11-1-2009

New Effect Size Rules of Thumb

Shlomo S. Sawilowsky
Wayne State University, shlomo@wayne.edu

Journal of Modern Applied Statistical Methods
November 2009, Vol. 8, No. 2, 597-599

1538 – 9472/09/$95.00

New Effect Size Rules of Thumb

Shlomo S. Sawilowsky
Wayne State University

Recommendations to expand Cohen's (1988) rules of thumb for interpreting effect sizes are given to include very small, very large, and huge effect sizes. The reasons for the expansion, and implications for designing Monte Carlo studies, are discussed.

Key words: Effect size, d, Monte Carlo simulation.

Introduction

Some primary considerations for conducting an appropriate Monte Carlo simulation were explicated in Sawilowsky (2003). For convenience, the list is repeated:

- the pseudo-random number generator has certain characteristics (e. g. a long period before repeating values);
- the pseudo-random number generator produces values that pass tests for randomness;
- the number of repetitions of the experiment is sufficiently large to ensure accuracy of results;
- the proper sampling technique is used;
- the algorithm used is valid for what is being modeled; and
- the study simulates the phenomenon in question.

The purpose of this article is to add the following two considerations:

- avoid the use of so-called true random number generators if the randomization process requires replication; and
- ensure study parameters are comprehensive, which necessitates new effect size rules of thumb.

Shlomo Sawilowsky is a WSU Distinguished Faculty Fellow, and Professor of Evaluation and Research. He is the founding editor of *JMASM*. Email: shlomo@wayne.edu.

Regarding the first addition, so-called true random number generators are based on sampling atmospheric or thermal noise, quantum optics, radioactive decay, or other such physical and deterministic phenomena. They aren't seeded, as are pseudo-random number generators, and hence it isn't possible to replicate the sequences they produce. The unscrupulous could make minor substitutions in the sequence to bias the results in such a way that may not be detectable by generic tests for randomness.

Lotteries, military conscriptions, or the like may attempt to overcome this limitation by having the public witness the process via direct observation, which is more compelling than video records that are easily alterable. However, in applications where transparency via replication is essential, such as random sampling in a study commissioned to support allegations in a lawsuit, the use of true random number generators are inappropriate. Thus, if the Monte Carlo study is also a simulation the appropriate number generator, so-called true or pseudo, must be chosen.

Regarding the second addition, Monte Carlo studies conducted on statistical tests' robustness and power properties require choices pertaining to sample sizes, alpha levels, number of tails, choice of competing statistics, inter-correlations of data structures, etc. The study parameters need not, however, be restricted to commonly occurring conditions. In Sawilowsky (1985), the rank transform was studied in the context of a 2×2×2 ANOVA employing sample sizes of 2 to 100 per cell. It is perhaps as unlikely that a classroom or clinic would contain

N=2 study participants as it is that there would be N=100 per cell. Those study parameters were chosen because they represented the minimum and the maximum sample sizes that could be handled given the constraints of the time-share mainframe computing resources available at that time. Prudence dictated sample sizes also be chosen between the two extremes to ensure there were no anomalies in the middle of the robustness rates or power spectrum.

Another important study parameter that must be considered in designing Monte Carlo simulations, which thanks to Cohen (e.g., 1962, 1969, 1977, 1988) has come to be the sin qua non of research design, is the effect size (for an overview, see Sawilowsky, Sawilowsky, & Grissom, in press). Previously, I discussed my conversations with Cohen on developing an encyclopedia of effect sizes:

> I had a series of written and telephone conversations with, and initiated by, Jacob Cohen. He recognized the weaknesses in educated guessing (Cohen, 1988, p. 12) or using his rules of thumb for small, medium, and large effect sizes (p. 532). I suggested cataloging and cross-referencing effect size information for sample size estimation and power analysis as a more deliberate alternative.
>
> Cohen expressed keen interest in this project. His support led to me to delivering a paper at the annual meeting of the AERA on the topic of a possible encyclopedia of effect sizes for education and psychology (Sawilowsky, 1996). The idea was to create something like the "physician's desk reference", but instead of medicines, the publication would be based on effect sizes. (Sawilowsky, 2003, p. 131).

In the context of the two independent sample layout, Cohen (1988) defined small, medium, and large effect sizes as d = .2, .5, and .8, respectively. Cohen (1988) warned about being flexible with these values and them becoming de facto standards for research. (See also Lenth, 2001.) Nevertheless, both warnings are summarily ignored today. That issue cannot be resolved here, but an important lesson that can be addressed is redressing the assumption in designing Monte Carlo studies that the effect size parameters need only conform to the minimum and maximum values of .2 and .8.

For example, when advising a former doctoral student on how to deconstruct the comparative power of the independent t test vs. the Wilcoxon test (Bridge, 2007), it was necessary to model very small effect sizes (e.g., .001, .01). This led to disproving the notion that when the former test fails to reject and the later test rejects it is because the latter is actually detecting a shift in scale instead of a shift in location. It would not have been possible to demonstrate this had the Monte Carlo study began by modeling effect sizes at .2.

Similarly, in the Monte Carlo study in 1985 mentioned above, I modeled what I called a very large effect size equivalent to d = 1.2. This was done because Walberg's (1984) collection of effect sizes pertaining to student learning outcomes included a magnitude of about 1.2 for the use of positive reinforcement as the intervention. Subsequently, in Monte Carlo studies I have conducted, and those conducted by my doctoral students that I supervised, the effect size parameters were extended to 1.2.

As the pursuit of quantifying effect sizes continued even larger effect sizes were obtained by researchers. For example, the use of cues as instructional strategies (d=1.25, Walberg & Lai, 1999), the student variable of prior knowledge (d = 1.43, Marzano, 2000, p. 69), and identifying similarities and differences (d = 1.6, Marzano, 2000, p. 63), exceeded what I defined as very large.

Incredibly, effect sizes on the use of mentoring as an instructional strategy to improve academic achievement have been reported in various studies and research textbooks to be as large as 2.0! The existence of such values, well beyond any rule of thumb heretofore published, has led to researchers presuming the studies yielding such results were flawed.

For example, when DuBois, et al. (2002) were confronted with study findings of huge effect sizes in their meta-analysis of mentoring, they resorted to attributing them as outliers and

deleting them from their study. This was just the first step to ignore the obvious. They then resorted to Winsorizing remaining "large effect sizes [as a] safeguard against these extreme values having undue influence," (p. 167). I have long railed against excommunicating raw data with a large percentage of extreme values as outliers, preferring to re-conceptualize the population as a mixed normal instead of a contaminated normal (assuming the underlying distribution is presumed to be Gaussian; the principle holds regardless of the parent population).

Recently, Hattie (2009) collected 800 meta-analyses that "encompassed 52,637 studies, and provided 146,142 effect sizes" (p. 15) pertaining to academic achievement. Figure 2.2 in Hattie (2009, p. 16) indicated about 75 studies with effect sizes greater than 1. Most fall in the bins of 1.05 to 1.09 and 1.15 to 1.19, but a few also fall in the 2.0+ bin.

Conclusion

Based on current research findings in the applied literature, it seems appropriate to revise the rules of thumb for effect sizes to now define d (.01) = very small, d (.2) = small, d (.5) = medium, d (.8) = large, d (1.2) = very large, and d (2.0) = huge. Hence, the list of conditions of an appropriate Monte Carlo study or simulation (Sawilowsky, 2003) should be expanded to incorporate these new minimum and maximum effect sizes, as well as appropriate values between the two end points.

References

Bridge, T. J. (2007). *Deconstructing the comparative power of the independent samples t test vs the Wilcoxon Mann-Whitney test for shift in location*. Unpublished doctoral dissertation, Detroit, MI: Wayne State University.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal Social Psychology*, *65*, 145-153.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. San Diego: Academic Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*, Rev. Ed. San Diego: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, NJ: Erlbaum.

Dubois, D. L., Holloway, B. E., Valentine, J. C., & Cooper, H. (2002). Effectiveness of mentoring programs for youth: A meta-analytic review. *American Journal of Community Psychology*, 30 (*2*), 157-197.

Hattie, J. (2009). Visible learning: a synthesis of over 800 meta-analyses relating to achievement. Park Square, OX: Rutledge.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, *55*, 187-193.

Marzano. (2000). A new era of school reform: Going where the research takes us, p. 63. Auroa, CO: Mid-Continent Research for Education and learning.

Sawilowsky, S. (2003a). You think you've got trivials? *Journal of Modern Applied Statistical Methods*, *2*(*1*) 218-225.

Sawilowsky, S. (2003b). A different future for social and behavioral science research. *Journal of Modern Applied Statistical Methods*, *2*(*1*), 128-132.

Sawilowsky, S. S., Sawilowsky, J., & Grissom, R. J. (in press). Effect size. *International Encyclopedia of Statistical Scienc*e. NY: Springer.

Walberg, H. J. (1984). Improving the productivity of America's schools. *Educational Leadership*, *41*(*8*), 19-27.

Walberg, H. J., & Lai, J-S. (1999). *Handbook of educational policy*. (G. J. Cizek, Ed.). San Diego, Academic Press, 419-453.