Critique of Cumming's "New Statistics" for Psychological Research: A Perspective from Outside Psychology

James E. Kennedy
Version of 10/31/2016

This paper is available on the internet in pdf and HTML at
http://jeksite.org/psi/critique_new_stat.pdf and http://jeksite.org/psi/critique_new_stat.htm

My graduate statistical training was in biostatistics rather than psychology. The typical context for statistical analysis was questions like: should a physician prescribe a certain treatment to a patient, or should a regulatory agency approve a treatment for use? The primary purpose of research and analysis was to provide a discrete yes/no answer for these types of questions.

The statistical classes taught that there were two aspects of statistical analysis: hypothesis testing and estimation. These answered different questions. Hypothesis testing addressed "does an effect occur?" Estimation addressed "how large is the effect?" We were taught that the two questions are closely related and that both are essential for scientific understanding. A properly designed hypothesis test includes a power analysis based on certain effect sizes. The more precise question is "does an effect of size X or larger occur?" Similarly, the confidence intervals for estimates of the size of an effect can be used to obtain the yes/no answer for a hypothesis test. Hypothesis testing was given a prominent role because the logic of the yes/no decision was conceptually straightforward, well developed mathematically, and fit the overall purposes of the research.

My subsequent work included analyzing data for environmental and public health regulations, academic medical research, and regulated pharmaceutical research. Regulated medical research has significantly higher methodological standards than academic medical research, but my experience in both was that well-powered confirmatory research was regarded as the standard for scientific evidence. Research on therapeutic interventions typically is divided into two stages: initial small exploratory studies and subsequent confirmatory studies designed with power analysis for strong conclusions.

I have also had an interest in parapsychology and have followed the research and associated controversies. In 2004 I published a paper arguing that the usual research methods for academic psychology were not adequate for a controversial subject like parapsychology and that pre-registered, well-powered confirmatory research was needed (Kennedy, 2004). That proposal received little interest at that time, but the recent methodological revolution in psychology has finally made that perspective fashionable (Open Science Collaboration, 2012; Wagenmakers, Wetzels, Borsboom, van der Maas & Kevit, 2012). In the course of following these developments, I have become aware of the "new statistics" that Cumming (2012, 2014) advocates for psychological research.

The “new statistics” focus on estimation and meta-analyses and openly denigrate hypothesis testing and associated “dichotomous” thinking. Several points need to be made about this approach.

Confirmatory Research and Power Analysis

Good confirmatory research, like good science in general, is based on making and testing specific predictions. Exploratory research focuses on estimating effect sizes without specific predictions. Cumming’s new statistics attempt to set exploratory research methods as the standard for scientific evidence and avoid the dichotomous thinking associated with testing specific predictions. That is a major retreat from the basic principles of good science and allows researchers to avoid confronting the possibility that their ideas may not be valid.

Psychological researchers have generally not distinguished between exploratory and confirmatory research. This resulted in a research culture that focused on exploratory research. Exploratory research is usually the creative step that is the starting point for a line of research. However, the flexibility of exploratory research makes it prone to various biases and questionable research practices (John, Loewebsteun & Prelec, 2012; Simmons, Nelson & Simonsohn, 2011). Properly conducted confirmatory research eliminates these problematic practices and provides the convincing evidence that makes science valid and self-correcting.

One conspicuous symptom of the emphasis on exploratory research is that formal power analysis has been rarely used to determine sample size in psychological research. Disregard for power analysis is common for exploratory research, but is not appropriate for confirmatory research. The lack of confirmatory research designed with power analysis is a major cause of the problems with hypothesis testing that Cumming describes.

Cohen (1990) concisely summarized the need for power analysis as “failure to subject your research plans to power analysis is simply irrational” (p. 1310). Textbooks on statistics for psychologists have long explained the central role of power analysis and inferential errors (e.g., Hayes, 1963; Keppel, 1973), yet this fundamental principle has been routinely ignored in practice (Bakker, van Dijik & Wicherts, 2012; Cumming, 2012; Kennedy, 2013). Power analysis evaluates the probability that a study will provide correct inferences and establishes the statistical validity of a planned analysis. Power analysis was used during study planning for all of the confirmatory medical studies I was involved with, and for many of the exploratory studies. The logic of hypothesis testing breaks down if power analysis is not part of the study design. Cumming appears to be largely blaming the statistical methods rather than the research culture for the irrational statistical practices (as characterized by Cohen) that have been common in psychology.

Studies with low power or unknown power are basically a biased form of research. An underpowered study can provide evidence that the alternative or experimental hypothesis is true, but cannot provide evidence that the experimental hypothesis is false. Significant results are interpreted as supporting the alternative hypothesis, but non-significant results are ambiguous because the outcome could be due to low power rather than to the experimental hypothesis being false.

The effect size used in the power analysis is a prediction about what will happen in an experiment. For a study with high power, a nonsignificant result provides evidence that the predicted effect size specified in the power analysis is false. A nonsignificant result does not provide direct evidence that the null hypothesis is true because a small, nonzero effect size could be true. The null hypothesis is basically irrelevant other than for developing a test for the effect predicted with the power analysis.

Researchers who abuse hypothesis tests focus on the null hypothesis and ignore statistical power. This prevents evidence that the experimental hypothesis is false and avoids the confirmatory question of whether the researchers can make reliable predictions about the effect. Most criticisms of hypothesis testing are actually criticisms of this abuse and are not applicable for proper applications of hypothesis tests with power analysis. Lack of falsifiable predictions is a common denominator for the past abuse of hypothesis testing and for Cumming's "new statistics."

Power analysis also has a fundamental role in confirmatory studies analyzed with Bayesian methods. Most writings on Bayesian analysis have focused on exploratory research and have not addressed power analysis. However, power analysis is needed for confirmatory research with Bayesian analyses (U.S. Food and Drug Administration, 2010; Kennedy, 2015; Kruschke, 2011). Commonly used Bayesian hypothesis testing methods can be biased (Kennedy, 2015; Kruschke, 2011 p. 311). Power analysis at the planning stage reveals these biases and is essential for confirmatory research.

Psychological researchers tend to discuss research "replications," whereas medical researchers discuss "confirmatory studies." The difference between these terms is that confirmation generally implies better methodological standards than were used in the earlier exploratory studies, whereas replication implies repeating the study with the same or similar methodology. Power analysis is often ignored for replication studies, but is not ignored for confirmatory studies. The term replication promotes the lack of distinction between exploratory and confirmatory research methods.

Power analysis is one of several methodological factors that distinguish confirmatory research from exploratory research. Other components of confirmatory methodology include: pre-specification of the statistical methods and the criteria for acceptable evidence, public pre-registration of experiments, experimental procedures that make intentional or unintentional data alterations by one person difficult, documented formal validation of software, and sharing data for analyses by others (Kennedy, 2013, 2014; KPU Registry, 2014).

The "new statistics" continue to blur the distinction between exploratory and confirmatory research and explicitly avoid making and testing specific predictions. This strategy continues the underlying scientific weakness of past research practices.

Limitations of Meta-Analysis

Cumming's discussion of "new statistics" does not address the limitations of meta-analysis. Typical retrospective meta-analysis is a form of post hoc analysis that has many of the intrinsic limitations of other types of post hoc analyses. Meta-analysis involves many decisions, including what studies to include, what effect size to use, what type of analyses to perform, how to

evaluate and correct small study effects, how to handle heterogeneity, and what moderating factors to consider. Different choices for these decisions can produce different results. Analysts typically have substantial knowledge of the database when they make these decisions, which inevitably introduce potential for bias. In addition, the evaluation of moderating factors, including methodological issues, is correlational analysis of observational data (Cooper & Hedges, 2009; Kennedy, 2013). Researchers who quickly recognize the limitations of correlational analysis of observational data in an individual study have been slow to recognize that the same limitations apply for many evaluations in meta-analyses.

Ferguson and Heene (2012, p. 558) recently commented that they "have seldom seen a meta-analysis resolve a controversial debate in a field. ... [W]e observe that the notion that meta-analyses are arbiters of data-driven debates does not appear to hold true." Although parapsychology was not discussed by Ferguson and Heene, the endless debates about meta-analyses in parapsychology are a clear example (Bosch, Steinkamp, & Boller, 2006; Hyman, 2010; Kennedy, 2013; Radin, Nelson, Dobyns, & Houtkooper, 2006; Schmeidler & Edge, 1999; Storm, 2000; Storm, Tressoldi, & Di Risio, 2010). Like other types of post hoc analyses, meta-analyses can be useful, but are not effective at resolving scientific controversies.

The strongest evidence for an effect comes from a group of well-powered confirmatory studies that produce reliable results (Kennedy, 2013). Ioannidis (2005) developed a measure of the "positive predictive value" PPV for different research methods. He estimated that a well-powered experiment has a PPV of .85, whereas a meta-analysis of underpowered studies has a PPV of only .41. After describing the post hoc, observational aspects of meta-analysis, Cooper and Hedges (2009) emphasized that "a research synthesis should never be considered a replacement for new primary research" (p. 564).

The "new statistics" overemphasize meta-analyses and underemphasize the need for strong confirmatory studies. This strategy is counterproductive when the validity of a controversial hypothesis is in question.

Scientific Conclusions or Endless Research

Cumming strongly advocates that researchers avoid "dichotomous thinking," and thus avoid yes/no conclusions about the validity of effects. This strategy focuses on continued collection of data without criteria for making inferences about hypotheses.

To put matters in perspective, the last medical research I was involved with was treating patients whose life expectancy was 4.5 months. That research was conducted with a sense of urgency and a dichotomous yes/no decision was needed. Academic research strategies that have the primary goal of generating publications and typically recommend further studies were not appropriate in that situation.

The ultimate driving force for the unfortunate statistical practices in psychological research appears to be the incentives and goals of the research culture (Nosek, Spies & Motyl, 2012). A primary goal of psychological research has been to generate publications that have become required for professional survival. The competitive research culture became fixated on *p* values with little regard for statistical power, inferential errors, or the distinction between exploratory and confirmatory research.

Changing the focus from hypothesis testing to estimation does not address the root cause of the problems. The research culture needs to recognize and reward well-designed confirmatory research. However, the "new statistics" avoid drawing inferences about the validity of effects as occurs with confirmatory research and promote continued ambiguity between exploration and confirmation.

At some point the "new statistics" will be recognized as shifting from one excess to another. This strategy would promote the generation of academic publications, but would continue to miss the balanced middle ground that is needed for efficient, effective scientific research.

Conclusions

Carefully designed, well-powered, pre-registered confirmatory studies that are analyzed with hypothesis tests are optimal for (a) research when human life is directly involved and answers are urgently needed, (b) controversial research such as parapsychology, and (c) any case when researchers want to provide the strongest evidence that they understand and can control an effect. These hypothesis tests require that researchers make and evaluate predictions about an effect. The hypothesis tests can be classical or Bayesian, but appropriate power analysis is needed with either approach. If reliable results cannot be demonstrated with a group of well-designed, well-powered confirmatory studies, the effects, or the researchers' understanding of the effects, should not be considered valid.

Estimation methods that exclude hypothesis tests are appropriate in the exploratory stage of research, and/or when the basic validity of an effect is not in question. Estimation methods are particularly appealing when the priority is to maximize the number of publications for academic research that does not need yes/no answers about the validity of effects. Psychological research often has these characteristics. The "new statistics" for psychologists appear to be focused more on promoting publications than on drawing strong inferences about the validity of effects. The focus on estimation and meta-analysis continues to overemphasize exploratory research and to underemphasize the need for strong confirmatory studies. The limitations of this approach will become increasingly apparent over time, and the evolution of more balanced statistical perspectives can be expected.

References

Bakker, M., van Dijik, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7,* 543–554. Retrieved from http://pps.sagepub.com/content/7/6/543.full.pdf+html

Bosch, H., Steinkamp, E., & Boller, E. (2006). In the eye of the beholder: Reply to Wilson and Shadish (2006) and Radin, Nelson, Dobyns, and Houtkooper (2006). *Psychological Bulletin, 132*, 533-537.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Cooper, H., & Hedges, L. V. (2009). Potential and limitations. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.) (pp. 561-572). New York: Sage.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7-29. http://dx.doi.org/10.1177/0956797613504966

Ferguson, C.J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*, 555-561. Retrieved from http://pps.sagepub.com/content/7/6/555.full.pdf+html

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

Hyman, R. (2010). Meta-analysis that conceals more than it reveals: Comment on Storm et al. (2010). *Psychological Bulletin, 136*, 486-490.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*, 0696-0701. Retrieved from http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124.

John, L.K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524-532. http://dx.doi.org/10.1177/0956797611430953

Kennedy, J. E. (2004). A proposal and challenge for proponents and skeptics of psi. *Journal of Parapsychology, 68*, 157-167. Retrieved from http:/jeksite.org/psi/jp04.pdf

Kennedy, J. E. (2013). Can parapsychology move beyond the controversies of retrospective meta-analysis? *Journal of Parapsychology*, 77, 21-35. Retrieved from http://jeksite.org/psi/jp13a.pdf

Kennedy, J. E. (2014). Experimenter misconduct in parapsychology: Analysis manipulation and fraud. Available at http://jeksite.org/psi/misconduct.pdf

Kennedy, J.E. (2015). Beware of inferential errors and low power with Bayesian analyses: Power analysis is needed for confirmatory research. *Journal of Parapsychology*, 79, 53-64. Retrieved from http://jeksite.org/psi/confirm_bayes.pdf

Keppel, G. (1973), *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ: Prentice-Hall.

KPU Registry. (2014). Exploratory and confirmatory analyses. Retrieved from http://www.koestler-parapsychology.psy.ed.ac.uk/Documents/explore_confirm.pdf

Kruschke, J.K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press.

Nosek, B.A., Spies, J.R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615-631. http://dx.doi.org/10.1177/1745691612459058

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*, 657–660. Retrieved from http://pps.sagepub.com/content/7/6/657.full

Radin, D., Nelson, R., Dobyns, Y., & Houtkooper, J. (2006). Reexamining psychokinesis: Comment on Bosch, Steinkamp, and Boller (2006). *Psychological Bulletin, 132*, 529-532.

Schmeidler, G. R., & Edge, H. (1999). Should ganzfeld research continue to be crucial in the search for a replicable psi effect? Part II. Edited ganzfeld debate. *Journal of Parapsychology*, *63*, 335-388.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. http://dx.doi.org/10.1177/0956797611417632

Storm, L. (2000). Research note: Replicable evidence of psi: A revision of Milton's (1999) meta-analysis of ganzfeld databases. *Journal of Parapsychology*, *64*, 411-416.

Storm, L., Tressoldi, P. E., & Di Risio, L. (2010b). A meta-analysis with nothing to hide: Reply to Hyman (2010). *Psychological Bulletin*, *136*, 491-494.

U.S. Food and Drug Administration, (2010). Guidance on the use of Bayesian statistics in medical device clinical Trials. Retrieved from http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf

Wagenmakers, E., Wetzels, R., Borsboom, D., van der Maas, H. J., & Kevit, R. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7,* 632–638. Retrieved from http://pps.sagepub.com/content/7/6/632.full.pdf+html

Other Methodology Articles