New Statistical Rituals for Old

Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis
By Geoff Cumming
New York: Routledge, 2012. 519 pp. ISBN 978-0-415-87967-5 (hardcover); ISBN 978-0-415-87968-2 (paperback). $100.00, hardcover; $46.95, paperback

Reviewed by Michael Palij

Michael Palij, New York University, 7 East 12th Street, 7th Floor, New York, NY 10003,
E-mail: mp26@nyu.edu

Date: March 12, 2012

PsycCRITIQUES 01/2012; 57(24). DOI:10.1037/a0028079

Critics never tire of pointing out that psychologists are notoriously lousy at the use of statistics, in part because they were never taught the proper statistics and they use statistical analysis as a form of automatic inference making based on "statistical rituals." As Gigerenzer (2004, p. 587) notes, "Statistical rituals largely eliminate statistical thinking in the social sciences." The focus of Gigerenzer's ire are the rituals associated with null hypothesis statistical testing (NHST). Geoff Cumming agrees with this view and goes further by asserting that NHST promotes a set of procedures that apparently corrupt the mind and lead to confusion and error. Cumming attempts to remedy the situation with the "new statistics" in his book *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*.

Then again, perhaps things are not as bad as critics assert even if there is trouble in River City. There is agreement that aspects of NHST are problematic and that practices should change (e.g., Wilkinson, & Task Force on Statistical Inference, 1999), but not all suggested changes may be for the better or even correct (remember P_{rep}; see Maraun & Gabriel, 2010). There is always the danger of substituting new rituals for old because (a) as with NHST, there is confusion about what is correct practice, and (b) as with NHST, practitioners are confused about what the practices or rituals actually do and mean.

Cumming's "New Statistics" are, as he acknowledges, not new at all — no Bayesian analysis or scary mathematical modeling of phenomena or other approaches (e.g., Ludbrock and Dudley, 1998). Instead, it is based on the use of effect sizes (ES), confidence intervals (CI), and meta-analyses instead of NHST (i.e., the "Old Statistics"). These procedures have been around, in one form or another, for over 100 years yet have not found their way into the everyday statistical practices of psychologists, though many researchers are aware of them and use them to some degree. So, what is really new?

What Cumming is referring to as "new" can be thought of as a project that has two parts: (a) establishing that the use of NHST leads to confusion and error to a greater degree than the use of ES/CI (e.g., NHST encourages "dichotomous thinking" because of the yes-no decision aspect while ES/CI encourages thinking in terms of "how much?"), and (b) research on statistical cognition that focuses on determining how statistical concepts should be taught, what is the most effective mode for their presentation, and fostering correct statistical thinking and inference (which would begin with the exclusive use of ES/CI). In my view, the latter is not necessarily connected to the former point (i.e., NHST may be clearer than ES/CI in certain cases).

This volume's real strength is that it provides a good overview of the different types of ES, how to view CI in a variety of situations, and how to conduct meta-analyses. Cumming even acknowledges that there are problems with the "new statistics" that have to be worked out before progress can be made (e.g., confusion over the ES Cohen's *d* when it is difficult to determine which version is being reported; see "Names and Symbols" pp. 295-296). One can find similar information elsewhere but having all this information in one volume is a good thing; however, I would suggest that those interested in meta-analysis examine the Cochrane Collaboration resources (www.cochrane.org) and check the recent literature on meta-analysis (e.g., Ioannides 2005).

Also, Cumming has created the freeware Excel add-on ESCI (Ess-key; Exploratory Software for Confidence Intervals available at www.thenewstatistics.com), which he uses in his examples. One does not have to agree with Cumming to find ESCI useful.

I think that there is much merit in Cumming's text and that it would make a good core text in undergraduate or graduate statistics courses ***after*** the traditional introductory course on psychological statistics. I can readily see how one might run a seminar using this text and an

appropriate list of readings (including those "odd" articles by researchers who have not yet gotten "religion" and are still advocating for NHST such as Hagen, 1997). Cumming raises a number of questions about what constitutes good statistical practice that students and even researchers should be exposed to and think about. But I think that the "take away message" from such a course should include the following: (a) there is no single set of statistical practices that everyone agrees with but one should be able to recognize bad statistical practices regardless; (b) ultimately, the statistical procedures one uses should be those that help to provide greater insight into the phenomena being studied; and (c) as with other empirical disciplines, new developments should be expected in statistical analysis in the future that may change how data are analyzed as well as how one can/should conceptualize the phenomena under investigation.

I now raise some issues of my own.

Issue #1: Is This "New" versus "Old" Statistics Argument Really Important?

Many of the issues Cumming raises have been raised before (e.g., Kirk 1996), so, what's the big deal? Consider the following comments by Ray Nickerson (2000) in his review of NHST:

> One of the people who gave me very useful feedback on a draft of this article questioned the accuracy of my claim that NHST is very controversial. "I think the impression that NHST is very controversial comes from focusing on the collection of articles you review—the product of a batch of authors arguing with each other and rarely even glancing at actual researchers outside the circle except to lament how little the researchers seem to benefit from all the sage advice being aimed by the debaters at both sides of almost every issue." The implication seems to be that the "controversy" is largely a manufactured one, of interest primarily—if not only—to those relatively few authors

> who benefit from keeping it alive. I must admit that this comment, from a psychologist for whom I have the highest esteem, gave me some pause about the wisdom of investing more time and effort in this article. I am convinced, however, that the controversy is real enough and that it deserves more attention from users of NHST than it has received. (Nickerson 2002, p. 241, footnote 2)

It should also be noted that though psychologists have presented NHST and ES/CI as opposing frameworks, this does not have to be the case. Consider Krantz's (1999) comments in his review of the book *What If There Were No Significance Tests?*, which was written for the *Journal of the American Statistical Association* (JASA):

> The book was edited and written by psychologists, and its title was well designed to be shocking to most psychologists. The difficulty in reviewing it for JASA is that the issue debated may seem rather trivial to many statisticians. The very existence of two divergent groups of experts, one group who view this issue as vitally important and one who might regard it as trivial, seemed to me an important aspect of modern statistical practice. (p. 1372)

Krantz was referring to the fact that most statisticians would not see the use of NHST and ES/CIs as an "either/or" issue, a position that Cumming seems to promote (i.e., "Just say NO to NHST").

Now, it may be true that many researchers have failed to report ES/CI (or report the wrong ones when they do), but there is nothing in NHST that prohibits the use of ES/CI. Perhaps when ES/CI play a significant role in the theories that researchers use, they will be more frequently used and reported in meaningful ways. In signal detection theory (SDT), the sensitivity parameter **d'** is almost equivalent to the ES measure **d,** and it would be hard to use

SDT without reference to **d'** or an equivalent. When ES/CI plays an equivalent role in psychological theories, they will then be similarly indispensable.

Issue #2: ES/CI and Meta-Analysis for Disappearing Effects or "The Decline Effect"

Cumming makes a number of assumptions in his presentation, and one implicit assumption is that all effect sizes, even if they are not statistically significant, are "real." Cumming's first example in his book uses two studies, one with statistically significant results, and the other with nonsignificant results. Cumming believes that an ES/CI and meta-analysis approach solves problems of such "inconsistency of results," apparently under the assumption that both are true effects (see Coulson, Healey, Fidler, & Cumming 2011 for details). This example reminded me of other situations where an initial statistically significant effect is found but the effect is reduced with each replication and ultimately is not replicated. What is one to make of these situations?

Such a pattern appears in Dworkin and Miller's (1986) Figure 1 that describes effect size as a function of year of publication for research that reported the operant conditioning of the autonomic nervous system. After a certain year no one was able to reproduce the effect. Dworkin and Miller investigated this and reported that after tremendous effort at replication, they were unable to do so. They concluded that the original results were real but the replications lacked something present in the original situations. Others have made less charitable assessments. As far as I know, no one has replicated the original finding after Dworkin and Miller's report. This raises the question of what were those effects?

The pattern shown by Dworkin and Miller is not unique and has been shown to occur in other situations. Ioannides (2005) has probably been most provocative in describing this phenomena and how frequently it occurs. Schooler (2011) has observed this result in his own

research on the verbal overshadowing effect and had pointed out that it was first described in detail by J. B. Rhine in studies on parapsychology (see Jonah Lehrer's 2010 *New Yorker* article for a longer discussion of this pattern). Rhine noted that in many of his studies a person would strongly show a psychic ability at first but it got smaller under repeated testing – a pattern he called the "decline effect." Apparently it occurred so often that it became an expected result for people who showed psychic abilities.

There are two questions here: (1) are we dealing with real effects, artifacts, or Type I errors, and (2) what role should ES/CI and meta-analyses play in these situations? Francis (2012) provides one solution using a test by Ioannides for the expected number of false positives in publications (used as an adjunct in meta-analysis). He analyzes the results from replications of Schooler's verbal overshadowing experiment as well as the PSI experiments reported by Bem that purport to show how the future influences the present. The analyses suggest that there is no "there" or ES there (i.e., these are false positive results). I recommend Francis's paper because he seems to have come up with a method that explains the decline effect and shows what other analyses have gotten wrong (spoiler alert: he did not get advice from the future, but NHST is involved). It should also serve as a caution about the uncritical use of ES/CI and meta-analysis.

Epilogue

I have memory that may be unreliable but is relevant here. Back in the 1970s, just as I was starting graduate studies, I attended a symposium at the New York Academy of Sciences where David Bakan was presenting on the problems of NHST. At the end of the session, I went up to Bakan and asked him questions that ultimately led to "what's the best attitude to have in these situations?" His response then was wise and to the point. He said: "Listen, just don't be a schmuck." Good advice anytime.

References

Coulson, M., Healey, M., Fidler, F. & Cumming, G. (2010). Confidence Intervals Permit, but Do Not Guarantee, Better Inference than Statistical Significance Testing. *Frontiers in Psychology*,1, 26. doi: 10.3389/fpsyg.2010.00026

Dworkin, B. R., & Miller, N. E. (1986). Failure to replicate visceral learning in the acute curarized rat preparation. *Behavioral Neuroscience, 100*(3), 299-314. doi:10.1037/0735-7044.100.3.299

Francis, G. (2012 online). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review, 19*, 1-6. doi:10.3758/s13423-012-0227-9

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Scoio-Economics, 33*, 587-606. doi:10.1016/j.socee.2004.09.033

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52*(1), 15-24.

Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Med 2*(8): e124. doi:10.1371/journal.pmed.0020124

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746-759. doi:10.1177/0013164496056005002

Krantz, D.H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 94*(448), 1372-1381.

Lehrer, J. (2010, December 13). The truth wears off. *The New Yorker, 86*(40), 52.

Ludbrook, J. & Dudley, H. (1999). Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician, 52*(2), 127-132.

Maraun, M. & Gabriel, S. (2010). Killeen's (2005) p_{rep} coefficient: Logical and mathematical problems. *Psychological Methods, 15*(2), 182-191. doi: 10.1037/a0016955

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241-301. doi:10.1037/1082-989X.5.2.241

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature, 470*, 437. doi:10.1038/470437a

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-604. doi:10.1037/0003-066X.54.8.594