null hypothesis that $\mathrm{P}^2 = 0$ is given as

$$F = \frac{R^2/K}{(1-R^2)/(N-K-1)} = \left(\frac{R^2}{1-R^2}\right)\left(\frac{N-K-1}{K}\right), \tag{68}$$

where R^2 is the sample estimate of P^2 (i.e., the observed proportion of variance in Y accounted for by the K predictors) with K and $N-K-1$ degrees of freedom. A noncentral χ^2 in the numerator and a central χ^2 in the denominator of an F-ratio has a sampling distribution that follows a noncentral F-distribution with the same numerator and denominator degrees of freedom as the χ^2s and noncentrality parameter equal to that of the χ^2 noncentrality parameter in the numerator (Patel and Read 1982; Stuart *et al.* 1999; Rencher 2000). Thus, when the null hypothesis is false (i.e., $\mathrm{P}^2 > 0$), the sampling distribution of the F-test statistic in Equation 68 follows a noncentral F-distribution with K and $N-K-1$ degrees of freedom and noncentrality parameter Λ. It can be shown (e.g., Rencher 2000; Stuart *et al.* 1999) that when the null hypothesis that $\mathrm{P}^2 = 0$ is false the noncentrality parameter of the F-test statistic for the test of the null hypothesis that $\mathrm{P}^2 = 0$ can be written as

$$\Lambda = \frac{\boldsymbol{\mathcal{B}}'\boldsymbol{\Sigma}_{\mathbf{X}}N\boldsymbol{\mathcal{B}}}{\sigma_Y^2(1-\mathrm{P}^2)}. \tag{69}$$

Substituting $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\sigma}_{\mathbf{X}Y}$ for the definition of $\boldsymbol{\mathcal{B}}$ from Equation 66, Equation 69 can be rewritten as

$$\Lambda = \frac{\boldsymbol{\sigma}_{\mathbf{X}Y}'\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Sigma}_{\mathbf{X}}N\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\sigma}_{\mathbf{X}Y}}{\sigma_Y^2(1-\mathrm{P}^2)}. \tag{70}$$

Equation 70 can itself be reduced and rewritten as

$$\Lambda = \left(\frac{\boldsymbol{\sigma}_{\mathrm{X}Y}'\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\sigma}_{\mathrm{X}Y}}{\sigma_Y^2}\right)\left(\frac{N}{1-\mathrm{P}^2}\right), \tag{71}$$

which reduces, due to Equation 65, to

$$\Lambda = \left(\frac{\mathrm{P}^2}{1}\right)\left(\frac{N}{1-\mathrm{P}^2}\right), \tag{72}$$

and finally

$$\Lambda = \left(\frac{\mathrm{P}^2}{1-\mathrm{P}^2}\right)N. \tag{73}$$

The quantity in parentheses in Equation 73 is

$$\phi^2 = \frac{\mathrm{P}^2}{1-\mathrm{P}^2}, \tag{74}$$

the same signal-to-noise ratio discussed in the ANOVA section (see Equation 46). Notice that P^2 can be written as a function of Λ by solving for P^2 in Equation 73:

$$\mathrm{P}^2 = \frac{\Lambda}{N+\Lambda}. \tag{75}$$

Given that the relationship between the F-test statistic and Λ as well as the relationship between P^2 and Λ, forming confidence intervals for P^2 is simply a matter of solving for Λ_L and Λ_U, so that $p(F|\Lambda_\mathrm{L}) = 1 - \alpha_\mathrm{L}$ and $p(F|\Lambda_\mathrm{U}) = \alpha_\mathrm{U}$ can be found. The confidence limits for Λ can then be transformed to the units of P^2 with Equation 75 in accord with the confidence interval transformation principle as

$$p\left[\frac{\Lambda_\mathrm{L}}{N + \Lambda_\mathrm{L}} \leq \mathrm{P}^2 \leq \frac{\Lambda_\mathrm{U}}{N + \Lambda_\mathrm{U}}\right] = 1 - \alpha. \tag{76}$$

Of course, the square root of the confidence limits can be taken so that the confidence interval is for the multiple correlation coefficient, P, which would be given as

$$p\left[\sqrt{\frac{\Lambda_\mathrm{L}}{N + \Lambda_\mathrm{L}}} \leq \mathrm{P} \leq \sqrt{\frac{\Lambda_\mathrm{U}}{N + \Lambda_\mathrm{U}}}\right] = 1 - \alpha. \tag{77}$$

Although not often discussed in applied texts, there are differences in the sampling distribution of R^2 when predictors are fixed compared to when predictors are random (e.g., Sampson 1974; Gatsonis and Sampson 1989; Rencher 2000). Even though the sampling distributions of R^2 in the fixed and random predictor case are very similar when the null hypothesis that $\mathrm{P}^2 = 0$ is true, they can be quite different when the null hypothesis is false. The method discussed thus far for confidence interval formation assumed that the predictors were fixed. However, in most applications of multiple regression in the BESS, regressor variables are random. Basing the confidence interval formation procedure on fixed regressors when regressors are random will thus lead to a nominal (specified) Type I error rate that differs from the empirical (actual) Type I error rate, something that is never desirable.

Confidence intervals for P^2 under the case of random regressors have not often been considered. Lee (1971) and Ding (1996) provide algorithms for computing various properties of the distribution of R^2 for random regressors, which can be used when forming confidence intervals. The method of Lee (1971) was implemented by Steiger and Fouladi (1992) in an early stand-alone program, **R2**. Algina and Olejnik (2000) provide a summary of the approach suggested by Lee (1971) as well as a SAS script that can be used to implement the procedure discussed in Lee (1971). **MBESS** also allows for confidence intervals for P^2 in the context of random predictor variables based on Lee (1971) and Algina and Olejnik (2000), which gives results that are consistent with the **R2** program of Steiger and Fouladi (1992). The confidence interval procedure for P^2 when regressors are random is conceptually similar to that discussed in the fixed case, however, the sampling distribution of the F-test statistics does not follow a noncentral F-distribution. Rather, Fisher (1928) showed that the sampling distribution of R^2 follows a Gauss hypergeometric distribution when predictors are random. Lee (1971) used an approximate of the sampling distribution of $R^2/(1 - R^2)$ (i.e., the sample estimate of ϕ^2 from Equation 74), which is monotonically related to the sampling distribution of R^2 when predictors are random. The sampling distribution of the observed signal-to-noise ratio is estimated with a three moment approximation using noncentral F-distributions based on an iterative scheme. The method of Lee (1971) is quite accurate in terms of the empirical and nominal level of confidence interval coverage. Although not many details have been included, the technical underpinnings of confidence intervals for P^2 when regressors are random is quite difficult. As will be shown, **MBESS** allows the regressors to be regarded as fixed or random depending on the specifications given. As might be expected due to the increased randomness

of the design, confidence intervals tend to be wider when regressors are random compared to when regressors are fixed, holding all other factors constant.

CIs for standardized omnibus effects in multiple regression with **MBESS**

The function `ci.R2()` from **MBESS** can be used to form confidence intervals for the population squared multiple correlation coefficient (i.e., ρ^2). Although many optional specifications exist, when predictors were fixed (i.e., when the predictors are not random), a basic specification would be of the form

`ci.R2(R2=`R^2`, N=`N`, k=`k`, conf.level=`$1-\alpha$`, Random.Regressors=FALSE)`,

where `R2` is the observed squared multiple correlation coefficient, `N` is the sample size, `k` is the number of regressors, `conf.level` is the desired confidence interval coverage, and `Random.Regressors` defines whether the regressors are fixed or random (with a `TRUE` statement for random regressors and a `FALSE` statement for fixed regressors).

As is more typically the case in the BESS, regressors are random and as such the default for `Random.Regressors` is `TRUE` and need not be specified. Thus, the following would provide a confidence interval for the squared multiple correlation coefficient when regressors are random

`ci.R2(R2=`R^2`, N=`N`, k=`k`, conf.level=`$1-\alpha$`)`.

Although one could easily take the square root of the confidence limits for ρ^2 obtained from the `ci.R2()` function as the confidence limits for ρ, due to the confidence interval transformation principle, the function `ci.R()` provides such a confidence interval directly. The way in which `ci.R()` is used is nearly identical to `ci.R2()`, with the only difference being that R is specified instead of R^2:

`ci.R(R=`R`, N=`N`, k=`k`, conf.level=`$1-\alpha$`)`.

4.2. Standardized effect sizes for targeted effects in multiple regression

Standardized regression coefficients are often used in the context of multiple regression in the BESS. Of course, basing a multiple regression on variables that have been converted to z-scores does not affect the overall fit of the model (i.e., R^2 is left unchanged), but it can facilitate interpretation when predictor variables are on different scales and/or when the measurement scales are themselves arbitrary.

The process of standardizing a regression coefficient can proceed in two ways: using scores that have been converted to z-scores *or* by multiplying the unstandardized regression coefficient by the ratio of the standard deviation of X_k to the standard deviation of Y. Thus, in the population the standardized regression coefficient is given as

$$\beta_k = \mathcal{B}_k \frac{\sigma_{X_k}}{\sigma_Y} \tag{78}$$

for regressor k. As before, when an unstandardized estimate is divided or multiplied by a random quantity the test statistic can no longer be pivoted. The test statistic for the null hypothesis that $\beta_k = 0$ is given as

$$t = \frac{b_k}{s_{b_k}}, \tag{79}$$

where s_{b_k} is the standard error of b_k defined as

$$s_{b_k} = \sqrt{\frac{1 - R^2_{Y \cdot \mathbf{X}}}{(1 - R^2_{\mathrm{X}_k \cdot \mathbf{X}_{-k}})(N - K - 1)}}, \tag{80}$$

with $R^2_{\mathrm{X}_k \cdot \mathbf{X}_{-k}}$ being the squared multiple correlation coefficient when predictor X_k is the dependent variable predicted from the remaining $K - 1$ regressors. Alternatively $R^2_{\mathrm{X}_k \cdot \mathbf{X}_{-k}}$ can be obtained indirectly from the covariance matrix of the predictors, $\mathbf{S}_{\mathrm{XX}}$, as

$$R^2_{\mathrm{X}_k \cdot \mathbf{X}_{-k}} = 1 - (s_k^2 c_{kk})^{-1} \tag{81}$$

where c_{kk} is the diagonal element of $\mathbf{S}^{-1}_{\mathrm{XX}}$ (Harris 2001).

Since the noncentrality parameter of a t-distribution is the value of the t-test statistic if population values were substituted for the sample quantities, the noncentrality parameter for a standardized regression coefficient can be written as

$$\lambda_k = \phi_k \sqrt{N} \tag{82}$$

where

$$\phi_k = \beta_k \sqrt{\frac{1 - \mathrm{P}^2_{\mathrm{X}_k \cdot \mathbf{X}_{-k}}}{1 - \mathrm{P}^2_{Y \cdot \mathbf{X}}}}. \tag{83}$$

Because β_k can be written (e.g., Hays 1994) as

$$\beta_k = \sqrt{\frac{\mathrm{P}^2_{Y \cdot \mathbf{X}} - \mathrm{P}^2_{Y \cdot \mathbf{X}_{-k}}}{1 - \mathrm{P}^2_{\mathrm{X}_k \cdot \mathbf{X}_{-k}}}} \tag{84}$$

ϕ_k can be written as

$$\phi_k = \sqrt{\frac{\mathrm{P}^2_{Y \cdot \mathbf{X}} - \mathrm{P}^2_{Y \cdot \mathbf{X}_{-k}}}{1 - \mathrm{P}^2_{Y \cdot \mathbf{X}}}}. \tag{85}$$

Notice that ϕ_k is the square root of the signal-to-noise ratio for a targeted effect, which shows the contribution of the kth effect to the overall signal-to-noise ratio, ϕ^2.

Given the representation of λ_k in Equation 82, β_k can be solved for such that

$$\beta_k = \frac{\lambda_k}{\sqrt{N}} \sqrt{\frac{1 - \mathrm{P}^2_{Y \cdot \mathbf{X}}}{1 - \mathrm{P}^2_{\mathrm{X}_k \cdot \mathbf{X}_{-k}}}}. \tag{86}$$

Thus, forming a confidence interval for β_k involves transforming the limits of the confidence interval for λ, by way of Equation 86:

$$p\left[\frac{\lambda_{\mathrm{L}}}{\sqrt{N}} \sqrt{\frac{1 - \mathrm{P}^2_{Y \cdot \mathbf{X}}}{1 - \mathrm{P}^2_{\mathrm{X}_k \cdot \mathbf{X}_{-k}}}} \leq \beta_k \leq \frac{\lambda_{\mathrm{U}}}{\sqrt{N}} \sqrt{\frac{1 - \mathrm{P}^2_{Y \cdot \mathbf{X}}}{1 - \mathrm{P}^2_{\mathrm{X}_k \cdot \mathbf{X}_{-k}}}}\right] = 1 - \alpha. \tag{87}$$

CIs for targeted effects in multiple regression with **MBESS**

The function `ci.src()` from **MBESS** can be used to form confidence intervals for the population standardized regression coefficient (i.e., β_k). Although many optional specifications exist, a basic specification would be of the form

`ci.src(beta.k=`b_k`, SE.beta.k=`s_{b_k}`, N=`N`, k=`k`, conf.level=`$1-\alpha$`)`,

where `beta.k` is the observed standardized regression coefficient, `SE.beta.k` is the observed standard error of the standardized regression coefficient, `N` is the sample size, and `k` is the number of regressor variables.[3]

5. General CI procedures for standardized effects

Although other standardized effects that are beneficial in the BESS and could have been discussed, the ideas presented generalize to a wide variety of standardized effects (e.g., the coefficient of variation, the Mahalanobis distance, the change in the squared multiple correlation coefficient when one or more predictors is added (or removed) from a model, the root mean square error of approximation (in the context of a structural equation model), etc.). In accord with the confidence interval inversion and transformation principles, exact confidence intervals for many such effects can be formed. The specific R functions from **MBESS** used to form confidence intervals with `conf.limits.nct()` and `conf.limits.ncf()`. These functions provide confidence limits for noncentral t and F parameters, respectively. The confidence interval functions discussed in the article thus used these general functions and applied the necessary transformation of the confidence limits so that the scale of the confidence interval was in terms of the particular standardized effect size. Relatedly, the function `conf.limits.nc.chisq()` is also part of **MBESS**, where the function is used to form confidence intervals for noncentral χ^2 parameters. The ability to form confidence intervals for noncentrality parameters in **MBESS** is one of its most important features.

6. Discussion

The topic of confidence intervals for the general case of standardized effect sizes has not often been considered in the general statistics literature, where parameters of interest are almost always in an unstandardized form. Confidence intervals for standardized effects, have, however, recently generated much interest in the BESS (e.g., Algina *et al.* 2006; Cumming and Finch 2001; Kelley 2005; Kelley and Rausch 2006; Steiger and Fouladi 1997; Smithson 2001, 2003; Steiger 2004), where confidence intervals for effect sizes of primary importance are very strongly encouraged (e.g., Wilkinson and The American Psychological Association Task Force on Statistical Inference 1999; Task Force on Reporting of Research Methods in AERA Publications 2006). However, software to implement confidence intervals for standardized effect sizes has, for the most part, been in the form of specialized scripts for specific effect sizes or stand alone software packages. **MBESS**, however, is a package within the R statistical language and environment and thus can be seamlessly incorporated into data analysis in R. Furthermore, the functions contained within **MBESS** are designed to be user friendly and accept the necessary sufficient statistics as input. Thus, for those who prefer other data analysis software programs, the **MBESS** functions can be used within R with only the summary statistics provided from other software programs (as was shown in the examples). One need

[3]The argument `beta.k` and `SE.beta.k` should not be confused with the population standardized regression coefficient and its standard error, respectively. In the BESS "beta weight" is often used to refer to the standardized regression coefficient, and use of `beta.k` and `SE.beta.k` in the `ci.src()` function helps to avoid confusion with the unstandardized regression coefficient, `b.j` and `SE.b.j` from the function `ci.rc()` for forming confidence intervals for unstandardized regression coefficients.

not be a skilled R user in order to use R for confidence interval formation with **MBESS**, as the necessary R commands are very simple and rely only on summary statistics that are easily obtainable in R or from other programs. It is hoped that this article and the Methods for the Behavioral, Educational, and Social Sciences (**MBESS**; Kelley 2007b,a) R package will be helpful resources as the future of quantitative research (Thompson 2002) unfolds in the behavioral, educational, and social sciences, where inferences based on dichotomous outcomes from null hypothesis significance tests (reject or fail-to-reject) are replaced by effect sizes and their corresponding confidence intervals.

Acknowledgments

This work was supported in part by a Proffitt Fellowship for educational research.

References

Algina J, Keselman HJ, Penfield RD (2006). "Confidence Interval Coverage for Cohen's Effect Size Statistic." *Educational and Psychological Measurement*, **66**, 945–960.

Algina J, Olejnik S (2000). "Determining Sample Size for Accurate Estimation of the Squared Multiple Correlation Coefficient." *Multivariate Behavioral Research*, **35**, 119–136.

Casella G, Berger RL (2002). *Statistical Inference*. Duxbury Press, Pacific Grove, CA, 2nd edition.

Cohen J (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition.

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

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.

Ding CG (1996). "On the Computation of the Distribution of the Square of the Sample Multiple Correlation coefficient." *Computational Statistics & Data Analysis*, **22**, 345–350.

Fisher RA (1928). "The General Sampling Distribution of the Multiple Correlation Coefficient." *Proceedings of the Royal Society, Series A*, **121**, 654–673.

Fleishman AI (1980). "Confidence Intervals for Correlation Ratios." *Educational and Psychological Measurement*, **40**, 659–670.

Gatsonis C, Sampson AR (1989). "Multiple Correlation: Exact Power and Sample Size Calculations." *Psychological Bulletin*, **106**, 516–524.

Glass GV, McGaw B, Smith ML (1981). *Meta-analysis in social research*. Sage, Beverly Hills, CA.

Harris RJ (2001). *A Primer of Multivariate Statistics*. Lawrence Erlbaum Associates, Mahwah, NJ, 3rd edition.

Hays WL (1994). *Statistics*. Harcourt Brace College Publishers, New York, NY, 5th edition.

Hedges L, Olkin I (1985). *Statistical Methods for Meta-Analysis*. Academic Press, Orlando, FL.

Hunter JE, Schmidt FL (2004). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Sage, Newbury Park, CA.

Johnson NL, Kotz S, Balakrishnan N (1995). *Continuous Univariate Distributions*, volume 2. Wiley, New York, NY, 2nd edition.

Johnson NL, Welch BL (1940). "Applications of the Noncentral t-Distribution." *Biometrika*, **31**, 362–389.

Kelley K (2005). "The Effects of Nonnormal Distributions on Confidence Intervals Around the Standardized Mean Difference: Bootstrapping as an Alternative to Parametric Confidence Intervals." *Educational and Psychological Measurement*, **65**, 51–69.

Kelley K (2007a). "Methods for the Behavioral, Educational, and Educational Sciences: An R Package." *Behavior Research Methods*. In press.

Kelley K (2007b). **MBESS***: Methods for the Behavioral, Educational, and Social Sciences*. R package version 0.0.8, URL http://CRAN.R-project.org/.

Kelley K, Rausch JR (2006). "Sample Size Planning for the Standardized Mean Difference: Accuracy in Parameter Estimation via Narrow Confidence Intervals." *Psychological Methods*, **11**, 363–385.

Kempthorne O, Folks L (1971). *Probability, Statistics, and Data Analysis*. Iowa State University Press, Ames, IA.

Kline RB (2004). *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*. American Psychological Association, Washington, DC.

Krantz DH (1999). "The Null Hypothesis Testing Controversy in Psychology." *Journal of the American Statistical Association*, **94**, 1372–1381.

Lee YS (1971). "Tables of the Upper Percentage Points of the Multiple Correlation." *Biometrika*, **59**, 175–189.

Maxwell SE, Delaney HD (2004). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Erlbaum, Mahwah, NJ, 2nd edition.

Meehl PE (1997). "The Problem is Epistemology, Not Statistics: Replace Significance Tests by Confidence Intervals and Auantify Accuracy of Risky Numerical Predictions." In LL Harlow, SA Mulaik, JH Steiger (eds.), "What if There Where no Significance Tests?", pp. 393–426. Lawrence Erlbaum Associates, Mahwah, NJ.

Neyman J (1935). "On the Problem of Confidence Intervals." *The annals of mathematical statistics*, **6**, 111–116.

Neyman J (1937). "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability." *Philosophical Transaction of the Royal Society of London. Series A, Mathematical and Physical Sciences*, **236**, 333–380.

Patel JK, Read CB (1982). *Handbook of the Normal Distribution*. Marcel Dekker, Inc., New York, NY.

Powell BB (1996). *Homer and the Origin of the Greek Alphabet*. Cambridge University Press, Cambridge, MA, 2nd edition.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rencher AC (2000). *Linear Models in Statistics*. Wiley, New York, NY.

Sampson AR (1974). "A Tale of Two Regressions." *Journal of the American Statistical Association*, **69**, 682–689.

Schmidt FL (1996). "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers." *Psychological Methods*, **1**, 115–129.

Smithson M (2001). "Correct Confidence Intervals for Various Regression Effect Sizes and Parameters: The Importance of Noncentral Distributions in Computing Intervals." *Educational and Psychological Measurement*, **61**, 605–632.

Smithson M (2003). *Confidence Intervals*. Sage Publications, Thousand Oaks, CA.

Steiger JH (2004). "Beyond the F Test: Effect Size Confidence Intervals and Tests of Close Fit in the Analysis of Variance and Contrast Analysis." *Psychological Methods*, **9**, 164–182.

Steiger JH, Fouladi RT (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 JH, Fouladi RT (1997). "Noncentrality Interval Estimation and the Evaluation of Statistical Methods." In LL Harlow, SA Mulaik, JH Steiger (eds.), "What if There Where no Significance Tests?", pp. 221–257. Lawrence Erlbaum Associates, Mahwah, NJ.

Stuart A, Ord JK, Arnold S (1999). *Kendall's Advanced Theory of Statistics: Classical Inference and the Linear Model*, volume 2A. Oxford University Press, New York, NY, 6th edition.

Task Force on Reporting of Research Methods in AERA Publications (2006). *Standards for Reporting on Empirical Social Science Research in AERA Publications, American Educational*. American Educational Research Association, Washington, DC.

Thompson B (2002). "What Future Quantitative Social Science Research could Look Like: Confidence Intervals for Effect Sizes." *Educational Researcher*, **31**, 25–32.

Wilkinson L, The American Psychological Association Task Force on Statistical Inference (1999). "Statistical Methods in Psychology: Guidelines and Explanations." *American Psychologist*, **54**, 594–604.

Affiliation:

Ken Kelley
Indiana University
Inquiry Methodology Program
201 North Rose Avenue
Bloomington, IN 47405, United States of America
E-mail: KKIII@Indiana.Edu
URL: http://www.indiana.edu/~kenkel/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 20, Issue 8 *Submitted:* 2006-10-01
May 2007 *Accepted:* 2007-02-22