Psychological Methods
2008, Vol. 13, No. 3, 272–277

1082-989X/08/$12.00 DOI: 10.1037/a0013158

An Inferential Confidence Interval Method of Establishing Statistical Equivalence That Corrects Tryon's (2001) Reduction Factor

Warren W. Tryon and Charles Lewis
Fordham University

Evidence of group matching frequently takes the form of a nonsignificant test of statistical difference. Theoretical hypotheses of no difference are also tested in this way. These practices are flawed in that null hypothesis statistical testing provides evidence against the null hypothesis and failing to reject H_0 is not evidence supportive of it. Tests of statistical equivalence are needed. This article corrects the inferential confidence interval (ICI) reduction factor introduced by W. W. Tryon (2001) and uses it to extend his discussion of statistical equivalence. This method is shown to be algebraically equivalent with D. J. Schuirmann's (1987) use of 2 one-sided *t* tests, a highly regarded and accepted method of testing for statistical equivalence. The ICI method provides an intuitive graphic method for inferring statistical difference as well as equivalence. Trivial difference occurs when a test of difference and a test of equivalence are both passed. Statistical indeterminacy results when both tests are failed. Hybrid confidence intervals are introduced that impose ICI limits on standard confidence intervals. These intervals are recommended as replacements for error bars because they facilitate inferences.

Keywords: confidence intervals, statistical difference, equivalence, indeterminacy, error bars

Investigators are interested in establishing statistical equivalence in at least two broad cases. The first case is methodological and concerns group matching where investigators wish to demonstrate that two groups are equivalent with regard to a control variable. Group matching is often more practical than individual matching. Investigators typically present a nonsignificant null hypothesis statistical test, often a *t* test for independent groups, as evidence that group matching has been successful. The second case is theoretical and concerns hypotheses of no difference. Investigators sometimes hypothesize that two treatments, psychological and/or pharmacological, are equally effective or that a new treatment is as effective as a well-established treatment for the same condition. Other times, theoretical implications indicate that two groups should test the same or perform equivalently. Blackwelder (1982) correctly concluded that "*p* is a measure of evidence against the null hypothesis, not for it, and insufficient evidence to reject the null hypothesis does not imply sufficient evidence to accept it" (p. 346). An equivalence test is needed. Anderson and Hauck (1983); Blackwelder (1982); Detsky and Sackett (1985); Dunnett and Gent, (1977); Frick (1995); Makuch and Simon (1978); Metzler (1979); Rogers, Howard, and Vessey (1993); Schuirmann (1987); Selwyn, Dempster, and Hall (1981); Selwyn and Hall (1984); and Westlake (1972, 1976, 1979, 1981) are some of the authors who have discussed statistical equivalence.

In this article, we have four objectives: first, to correct Tryon's (2001) formula for a shrinkage factor, hereafter referred to as a *reduction factor*, for assessing statistical difference and equivalence; second, to report that the inferential confidence interval (ICI) method for assessing statistical equivalence is algebraically equivalent to Schuirmann's (1987) method;[1] third, to show that the ICI method enables two additional inferences of trivial difference and statistical indeterminacy; fourth, to introduce hybrid confidence intervals as replacements for error bars.

Equation 1 presents the standard formula for a confidence interval that we call the *descriptive confidence interval*[2] (DCI) for the population mean because it was designed to

Warren W. Tryon and Charles Lewis, Department of Psychology, Fordham University.

Correspondence concerning this article should be addressed to Warren W. Tryon, Department of Psychology, Fordham University, Bronx, NY 10458-9998. E-mail: wtryon@fordham.edu

[1] Rogers, Howard, and Vessey (1993) described essentially the same method of testing for statistical equivalence.

[2] Although all confidence intervals are inferential, the terms *descriptive* and *inferential* have been chosen to distinguish an interest in parameter and/or sample coverage from an interest in determining if a *t* test for independent groups would be statistically significant if it was conducted.

capture the true value of the parameter a stated proportion of the time. It is computed using the sample mean plus and minus $t_v^{\alpha/2}$ (the upper 100 · α/2 percentile of the t distribution with v degrees of freedom) estimated standard errors ($df = v = n - 1$) of the sample mean, where the estimated standard error of the sample mean equals the sample standard deviation (S) divided by the square root of the sample size (n), as follows:

$$\bar{Y} \pm t_v^{\alpha/2} S_{\bar{Y}} = \bar{Y} \pm t_v^{\alpha/2} \frac{S}{\sqrt{n}}. \quad (1)$$

Basing inferences of statistical difference between two means on DCIs is complicated by the fact that the arms of two DCIs can overlap by 58% and yet a standard t test can yield evidence of statistical difference, for example, $p = .047$ (Cumming & Finch, 2005, Figure 4). Schenker and Gentleman (2001) and Belia, Fidler, Williams, and Cumming (2005) have shown that investigators' natural tendency is to infer difference when two confidence intervals do not overlap and not otherwise. These studies document that even seasoned investigators cannot always adjust their decisions to compensate for confidence interval overlap. Cumming and Finch (2005) have attempted to educate investigators by providing rules to guide correct inference. Tryon (2001) took a human factors approach to this problem, capitalized on people's intuition, and created reduced ICIs for a pair of means that abut when the two groups just differ by a t test that is significant at $p = .05$, as illustrated in Figure 1.

The difference between the two means illustrated in Figure 1 equals the sum of the half widths of the two reduced ICIs. The objective illustrated in Figure 1 was to determine the critical t value (t_x) that made the two means just significantly different at the stated significance level. Tryon's (2001) Equation 2 began with the equation for the independent groups t test, substituted the confidence interval expression for the difference between the two means depicted in Figure 1 (assuming $\bar{Y}_2 > \bar{Y}_1$), and factored out the t value (t_x) needed to ensure that two nonoverlaping confidence intervals would be statistically significant at the stated probability level as follows:

$$t_v^{\alpha/2} = \frac{\bar{Y}_2 - \bar{Y}_1}{S_{\bar{Y}_2 - \bar{Y}_1}} = \frac{t_x S_{\bar{Y}_1} + t_x S_{\bar{Y}_2}}{S_{\bar{Y}_2 - \bar{Y}_1}} = \frac{t_x (S_{\bar{Y}_1} + S_{\bar{Y}_2})}{S_{\bar{Y}_2 - \bar{Y}_1}} = t_x \frac{S_{\bar{Y}_1} + S_{\bar{Y}_2}}{S_{\bar{Y}_2 - \bar{Y}_1}}. \quad (2)$$

Tryon (2001) presented Equation 3, which solved for t_x and defined E as the ratio of the standard error of the difference between two groups to the sum of the standard errors of both groups. The symbol E (for experimental design) was used to draw attention to the fact that the tabled value of the t distribution is being reduced by a composite factor that is a function of the experimental design as follows:

$$t_x = t_v^{\alpha/2} \frac{S_{\bar{Y}_2 - \bar{Y}_1}}{S_{\bar{Y}_1} + S_{\bar{Y}_2}} = t_v^{\alpha/2} E. \quad (3)$$

The flaw in this derivation is that factoring t_x in Equation 2 requires that a single t value for t_x be used, yet Tryon's (2001) numerical examples made use of different t values when sample sizes were unequal. This flaw can be corrected by recognizing that three t values are relevant: the t value for each of the two groups, $t_{v_1}^{\alpha/2}$ and $t_{v_2}^{\alpha/2}$, respectively, that are the same when $n_1 = n_2$, and the t value for the difference between the two means, $t_{v_{12}}^{\alpha/2}$, that depends on $df = v_{12} = v_1 + v_2 = (n_1 - 1) + (n_2 - 1)$ when the two population variances are assumed to be equal and on Satterthwaite's (1946) correction when they are not assumed equal (see Equation 4). Specifically,

$$df = \frac{(S_{\bar{Y}_1}^2 + S_{\bar{Y}_2}^2)^2}{\frac{(S_{\bar{Y}_1}^2)^2}{n_1 - 1} + \frac{(S_{\bar{Y}_2}^2)^2}{n_2 - 1}} = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}. \quad (4)$$

To correct the derivation, first assume that Group 2 has a higher sample mean than Group 1. The descriptive confidence interval for Group 1 ranges from $\bar{Y}_1 - t_{v_1}^{\alpha/2} S_{\bar{Y}_1}$ to $\bar{Y}_1 + t_{v_1}^{\alpha/2} S_{\bar{Y}_1}$, and the descriptive confidence interval for Group 2 ranges from $\bar{Y}_2 - t_{v_2}^{\alpha/2} S_{\bar{Y}_2}$ to $\bar{Y}_2 + t_{v_2}^{\alpha/2} S_{\bar{Y}_2}$. The corresponding inferential confidence interval for Group 1 ranges

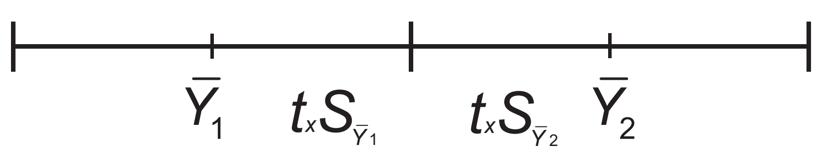

Figure 1. Inferential confidence intervals (ICIs) were designed to facilitate comparing means. This figure illustrates the computation of two such intervals where the requisite number (t_x) of standard errors are added to and subtracted from each of two means resulting in ICIs that abut when a t test is just statistically significant.

from $\bar{Y}_1 - Et_{v_1}^{\alpha/2}S_{\bar{Y}_1}$ to $\bar{Y}_1 + Et_{v_1}^{\alpha/2}S_{\bar{Y}_1}$, and the inferential confidence interval for Group 2 ranges from $\bar{Y}_2 - Et_{v_2}^{\alpha/2}S_{\bar{Y}_2}$ to $\bar{Y}_2 + Et_{v_2}^{\alpha/2}S_{\bar{Y}_2}$ where E is the appropriate reduction factor. The desired result is for the t test for two independent groups to be significant if and only if the difference between the lower limit of the ICI for Group 2 minus the upper limit of the ICI for Group 1 is greater than or equal to zero. Equation 5 expresses the identity that results when the t test is just significant, whereas Equation 6 uses E as the reduction factor by which each descriptive confidence interval must be reduced so that the two ICIs abut when the t test is just significant. Specifically,

$$t_{v_{12}}^{\alpha/2} = \frac{\bar{Y}_2 - \bar{Y}_1}{S_{\bar{Y}_2-\bar{Y}_1}} \tag{5}$$

if and only if

$$(\bar{Y}_2 - Et_{v_2}^{\alpha/2}S_{\bar{Y}_2}) - (\bar{Y}_1 + Et_{v_1}^{\alpha/2}S_{\bar{Y}_1}) = 0. \tag{6}$$

Solving Equation 6 for the difference between the two means gives Equation 7:

$$\bar{Y}_2 - \bar{Y}_1 = E(t_{v_1}^{\alpha/2}S_{\bar{Y}_1} + t_{v_2}^{\alpha/2}S_{\bar{Y}_2}). \tag{7}$$

Substituting the right-hand portion of this equation for the difference between the two means in the numerator of the right-hand portion of Equation 5 yields Equation 8,

$$t_{v_{12}}^{\alpha/2} = \frac{E(t_{v_1}^{\alpha/2}S_{\bar{Y}_1} + t_{v_2}^{\alpha/2}S_{\bar{Y}_2})}{S_{\bar{Y}_2-\bar{Y}_1}}, \tag{8}$$

where the denominator may be based on pooled or unpooled within-group variance estimates and $t_{v_{12}}^{\alpha/2}$ is modified accordingly. Solving Equation 8 for E yields Equation 9:

$$E = \frac{t_{v_{12}}^{\alpha/2}S_{\bar{Y}_2-\bar{Y}_1}}{t_{v_1}^{\alpha/2}S_{\bar{Y}_1} + t_{v_2}^{\alpha/2}S_{\bar{Y}_2}}. \tag{9}$$

This corresponds to Tryon's (2001) formula for E times the ratio $t_{v_{12}}^{\alpha/2}/t_{v_1}^{\alpha/2}$ when $n_1 = n_2$. This modification becomes trivial for $n > 30$ but is increasingly important as sample sizes become smaller. No simple relationship between the current result and the earlier formula exists when $n_1 \neq n_2$.

The above derivation has been carried out for the case of two independent samples. It is worth noting that the extension to the case of dependent samples is straightforward. The estimated standard error of the difference between sample means for two independent samples is given as Equation 10:

$$S_{\bar{Y}_2-\bar{Y}_1} = \sqrt{S_{\bar{Y}_1}^2 + S_{\bar{Y}_1}^2}. \tag{10}$$

The estimated standard error for the difference between sample means for dependent samples is given as Equation 11 with degrees of freedom equal to $n - 1$,

$$S_{\bar{Y}_2-\bar{Y}_1} = \sqrt{S_{\bar{Y}_1}^2 + S_{\bar{Y}_2}^2 - 2r_{12}S_{\bar{Y}_1}S_{\bar{Y}_2}}, \tag{11}$$

where r_{12} denotes the correlation between responses in the two dependent samples. Note that $S_{\bar{Y}_2-\bar{Y}_1}$ and therefore E is smaller to the degree that the responses in the two samples are more positively correlated. Also note that because $v_{12} = v_1 = v_2 = n - 1$ in the dependent samples case, the expression for E in Equation 9 simplifies to the one given by Tryon (2001).

Statistical Equivalence

Figure 2 defines two values of what we call the *equivalence range* (*eRg*)[3] in terms of two hybrid confidence intervals with inner ICI limits and outer DCI limits: $\bar{Y}_1$ is bounded by upper and lower ICI limits of U_1 and L_1 and $\bar{Y}_2$ is bounded by upper and lower ICI limits of U_2 and L_2 where $\bar{Y}_2 > \bar{Y}_1$. Statistical equivalence occurs when $eRg \leq \Delta$, an amount that is considered to be inconsequential on substantive grounds that have been established apart from the analysis at hand by professional consensus or other means. The value of *eRg* is always the larger of eRg_1 and eRg_2, which in Figure 2 is eRg_1. Note that if $\bar{Y}_2 > \bar{Y}_1$, then $eRg_1 > eRg_2$, so statistical equivalence occurs if and only if

$$(\bar{Y}_2 - \bar{Y}_1) + E(t_{v_2}^{\alpha/2}S_{\bar{Y}_2} + t_{v_1}^{\alpha/2}S_{\bar{Y}_1}) \leq \Delta. \tag{12}$$

Schuirmann (1987) proposed a double t test method of certifying statistical equivalence based on the following two null hypotheses:

1. H_{01}: $\mu_2 - \mu_1 \geq \Delta$. H_{01} is rejected if $t_1 \leq -t_{v_{12}}^{\alpha}$ where

$$t_1 = \frac{(\bar{Y}_2 - \bar{Y}_1) - \Delta}{S_{\bar{Y}_2-\bar{Y}_1}}. \tag{13}$$

2. H_{02}: $\mu_2 - \mu_1 \leq -\Delta$. H_{02} is rejected if $t_2 \geq t_{v_{12}}^{\alpha}$ where

$$t_2 = \frac{(\bar{Y}_2 - \bar{Y}_1) - (-\Delta)}{S_{\bar{Y}_2-\bar{Y}_1}}. \tag{14}$$

The value of t_1 becomes more negative as Δ exceeds the positive difference between the two means. Rejecting H_{01} certifies that the difference between the means is smaller than Δ.

[3] We thank Tammy Trierweiler for suggesting the *eRg* symbol for equivalence range. The *g* in *eRg* is part of the abbreviation of the word *range*.

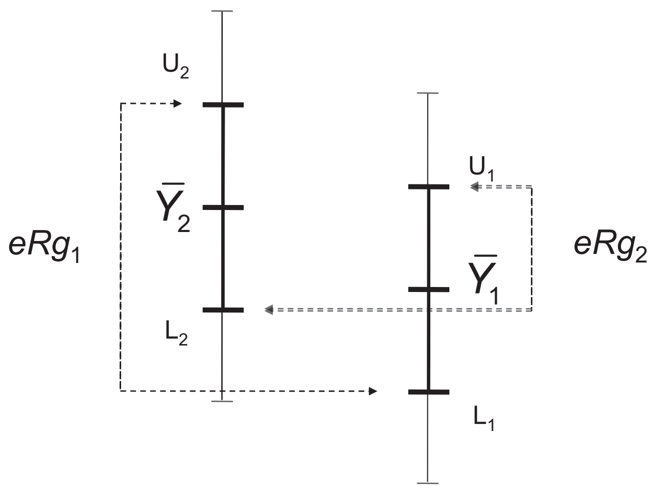

Figure 2. This figure graphically defines the equivalence range (*eRg*) as the larger of two differences between the endpoints of two ICIs designated by the heavy lines. Delta is a value that, on substantive grounds, is agreed to be an inconsequential amount. Statistical equivalence occurs when $eRg \leq \Delta$. The light lines designate the upper and lower limits of standard 95% confidence intervals. The proportions in this figure reflect an example provided by Tryon (2001, p. 376).

The value of t_2 becomes more positive as Δ increases. Rejecting H_{02} certifies that the difference between the means is greater than $-\Delta$. A large positive difference between the means would also reject H_{02} and a large negative difference would also reject H_{01}. But both null hypotheses must be rejected to establish statistical equivalence. Rejecting both null hypotheses establishes statistical equivalence within the interval $[-\Delta, \Delta]$.

Schuirmann (1987) tested each of these null hypotheses at the $100(1 - \alpha)$ percentile value of $t^{\alpha}_{v_{12}}$ rather than the $100(1 - \alpha/2)$ percentile. This turns out to control the overall Type I error rate (defined as the probability of claiming equivalence when the absolute difference between the population means exceeds Δ) at α that occurs when either of the null hypotheses being tested is true. If each tail of an ordinary two-tailed test is tested at level α, the resulting combined test has level 2α. Schuirmann's equivalence test performs two one-tailed tests, each at level α, but the resulting combined test still has level α rather than 2α. This is surprising until one realizes that, unlike the ordinary two-tailed test, the null hypotheses tested with Schuirmann's test have no overlap. For the ordinary test, when $\mu_2 - \mu_1 = 0$, a Type I error can be made in either tail, so the total Type I error probability must be the sum of the two tail probabilities. For Schuirmann's test, if in the population $\mu_2 - \mu_1 > \Delta$, then a Type I error can only be made with the test of H_{01}: $\mu_2 - \mu_1 \geq \Delta$ because in this case H_{02}: $\mu_2 - \mu_1 \leq -\Delta$ must be false. Similarly, if in the population $\mu_2 - \mu_1 < -\Delta$, then a Type I error can only be made with the test of H_{02}: $\mu_2 - \mu_1 \leq -\Delta$ because H_{01}: $\mu_2 - \mu_1 \geq \Delta$ must be false. Therefore, the total Type I error probability can never exceed the alpha level for each of the tests separately.

In practice, Schuirmann's two tests to establish statistical equivalence can be reduced to a single test. Suppose $\bar{Y}_2 > \bar{Y}_1$. Then, if one is to check for statistical equivalence, only H_{01}: $\mu_2 - \mu_1 \geq \Delta$ needs to be tested. If H_{01} is not rejected, then statistical equivalence does not apply and there is no need to test H_{02}. However, if H_{01} is rejected, then the test of H_{02} must also lead to a rejection. To see this, note that H_{01} is rejected if and only if

$$\frac{(\bar{Y}_2 - \bar{Y}_1) - \Delta}{S_{\bar{Y}_2 - \bar{Y}_1}} \leq -t^{\alpha}_{v_{12}}. \tag{15}$$

Equation 15 is equivalent to

$$(\bar{Y}_2 - \bar{Y}_1) + t^{\alpha}_{v_{12}} S_{\bar{Y}_2 - \bar{Y}_1} \leq \Delta. \tag{16}$$

Because $\bar{Y}_2 - \bar{Y}_1 > 0$,

$$-(\bar{Y}_2 - \bar{Y}_1) + t^{\alpha}_{v_{12}} S_{\bar{Y}_2 - \bar{Y}_1} < (\bar{Y}_2 - \bar{Y}_1) + t^{\alpha}_{v_{12}} S_{\bar{Y}_2 - \bar{Y}_1}, \tag{17}$$

so, from Equations 16 and 17,

$$-(\bar{Y}_2 - \bar{Y}_1) + t^{\alpha}_{v_{12}} S_{\bar{Y}_2 - \bar{Y}_1} \leq \Delta. \tag{18}$$

Rearranging terms in Equation 18 gives

$$\frac{(\bar{Y}_2 - \bar{Y}_1) + \Delta}{S_{\bar{Y}_2 - \bar{Y}_1}} \geq t^{\alpha}_{v_{12}}. \tag{19}$$

On the basis of Equation 14, this is just what is required to reject H_{02} and, thus, establish statistical equivalence. To summarize, if $\bar{Y}_2 > \bar{Y}_1$, Schuirmann's procedure implies that statistical equivalence is established if and only if the inequality in Equation 16 holds.

Next, we show that Equation 16 is equivalent to requiring that a $100(1 - 2\alpha)$ version of *eRg* (denoted by $eRg^{2\alpha}$) is less than or equal to Δ. The intention here is to use Schuirmann's method as a benchmark to validate the ICI approach. First, define a 2α version of the reduction factor E:

$$E^{2\alpha} = \frac{t^{\alpha}_{v_{12}} S_{\bar{Y}_2 - \bar{Y}_1}}{t^{\alpha}_{v_1} S_{\bar{Y}_1} + t^{\alpha}_{v_1} S_{\bar{Y}_2}}. \tag{20}$$

Rearrange the terms in Equation 20 to give

$$t^{\alpha}_{v_{12}} S_{\bar{Y}_2 - \bar{Y}_1} = E^{2\alpha}(t^{\alpha}_{v_1} S_{\bar{Y}_1} + t^{\alpha}_{v_2} S_{\bar{Y}_2}). \tag{21}$$

Assuming $\bar{Y}_2 > \bar{Y}_1$, use Equation 21 to rewrite Equation 16 as

$$(\bar{Y}_2 - \bar{Y}_1) + E^{2\alpha}(t^{\alpha}_{v_1} S_{\bar{Y}_1} + t^{\alpha}_{v_2} S_{\bar{Y}_2}) \leq \Delta. \tag{22}$$

On the basis of Equation 12, Equation 22 may be rewritten as

$$eRg^{2\alpha} \leq \Delta. \tag{23}$$

To summarize, statistical equivalence is established by Schuirmann's procedure if and only if it is established by a 100(1 − 2α) version of Tryon's (2001) ICI test for statistical equivalence. That is, both methods reach the same conclusion. This relationship between the two methods may also be stated as follows: The standard 100(1 − α) ICI procedure establishes statistical equivalence if and only if an α/2 version of Schuirmann's test does so. Algebraic equivalence confers no advantage or disadvantage but in this case demonstrates a connection between a new method and a well-established one. The primary advantage of the ICI method is that it provides a single context in which to evaluate statistical difference, equivalence, trivial difference, and indeterminacy.

Trivial Difference and Indeterminacy

An advantage of the ICI method is that it enables one to evaluate both statistical difference and equivalence within the same graphic or numeric framework. It introduces two additional inferences not generally discussed. Trivial difference occurs when both the test of statistical difference and the test of statistical equivalence are passed. This occurs when the ICIs do not overlap but are sufficiently close together so that $eRg \leq \Delta$. Statistical indeterminacy occurs when both the test of statistical difference and the test of statistical equivalence are failed. No conclusions can be drawn. More data are required.

Hybrid Intervals Replace Error Bars

Hybrid confidence intervals impose ICI limits on standard confidence intervals as illustrated by the heavy lines in Figure 2. It is recommended that investigators use these intervals instead of error bars to illustrate the variability of their data because the inner, more visually prominent ICI limits facilitate drawing inferences regarding statistical difference, equivalence, indeterminacy, and trivial difference, whereas the DCI extensions indicated by the lighter lines provide additional coverage information. We believe that these hybrid confidence intervals better serve the scientific community than do error bars or other methods.

Multiple Comparisons

The inner ICI bars depend on multiple comparisons in two ways. First, the values of t used in Equation 9 to calculate E depend on the sample sizes of the two groups being compared when there are fewer than 120 degrees of freedom per group. A different value of E is used when comparing a group of 20 with a group of 30 than when comparing a group of 20 with a group of 40. This variation disappears when all samples exceed 121 participants and t values essentially become z values. Second, the values of t used in Equation 9 to calculate E depend on the number of comparisons being made. The Bonferroni inequality can be used to control the Type I error rate. With k groups, there are $k(k - 1)/2$ possible pairwise comparisons. Equation 24 describes the alpha level that should be used for the t value in the numerator of Equation 9 when conducting two-tailed tests among k groups:

$$\frac{\alpha/2}{k(k-1)/2} = \frac{\alpha}{k(k-1)}. \tag{24}$$

The URL http://statpages.org/pdfs.html locates a calculator that yields the t value for any valid combination of alpha value and degrees of freedom. Alternatively, one can use the Excel function TINV(alpha,df) to calculate two-tailed t values or TINV(2*alpha,df) to calculate one-tailed t values. The inner ICI bars also depend on the standard deviations of the groups being compared. Good experimental design seeks to equate them or use proportionately larger sample sizes in groups with larger standard deviations. Alternatively, one can create ICI limits based on average E values as recommended by Goldstein and Healy (1995).

References

Anderson, S., & Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics: Theory and Methods, 12,* 2663–2692.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10,* 389–396.

Blackwelder, W. C. (1982). "Proving the null hypothesis" in clinical trials. *Controlled Clinical Trials, 3,* 345–353.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60,* 170–180.

Detsky, A. S., & Sackett, D. L. (1985). When was a "negative" clinical trial big enough? How many patients you needed depends on what you found. *Archives of Internal Medicine, 145,* 709–712.

Dunnett, C. W., & Gent, M. (1977). Significance testing to establish equivalence between treatments with special reference to data in the form of 2 × 2 tables. *Biometrics, 33,* 593–602.

Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition, 23,* 132–138.

Goldstein, H., & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, Series A: Statistics in Society, 158,* 175–177.

Makuch, R., & Simon, R. (1978). Sample size requirements for evaluating a conservative therapy. *Cancer Treatment Reports, 62,* 1037–1040.

Metzler, C. M. (1979). Bioavailability—A problem in equivalence. *Biometrics, 30,* 309–317.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significant tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113,* 553–565.

Satterthwaite, F. W. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2,* 110–114.

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician, 55,* 182–186.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15,* 657–680.

Selwyn, M. R., Dempster, A. P., & Hall, N. R. (1981). A Bayesian approach to bioequivalence for the 2 × 2 changeover design. *Biometrics, 37,* 11–21.

Selwyn, M. R., & Hall, N. R. (1984). On Bayesian methods for bioequivalence. *Biometrics, 40,* 1103–1108.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6,* 371–386.

Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Science, 61,* 1340–1341.

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics, 32,* 741–744.

Westlake, W. J. (1979). Statistical aspects of comparative bioavailability trials. *Biometrics, 35,* 273–280.

Westlake, W. J. (1981). Bioequivalence testing—A need to rethink (Reader reaction response). *Biometrics, 37,* 591–593.

Received June 7, 2006
Revision received June 11, 2008
Accepted June 11, 2008 ■