DIGITALCOMMONS
@WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 2 | Issue 1 Article 24

5-1-2003

Was Monte Carlo Necessary?

Thomas R. Knapp
tknapp5@juno.com

Recommended Citation
Knapp, Thomas R. (2003) "Was Monte Carlo Necessary?," *Journal of Modern Applied Statistical Methods*: Vol. 2 : Iss. 1 , Article 24.
DOI: 10.22237/jmasm/1051748640
Available at: http://digitalcommons.wayne.edu/jmasm/vol2/iss1/24

Journal of Modern Applied Statistical Methods
May 2003, Vol. 2, No.1, 237-241

1538 – 9472/03/$30.00

Invited Debate: Comment
Was Monte Carlo Necessary?

Thomas R. Knapp
Kailua-Kona, Hawaii

In the critique that follows, I have attempted to summarize the principal disagreements between Sawilowsky and Roberts & Henson regarding the reporting and interpreting of statistically non-significant effect sizes, and to provide my own personal evaluations of their respective arguments.

Keywords: Non-significant effect sizes; Monte Carlo investigations

Introduction

There are three principal matters to consider. They are (in my opinion) in order of decreasing importance:

The Reporting and Interpreting of Non-significant Statistics.

I think that the matter of reporting, interpreting, publishing, etc. statistically non-significant effect sizes can be argued without appealing to the results of any Monte Carlo investigations. Indeed, that matter has been debated almost ad nauseam over the last half-century, as the reference to Melton (1962) in the exchange between Knapp & Sawilowsky (2001) and Thompson (2001) indicates.

Consider, for example, a researcher who draws a simple random sample from a population, assumes linearity and bivariate normality, calculates a Pearson product-moment correlation coefficient (one of the simplest and most important effect size measures) between two variables for the sample, tests it for statistical significance, and gets a p-value of .03.

Should that correlation be reported? Of course; the correlation between those two variables in that sample is ___. Should it be interpreted? Of course; that correlation is not statistically significant at the .01 level, is statistically significant at the .05 level, etc. (depending upon the value of alpha chosen at the outset of the study). [Or, if interval estimation is preferred, one's confidence is .99 (or .95, or whatever) that the interval from ___ to ___ covers the population correlation.]

Should that study be published? Aye, there's the rub. Melton wouldn't have (he insisted that p be less than .01); I presume Sawilowsky & Yoon wouldn't either; and I further presume that Roberts & Henson would – all other things being equal (good theory, design, measurement, etc.). If statistically non-significant findings are not published occasionally, the literature will have an imbalance of Type I errors.

One-sided vs Two-sided Inference

If I'm wrong and if one does need Monte Carlo evidence in order to decide whether or not a statistically non-significant effect size is of interest, should the focus be on one-sided inference or two-sided inference? Sawilowsky & Yoon (2002) chose two-sided inference and concentrated on the absolute value of Cohen's d. Roberts & Henson (2002) chose one-sided inference by concentrating on d's that were greater than or equal to 0 (with the alternative hypothesis taken to be that the experimental mean is greater than the control mean). I agree with Roberts & Henson, since that better reflects the more typical

Thomas R. Knapp, Ed. D. (Harvard, 1959) is Professor Emeritus of Education and Nursing, University of Rochester and The Ohio State University. Email him at tknapp5@juno.com

research hypothesis and is also simpler (it involves only two sampling distributions rather than three).

Technical Aspects of Monte Carlo Investigations

Sawilowsky & Yoon (2002) carried out one kind of Monte Carlo investigation. Roberts & Henson (2002) carried out another kind of Monte Carlo investigation. The particular details (number of replications, Fortran vs S-Plus, etc.) also differed. I have no idea who's right and who's wrong there.

Specific Comments

Sawilowsky & Yoon (2002)

1. They chose sample sizes of 10 and 10, and a power level of .2 "to mimic applied research" (p. 143). Those sample sizes strike me as too small for typical educational experiments, and there is a considerable amount of evidence (see, for example, Aberson, et al., 2002) that the average a priori power for published studies in education is approximately .5, not .2.

2. Their Monte Carlo investigation revealed an average obtained absolute effect size of .169 for statistically non-significant results when comparing the means of two samples of size 10 drawn from normal populations in which the population effect size was zero. I believe such a result could have been determined analytically (mathematically), and I also believe that .169 is actually too low. In the Appendix to this critique I have outlined a proof of those beliefs.

I conclude that the Sawilowsky & Yoon (2002) research was not necessary.

Roberts & Henson (2002)

Roberts & Henson (2002) were reacting to Sawilowsky & Yoon (2001), not Sawilowsky & Yoon (2002), but those two papers are almost identical.

1. In their opening sentence, Roberts & Henson (2002) referred to a controversy between "the role and function of effect sizes" and the use of "statistical significance tests" (p. 241). That is a false comparison. People who use statistical significance tests have almost always calculated some sorts of sample effect sizes *before* they carry out the significance tests (see the Pearson r example, above).

The general controversy involves whether or not significance tests should be prohibited; the specific controversy between Sawilowsky and Roberts & Henson involves whether or not statistically non-significant sample effect sizes should be reported and interpreted.

2. They (Roberts & Henson) went through an elaborate discussion of Thompson's (2002) recommendation of converting d to r, Friedman's (1968) formula for converting r to d, Ezekiel's (1930) correction formula, etc. That is unnecessary. All one needs to do is algebraically re-solve the d-to-r formula given by Cohen (1988) for r in terms of d (but see Aaron, Kromrey, & Ferron, 1998 regarding that formula - it only works for equal and large n's) and/or appeal to the work of Hedges (1981), Kraemer (1983), and Hedges & Olkin (1985) concerning the amount of bias in Cohen's d.

3. They then went on to report in three separate tables the results of their Monte Carlo investigation, for various values of Cohen's d in the population, various values of the population standard deviation (the mean for the control group was taken to be 100), and various sample sizes, including the $n_1 = n_2 = 10$ case that was of interest to Sawilowsky & Yoon. Several of those results are already reasonably well known.

The expected value (mean) of a sample r^2 is equal to 1/(N-1) when the population r^2 is equal to zero (see, for example, Marascuilo & Levin, 1983, p. 97), so the small differences between that expected value for $n_1 = n_2 = 10$ (an N = 20), i.e., .0526315..., and the mean sample r^2 for a population r^2 of 0 in their tables are all attributable to Monte Carlo sampling variation. Formulas for the expected value and sampling variance for Cohen's d can be found in Hedges (1981), in Kraemer (1983), and in Hedges & Olkin (1985, pp. 78-81), so their results for d differ from those derived mathematically also because of Monte Carlo sampling variation.

Some of the other results are a bit baffling. For example, why isn't the Bias row for d in each of those tables equal to the difference between the mean sample d and the d in the population? [Is it because of the discrepancies between the desired population d's and the Monte Carlo population d's to which they referred on page 247?] And how can the bias for the sample d for a population d of .20 be *greater* for n's of 100 than for n's of 50 in both Table 1 and Table 2?

4. In their concluding section Roberts & Henson (2002) claimed that "...replication of a given study is the only true way to evaluate possible generalizability" (p. 252). I agree (by definition). They went on to say that "Statistically nonsignificant effects may be fully replicable." Of course; if nothing is going on, nothing will keep getting replicated, but that doesn't help their argument.

I equally regretfully conclude that the Roberts & Henson (2002) research was also not necessary.

Sawilowsky (2003)

1. He drew several distinctions among simulation, Monte Carlo, Monte Carlo simulation, sampling with replacement vs. sampling without replacement, and characteristics of a "high quality Monte Carlo simulation" (p. 218) The first three and the fifth are apparently important to make in any Monte Carlo investigation (I leave that to others to decide). The fourth distinction (sampling with replacement vs. sampling without replacement) is of course always important to make, especially when it comes to sampling within sample and sampling between samples.

Under "Monte Carlo" he properly acknowledged that there are some situations, such as finding the definite integral from 0 to 1 of f(x) = x, where the Monte Carlo approach could be used but should not be. However, under "Sampling With vs. Without Replacement" he claimed that sampling without replacement is appropriate when sampling from a deck of cards. I disagree; such sampling can be either with or without replacement within sample - it all depends upon whether or not a sampled card gets replaced in the deck prior to the sampling of a subsequent card - but sampling must be with replacement between samples or you soon run out of cards to sample.

2. The remainder of his response is the heart of his paper (in my opinion). He first listed what he called "Nine Minor Criticisms" of Roberts & Henson (2002). I would have identified at least two of those (# 7 and # 8) as major criticisms. Why Roberts & Henson bothered with three different tables is beyond me (their rationale on page 246 is interesting but irrelevant, given that both d and r^2 are scale free); and I have already indicated above in my Comments #2 and #3 regarding their study that the bias in d had already been addressed analytically by Hedges (1981), by Kraemer (1983), and by Hedges & Olkin (1985).

Sawilowsky's "Major Criticism" apparently has to do with the kinds of results one might obtain when sampling from populations with d's of 0, and with the order in which the results appear. I found that section rather difficult to follow. I guess the point he's making is that the findings in Data Set B are more likely to be obtained and will look more impressive than the findings in Data Set A, but the obtained effect sizes in both data sets could easily be attributable to chance.

Roberts & Henson (2003)

1. At the beginning of their paper, Roberts & Henson (2003) stated that the first portion of Sawilowsky's (2003) paper "does not bear comment on" (p. 226). Although I don't particularly care for Monte Carlo investigations, Roberts and Henson apparently do (since their research was such an investigation), and Sawilowsky's claims concerning how a good Monte Carlo simulation should be carried out deserved a response. (They did comment on some technical Monte Carlo features in their responses to Sawilowsky's minor criticisms.)

2. They then went on to address all of Sawilowsky's minor criticisms. I have already implied my lack of interest in #2 and #5. And they appear to have accepted criticisms #1, 3, and 6. Where they disagreed most with Sawilowsky is with respect to criticisms #4, 7, 8, and 9. I shall accordingly concentrate on those matters.

As indicated above, I agree with them regarding negative values of d (#4). But I take exception to their responses to those last three criticisms. Their paragraph (regarding #7) that bears the heading "Redundancy is reinforcement!" (with an exclamation point yet) is bizarre. As Sawilowsky (2003) pointed out, and as I argued above, there was no good reason for including all three tables. Their sentence "We would argue that if the results were redundant then we would see exactly the same values in each of the tables, which we in fact did not." (p. 229) shows a lack of understanding of Monte Carlo. It is inherent in the method that you do not get "exactly" anything; it is subject to sampling variation just like any sample statistic is. And they missed the point regarding #8. The published work on the bias of d obviated the need for Monte Carlo. (I'm not sure what point they were trying to make regarding #9, other than the fact that Type II errors are possible.)

3. They concluded their paper by responding to Sawilowsky's (2003) major criticism. They may have been even more confused than I was by that section of Sawilowsky's paper, because they seemed to be talking about Type II error all over again, introducing several citations to the literature on misconceptions regarding significance testing, etc. rather than directly addressing Sawilowsky's examples of data sets that could be realized and how they should be interpreted.

References

Aaron, B., Kromrey, J.D. & Ferron, J.M. (November, 1998). Equating r-based and d-based effect size indices: Problems with a commonly recommended formula. Paper presented at the annual meeting of the Florida Educational Research Association, Orlando, FL. (ERIC Document Reproduction Service No. ED 433 353)

Aberson, C.L., Berger, D.E., Healy, M.R., & Romero, V.L. (2002). An interactive tutorial for teaching statistical power. *Journal of Statistics Education*, *10* (3) [Online].

Ezekiel, M. (1930). *Methods of correlational analysis*. New York: Wiley.

Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin*, *70*, 245-251.

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

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

Knapp, T.R., & Sawilowsky, S.S. (2001). Constructive criticisms of methodological and editorial practices. *Journal of Experimental Education*, *70* (1), 65-79.

Kraemer, H.C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. *Journal of Educational Statistics*, *8*, 93-101.

Marascuilo, L.A., & Levin, J.R. (1983). *Multivariate statistics in the social sciences*. Monterey, CA: Brooks/Cole.

Melton, A.W. (1962). Editorial. *Journal of Experimental Psychology*, *64*, 553-557.

Roberts, J.K., & Henson, R.K. (2002). Correction for bias in estimating effect sizes. *Educational and Psychological Measurement*, *62* (2), 241-253.

Roberts, J.K., & Henson, R.K. (2003). Not all effects are created equal: A rejoinder to Sawilowsky. *Journal of Modern Applied Statistical Methods*, *2*(1), 227-231.

Sawilowsky, S.S. (2003). You think you've got trivials? *Journal of Modern Applied Statistical Methods*, *2*(1), 218-226.

Sawilowsky, S.S., & Yoon, J.S. (August, 2001). The trouble with trivials (p>.05). Paper presented at the 53rd Session of the International Statistical Institute, Seoul, South Korea.

Sawilowsky, S.S., & Yoon, J.S. (2002). The trouble with trivials (p>.05). *Journal of Modern Applied Statistical Methods*, *1* (1), 143-144.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education*, *70* (1), 80-93.

Thompson, B. (2002). "Statistical", "practical", and "clinical". How many kinds of significance do counselors need to consider? *Journal of Counseling and Development*, *80*, 64-71.

Appendix

It can be shown (personal communication from Ingram Olkin, May 5, 2003) that the expected value of the absolute value of Cohen's d, i.e., E(|d|), can be expressed as an infinite series in terms of gamma functions of the two sample sizes and in terms of the population effect size. If the population effect size is equal to zero and $n_1 = n_2 = 10$ (the case of particular interest to Sawilowsky and one of the cases of interest to Roberts & Henson), E(|d|) is approximately .3726.

Kraemer (1983) showed that d follows the t sampling distribution with n_1 + n_2 -2 degrees of freedom and provided a formula for calculating the percentiles of that distribution. From the 97.5 th percentile it can be determined that the cut-off point for the .05 significance level is approximately .940 for the absolute value of d.

And, from the middle 95% of that distribution it can be determined that the mean of the "non-rejectable" absolute values of d is approximately .336 (not .169). By appealing to the formula for a weighted mean it can be further determined that the mean of the "rejectable" absolute values of d is approximately 1.076 (not .508).