PERSPECTIVE
After *p* Values: The New Statistics for Undergraduate Neuroscience Education

Robert J Calin-Jageman
Psychology Department, Dominican University, River Forest, IL 60305.

Statistical inference is a methodological cornerstone for neuroscience education. For many years this has meant inculcating neuroscience majors into null hypothesis significance testing with *p* values. There is increasing concern, however, about the pervasive misuse of *p* values. It is time to start planning statistics curricula for neuroscience majors that replaces or de-emphasizes *p* values. One promising alternative approach is what Cumming has dubbed the "New Statistics", an approach that emphasizes effect sizes, confidence intervals, meta-analysis, and open science. I give an example of the New Statistics in action and describe some of the key benefits of adopting this approach in neuroscience education.

Key words: inferential statistics, neuroscience education, null-hypothesis significance testing, Open Science, confidence intervals

---

This is the first in a new Stats Perspectives Series for JUNE.

Neuroscientists try to discern general principles of nervous system function but can collect only finite sets of data. Thus, inferential statistics serves as a foundation for neuroscience practice and neuroscience education.

It is unsettling to realize that the foundation is shaking. There is increasing caution about our field's reliance on Null Hypothesis Significance Testing (NHST). NHST tests against a null hypothesis that is unlikely to be exactly true in any case, emits *p* values that are routinely misinterpreted, and arbitrarily dichotomizes research results in a way that is surprisingly unreliable (Box 1 lists some key documents in the case against NHST). Recognizing these problems, the American Statistical Association (ASA) recently issued a statement on *p* values and hypothesis testing (Wasserstein and Lazar, 2016). It cautions that "scientific conclusions… should not be based only on whether a p-value passes a specific threshold" (p. 131). Because of the pervasive misuse of *p* values "statisticians often supplement or even replace *p* values with other approaches" (p. 132).

If you're like me, you were never taught any "other approaches" to *p* values. In undergrad and grad school, I was trained to use SPSS, to knowingly discuss the null hypothesis, and to feel appropriately elated if $p < 0.05$. It all seemed fine to me.

My satisfaction with NHST began to crumble when my post-doc advisor introduced me to a long and withering line of criticism against the approach (e.g., Cohen, 1994; Gigerenzer, 1993; Meehl, 1967). Discussing these articles during lab meetings felt exhilarating, transgressive, and sometimes humiliating. The problems were so clear once they were pointed out to me; how had I been so blind?

Although I left my post-doc a bit more clear-eyed about the manifold issues with the NHST approach, I still had no sense of what better approach to use. I began my first teaching appointment helping new students become mesmerized by *p* values. What else could I do?

Fortunately for us and for our science, there are some good answers to that question. The hegemony of the NHST approach is ending in part because excellent alternatives are becoming more known and usable every day (e.g., accessible Bayesian approaches: Kruschke and Liddell, 2017).

It will be some time before the statistical foundations of neuroscience stop shaking. From the proliferation of alternatives, it is difficult to predict which will be widely useful. Hopefully, no one approach will emerge as "the" way to do statistics in the way that NHST has reigned. Statistical pluralism seems essential for a field as broad and diverse as neuroscience. Still, the statement from the American Statistical Association should be enough to banish any lingering complacency with our current approach to statistics education: it is time to start moving away from the NHST approach in the neuroscience curriculum.

If you are willing to heed this call, and an alternative worth exploring is the "New Statistics" (Cumming, 2011; Cumming and Calin-Jageman, 2017). This approach (also known as the estimation approach) is based on four principles:

1. Ask quantitative questions and then make research conclusions that focus on effect sizes.
2. Countenance uncertainty by reporting and interpreting confidence intervals.
3. Seek replication and use meta-analysis as a matter of course.
4. Be open and complete in reporting all analyses, especially in distinguishing planned and exploratory analyses.

Box 1: Key sources in the case against *p* values

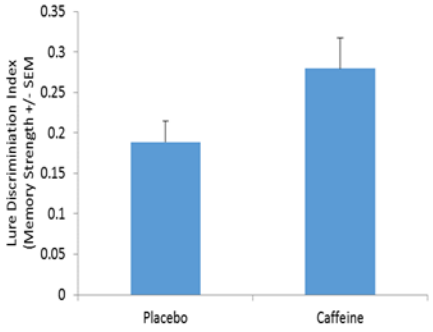
- Cumming G. 2008. Replication and p intervals. *Perspect Psychol Sci* 3: 286–300. PMID: 26158948
- Gigerenzer G. 2004. Mindless statistics. *J Socio Econ* 33: 587–606. DOI: 10.1016/j.socec.2004.09.033
- Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 22: 1359–66. PMID: 22006061
- Szucs D, Ioannidis JPA. 2017. When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Front Hum Neurosci* 11: 390. PMID: 28824397

Box 2: The New Statistics in Action

Borota et al. (2014) examined the influence of caffeine on memory consolidation. Participants studied images of objects and then received either 200mg of caffeine (*n* = 20) or a placebo (*n* = 24). The next day, memory was evaluated. At the request of reviewers, an extension study was conducted (*n* = 14-15/group). The analyses are from data reconstructed from Figure 2C. Figures are first study only.

The NHST approach:

Does caffeine affect memory consolidation? The 200mg group had significantly better memory scores ($t(42) = 2.0$, $p = 0.05$, figure) indicating that caffeine enhanced consolidation. Data for the extension study was combined with the first study. Analysis again showed that performance for the 200mg caffeine condition was higher than for placebo ($t(71) = 2.0$, $p = 0.049$, not shown). We conclude that caffeine enhances memory consolidation.



The research question is qualitative.

No uncertainty is expressed; results are either significant or not. The sampling error expected is not mentioned, though error bars show 1 SEM.

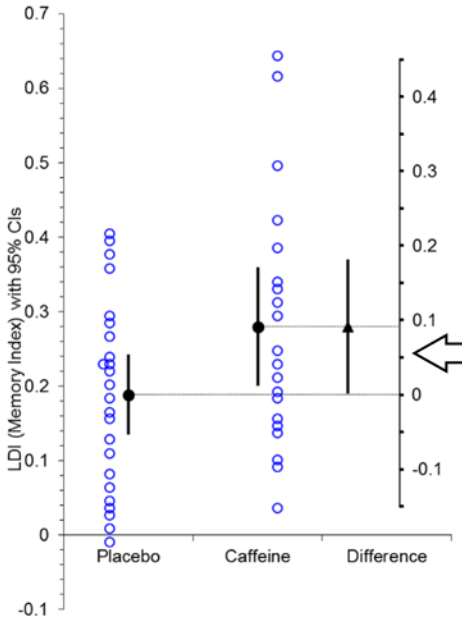
A bar chart does not directly represent the effect of interest, which is the **difference** between the two groups.

The New Statistics Approach:

To what extent does caffeine enhance memory consolidation? The 200mg group scored better than the placebo group by 48.3%* with a margin of error of 48% (95% CI [0.3%, 96%], Figure). This is consistent with caffeine having a very large impact on consolidation, but it is also consistent with caffeine having almost exactly no impact on consolidation. A meta-analysis was used to integrate the results of the first study with the extension study (not shown). This indicated a caffeine-induced improvement of 20% with a margin of error of 21% (95% CI[-1%, 41%]). This CI is too long to enable firm conclusions as it is consistent with a large benefit, but also consistent with no effect or even a very weak impairment. Caffeine is unlikely to strongly impair memory consolidation.

The research question is quantitative.

A 95% CI gives a quantitative expression of uncertainty related to sampling error. Conclusions should consider the whole range of the CI.

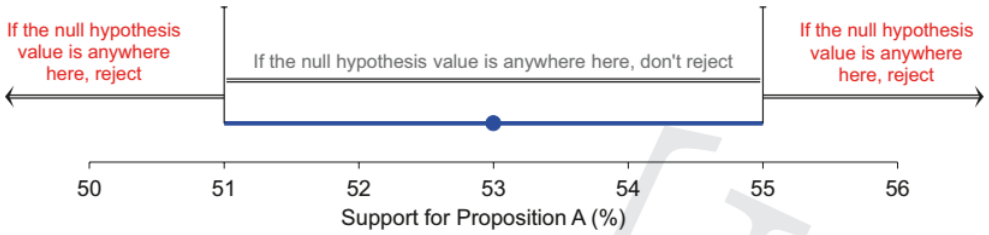


The figure graphically represents the key finding: the difference between the two groups and the uncertainty related to sampling error. Error bars are 95% CIs.

Explanation. The New Stats and NHST can yield different conclusions for any design—the key difference is countenancing uncertainty that NHST hides.

*For this example I report the data as % difference from placebo. It would be better to report raw score differences as well as a standardized effect size measure (e.g. Cohen's d). More on effect sizes in the next installment of the Stats Perspectives Series.

Box 3: Translating between Confidence Intervals and *p* values:



The 95% confidence interval is the collection of all the null hypotheses which are not rejected given alpha = 0.05. In other words, if the null is outside of the confidence interval, $p < 0.05$. Confidence intervals thus provide all the information a *p* value provides and more (but resist the urge to only use confidence intervals to make NHST judgements)

Box 2 gives an example of the New Statistics in action, showing how the same data might be interpreted with the familiar NHST approach and with the New Statistics approach (data are modelled after Borota et al., 2014). The example is one in which the two approaches lead to very different conclusions, much to the credit, I think, of the New Statistics approach.

Some fields have long since moved away from *p* values towards confidence intervals (e.g., International Committee of Medical Journal Editors, 1997). What, then, is "new" about the New Statistics? Really, just the push to make the same transition in the behavioral and life sciences.

Compared to *p* values and the NHST approach, the New Statistics offers several advantages (Cumming and Finch, 2001):

- Confidence intervals are easier to understand so they support better understanding and interpretation. In my experience, students find this approach *much* easier to learn and a higher percentage achieve the skills required to make thoughtful use of empirical data.
- Confidence intervals lend themselves readily to meta-analysis, fostering cumulative science that builds upon and synthesizes previous results.
- Confidence intervals help focus on the precision obtained in a study. It is easy to plan a study to obtain a desired precision and to judge the precision of a study once it is complete.
- The New Statistics is adaptable to different statistical philosophies. Frequentists can calculate and interpret confidence intervals; Bayesians can calculate and interpret credible intervals (Kruschke and Liddell, 2017).

Although these advantages may seem compelling, the challenge of learning and teaching a new approach to statistics may seem too daunting to contemplate. Fortunately, those trained in *p* values can very easily make the transition to the New Statistics. The mathematical foundations are the same for both approaches. For example, the confidence intervals in Box 2 were calculated from the standard error and the critical *t* value for the sample size obtained—that's just a new way to use the same information plugged into a typical *t* test. Because the mathematical foundations are the same, there is a direct link between confidence intervals and null-hypothesis testing (Box 3). It is easy to translate between the two approaches, and students can actually understand *p* values better if they learn about confidence intervals first (Box 3).

Ease of learning the New Statistics does not mean that transitioning your neuroscience curriculum away from *p* values will be easy. Curricula have many interlocking pieces. Changing the approach in your statistics coursework will require revised readings, materials, activities, quizzes, and exams. Often statistics instruction is designed for multiple majors, so advocating for and implementing a change may require building alliances across departments. In addition, changes can reverberate through your curriculum, as adopting the New Statistics in foundational coursework can also require updating upper-level lab assignments, research project rubrics, exit exams, and the like.

Fortunately, there is a growing ecosystem of resources to draw upon to help your program contemplate life after *p* values. This includes a crowd-sourced Open Science Framework project that collects resources to help instructors get started with the New Statistics (https://osf.io/muy6u/wiki/home/). The project's list of software resources is especially useful for getting started with the New Statistics. When it comes to publishing, don't stress. Editors and professional organizations are becoming familiar with effect sizes and confidence intervals and are often requiring this new approach to reporting results. When you submit manuscripts with your students, you can include *p* values as a supplement and/or mention that statistical significance can be determined by inspecting the confidence intervals reported (here are two examples from my lab: Herdegen et al., 2014; Conte et al., 2017).

The New Statistics is not a pancea. As with *p* values, students can stubbornly hold on to misconceptions about confidence intervals (Hoekstra et al., 2014). Moreover, even seasoned researchers can fall into the trap of not really interpreting confidence intervals, but instead using them as proxies for *p* values (Fidler et al., 2004). Still, we shouldn't let the perfect be the enemy of the good. Given the manifold and pervasive misuse of *p* values (Cumming, 2008; Simmons et al., 2011; Szucs and Ioannidis, 2017) our neuroscience majors will be better served by a statistics curriculum that includes or even focuses on alternative approaches. Are you ready to get started?

REFERENCES

Borota D, Murray E, Keceli G, Chang A, Watabe JM, Ly M, Toscano JP, Yassa MA (2014) Post-study caffeine administration enhances memory consolidation in humans. Nat Neurosci 17:201–203. http://www.ncbi.nlm.nih.gov/pubmed/24413697 (Accessed January 21, 2014).

Cohen J (1994) The earth is round (p < .05). Am Psychol 49:997–1003. http://doi.apa.org/getdoi.cfm?doi=10.1037/0003-066X.49.12.997.

Conte C, Herdegen S, Kamal S, Patel J, Patel U, Perez L, Rivota M, Calin-Jageman RJ, Calin-Jageman IE (2017) Transcriptional correlates of memory maintenance following long-term sensitization of Aplysia californica. Learn Mem 24:502–515. http://www.ncbi.nlm.nih.gov/pubmed/28916625.

Cumming G (2008) Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Perspect Psychol Sci 3:286–300.

Cumming G (2011) Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. New York: Routledge.

Cumming G, Calin-Jageman RJ (2017) Introduction to the new statistics: Estimation, open science, and beyond. New York: Routledge. http://thenewstatistics.com/itns/.

Cumming G, Finch SUE (2001) Four reasons to use CIs. Educ Psychol Meas 61:532–574.

Fidler F, Thomason N, Cumming G, Finch S, Leeman J (2004) Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine. Psychol Sci 15:119–126. http://pss.sagepub.com/content/15/2/119.short.

Gigerenzer G (1993) The superego, the ego, and the id in statistical reasoning. In Handboo for data analysis in the behavioral sciences: methodological issues (Keren G and Lewis C, eds), pp 311–339, Hillsdale, NJ: Erlbaum.

Gigerenzer G (2004) Mindless statistics. J Socio Econ 33:587–606. DOI: 10.1016/j.socec.2004.09.033.

Herdegen S, Holmes G, Cyriac A, Calin-Jageman IE, Calin-Jageman RJ (2014) Characterization of the rapid transcriptional response to long-term sensitization training in Aplysia californica. Neurobiol Learn Mem 116:27–35. http://linkinghub.elsevier.com/retrieve/pii/S1074742714001361.

Hoekstra R, Morey RD, Rouder JN, Wagenmakers E-J (2014) Robust misinterpretation of confidence intervals. Psychon Bull Rev 21:1157–1164. http://www.ncbi.nlm.nih.gov/pubmed/24420726 (Accessed July 14, 2014).

International Committee of Medical Journal Editors (1997) Uniform Requirements for Manuscripts Submitted to Biomedical Journals. N Engl J Med 336:309–316. http://www.nejm.org/doi/abs/10.1056/NEJM199701233360422.

Kruschke JK, Liddell TM (2017) The Bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychon Bull Rev. http://link.springer.com/10.3758/s13423-016-1221-4.

Meehl PE (1967) Theory-testing in psychology and physics: a methodological paradox. Philos Sci 34:103–115. http://www.journals.uchicago.edu/doi/10.1086/288135.

Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 22:1359–66. http://pss.sagepub.com/lookup/doi/10.1177/0956797611417632 (Accessed March 19, 2014).

Szucs D, Ioannidis JPA (2017) When null hypothesis significance testing is unsuitable for research: a reassessment. Front Hum Neurosci 11:390. http://journal.frontiersin.org/article/10.3389/fnhum.2017.00390/full.

Wasserstein RL, Lazar NA (2016) The ASA's statement on p -values: context, process, and purpose. Am Stat 70:129–133. https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108.

Potential Conflict of Interest Statement: Dr. Robert J Calin-Jageman is a co-author of a textbook that teaches the New Statistics Approach.

Received August 28, 2017; revised August 28, 2017; accepted October 02, 2017.

Thank you to Geoff Cumming inviting me to join the crusade and for so much helpful mentorship along the way.

Address correspondence to: Dr. Robert J Calin-Jageman, Psychology Department, 7900 West Division, River Forest, IL 60305. Email: rcalinjageman@dom.edu.