Confidence Interval Coverage for Cohen's Effect Size Statistic

James Algina

University of Florida

and

H. J. Keselman

University of Manitoba

Kelly (2005) made a number of important contributions to the literature pertaining to confidence intervals (CIs) for Cohen's (1988) effect size (ES) statistic. One important finding he noted was that a noncentral-t (NCT) based CI has inaccurate coverage when data are nonnormal. Further, he found, that when data are nonnormal, accurate coverage could be achieved by adopting bootstrap methods. Specifically, he found two methods to be effective: the percentile method and the bias-corrected and accelerated (BCA) bootstrap methods. Coverage for the BCA method was better than the percentile bootstrap method and accordingly Kelly recommended that researchers adopt the BCA CI for Cohen's ES statistic.

However, Kelly (a) explored a limited range of non-normality, (b) did not examine a complete comparison of how population values of ES and sample size affects coverage probability, and (c) in some cases, used sample sizes that would be quite large in educational and psychological research. Accordingly, our study was designed to generalize Kelly's results.

Theoretical Background

One of the most commonly reported ESs is Cohen's d:

$$d = \frac{\bar{Y}_2 - \bar{Y}_1}{S}$$

where $\bar{Y}_j$ is the mean for the *j*th level $j = 1,2$ and *S* is the square root of the pooled variance, which we refer to as the pooled standard deviation. The number of observations in a level is denoted by n_j $N = n_1 + n_2$. Cohen's *d* estimates

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

where μ_j is the population mean for the *j*th level and σ is the population standard deviation, assumed to be equal for both levels.

It is known (see, for example, Cumming & Finch, 2001 or Steiger & Fouladi, 1997) that when the sample data are drawn from normal distributions, the variances of the two populations are equal, and the scores are independently distributed, an exact CI for the population ES i.e., δ can be constructed by using the NCT distribution. Figure 1 presents a central and NCT distribution. The distribution of the right is an example of a NCT distribution and is the sampling distribution of the *t* statistic when δ is not equal to zero. It has two parameters. The first is the familiar degrees of freedom and is $N-2$ in our context. The second is the noncentrality parameter

$$\lambda = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\left(\frac{\mu_2 - \mu_1}{\sigma}\right) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\,\delta .$$

The noncentrality parameter controls the location of the NCT distribution. In fact, the mean of the NCT distribution is approximately equal to λ (Hedges, 1981), with the accuracy of the approximation improving as *N* increases. The central *t* distribution, the sampling distribution of the *t* statistic when $\delta = 0$, is the special case of the NCT distribution that occurs when δ, and therefore, λ are zero.

To find a 95% (for example) CI for δ, we first use the NCT distribution to find a 95% CI for λ. Then multiplying the limits of the interval for λ by $\sqrt{n_1 + n_2 \,/ n_1 n_2}$ a 95% CI for δ is obtained. The lower limit of the 95% CI for λ is the noncentrality parameter for the NCT distribution in which the calculated *t* statistic

$$t = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\left(\frac{\bar{Y}_2 - \bar{Y}_1}{S}\right)$$

is the .975 quantile. The upper limit of the 95% interval for λ is the noncentrality parameter for the NCT distribution in which the calculated *t* statistic is the .025 quantile of the distribution (see Steiger & Fouladi, 1997). Means and standard deviations for an example are provided in Table 1. Calculations show that $d = .97$ and $t = 3.14$. The *t* statistic, along with two NCT distributions, is depicted in Figure 2. As Figure 2, indicates, if $\lambda = 5.21$, then $t = 3.14$ is the .025 percentile. Therefore the upper limit of the CI for λ is 5.21. If $\lambda = 1.05$ then $t = 3.14$ is the .975 percentile and the lower limit of the CI for λ is 1.05. Multiplying both limits by $\sqrt{n_1 + n_2 \,/ n_1 n_2}$, .32 and 1.61 are obtained as the lower and upper limits, respectively, for a 95% CI for δ.

Methods

We investigated the robustness of the NCT distribution-based CIs for δ and to sampling from nonnormal distributions. Probability coverage was estimated for all combinations of the following three factors: population distribution (four cases from the family of g and h distributions), sample size: $n_1 = n_2 = 20$ to 100 in steps of 20, and population ESs δ : 0 and .2 to 1.4 in steps of .3. The nominal confidence level for all intervals was .95 and each condition was replicated 5000 times.

The data were generated from the g and h distribution (Hoaglin, 1985). Specifically, we chose to investigate four g and h distributions: (a) $g = h = 0$, the standard normal distribution $\gamma_1 = \gamma_2 = 0$, (b) $g = .76$ and $h = -.098$, a distribution

with skew and kurtosis equal to that for an exponential distribution $\gamma_1 = 2,\ \gamma_2 = 6$, (c) $g = 0$ and $h = .225$ $\gamma_1 = 0$ and $\gamma_2 = 154.84$, and (d) $g = .225$ and $h = .225$ ($\gamma_1 = 4.90$ and $\gamma_2 = 4673.80$). The coefficient γ_1 is a measure of skew and γ_2 is a measure of kurtosis. As indicated in the description of the first distribution, a normal distribution has $\gamma_1 = \ \gamma_2 = 0$. Distributions with positive skew typically have $\gamma_1 > 0$ and distributions with negative skew typically have $\gamma_1 < 0$. Short-tailed distributions, such as a uniform distribution, typically have $\gamma_2 < 0$ and long-tailed distributions, such as a *t* distribution, typically have $\gamma_2 > 0.$ The three nonnormal distributions are quite strongly nonnormal. We selected these because we wanted to find whether the CIs would work well over a wide range of distributions, not merely with distributions that are nearly normal.

To generate data from a g and h distribution, standard unit normal variables Z_{ij} were converted to g and h distributed random variables via

$$Y_{ij} = \frac{\exp\ gZ_{ij}\ - 1}{g} \exp\left(\frac{hZ_{ij}^2}{2}\right)$$

when both g and h were non-zero. When g was zero

$$Y_{ij} = Z_{ij} \exp\left(\frac{hZ_{ij}^2}{2}\right).$$

The Z_{ij} scores were generated by using RANNOR in SAS (SAS, 1999). For simulated participants in treatment 2, the Y_{i2} scores were transformed to

$$Y_{i2} + \sigma \times \delta .$$

These transformed scores were used in the CI for δ .

References

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC.

Bird, K. D. (2002). Confidence intervals for effect sizes in analysis of variance. *Educational and Psychological Measurement*, *62*, 197-226.

Cohen, J. (1965). Some statistical issues in psychological research. In B.B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95-121). New York: Academic Press.

Cumming G., & Finch S. 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.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (1986). *Robust statistics*. New York: Wiley .

Hays, W. L. (1963). *Statistics*. New York: Holt, Rinehart and Winston.

Hedges, L. V. (1981) Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107-128.

Hoaglin, D. C. (1983). Summarizing shape numerically: The g-and h distributions. In D. C. Hoaglin, F. Mosteller, & Tukey, J. W. (Eds.), *Data analysis for tables, trends, and shapes: Robust and exploratory techniques*. New York: Wiley.

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology*, *82*, 3-5.

SAS Institute Inc. (1999). *SAS/IML user's guide, version 8*. Cary, NC: Author.

Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. New York: Wiley.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. Harlow, S. Mulaik, & J. H. Steiger (eds.), *What if there were no significance tests?* Hillsdale, NJ: Erlbaum.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, *54*, 837-847.

Wilkinson, L. and the Task force on Statistical Inference (1999). Statistical methods in psychology journals. *American Psychologist*, *54*, 594-604.

Wilcox, R. R. (2001). *Fundamentals of modern statistical methods*. New York:Springer.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego: Academic Press.

Wilcox, R. R., & Keselman, H. J. (in press). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*.

Yuen, K. K., & Dixon, W. J. (1973). The approximate behaviour and performance of the two-sample trimmed *t*. *Biometrika*, *60*, 369-374.

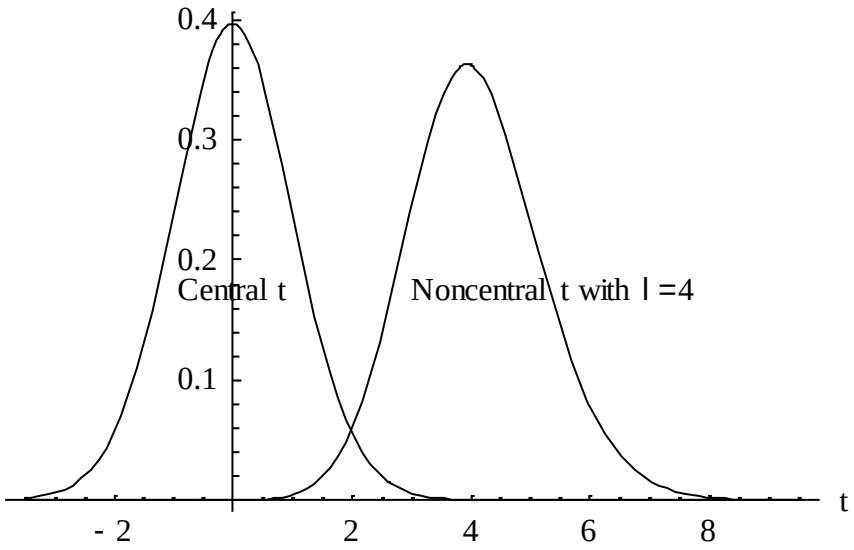

Figure 1. A central and a noncentral *t* distribution.

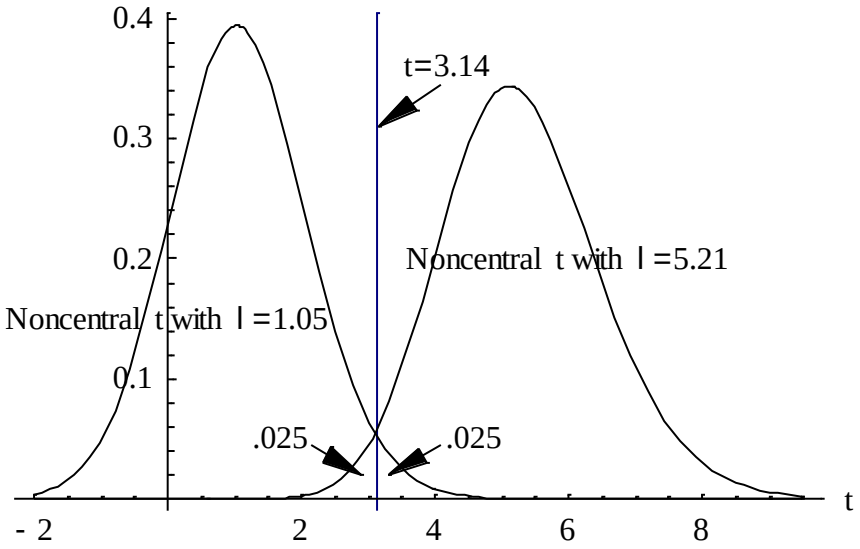

Figure 2. Graphical representation of finding a confidence interval for the noncentrality parameter

Our conclusion would be that with non-normal data BCA on delta may not work with sample sizes typical of those in the educational and psychological research [25 to 75 per group--If we need to go higher (100 per group I can do that)] and that researchers need to be very cautious in applying percentile or BCA to delta. If we want to we can conclude by pointing out that in the context of a repeated measures design Algina and Keselman (EPM, in press) introduced delta sub R and showed that percentile bootstrap intervals for delta sub R worked quite well and that Algina and Keselman (under review) have shown the same results in the context of an independent samples design. Taken together these results suggest replacing delta by delta sub R and using the percentile bootstrap interval.