Estimation statistics should replace significance testing

Adam Claridge-Chang[1, 2, 5] and Pryseley N. Assam[3, 4]

1. Program in Neuroscience and Behavioral Disorders, Duke-NUS Medical School, Singapore 138673
2. Institute for Molecular and Cell Biology, Singapore 138673
3. Singapore Clinical Research Institute, Singapore 138669
4. Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore 169857
5. Department of Physiology, National University of Singapore, Singapore 138673

Correspondence: claridge-chang.adam@duke-nus.edu.sg

FOR OVER FORTY YEARS, null hypothesis significance testing and P values have been questioned by statistical commentators, their utility criticized on philosophical and practical grounds (Halsey *et al.* 2015; Cohen 1994). Luckily, the preferred statistical methodology is accessible with modest re-training. An obstacle to the adoption of this alternative is a basic branding problem: it does not have a widely-used name. We suggest the most appropriate name for this superior approach is 'estimation statistics,' a term describing the methods that focus on the estimation of effect sizes (point estimates) and their confidence intervals (precision estimates). Estimation statistics offers several key benefits.

Estimation gives a more informative way to analyze and interpret results. For example, for an experiment with two independent groups the estimation counterpart to a t-test is to calculate the mean difference and its confidence interval (Cumming 2012). The mean difference (MD) is calculated as one mean minus the other, while its confidence interval falls between MD - (1.96 × SEMD) and MD + (1.96 × SEMD), where SEMD is the pooled standard error of MD (Altman *et al.* 2000). Knowing and thinking about the magnitude and precision of an effect is more useful to quantitative science than contemplating the probability of observing data of at least that extremity, assuming absolutely no effect. An old joke about the science of metal springs says that estimation reveals the proportionality between force and extension (Hooke's law), while P just tells you "*when you pull on it, it gets longer*" (Tukey 1969). Medical research has led the way in adopting estimation statistics. Using the effect size in the clinical context rightfully places the focus on the magnitude of a treatment's benefit. In basic research, using effect sizes facilitates quantitative comparisons and models (like Hooke's law of elasticity), and—importantly—encourages data analysts to think about the metrics they use and how they relate to the natural processes under study.

The second key benefit of estimation statistics is synthesizing data from published sources using the methods of systematic review and meta-analysis. Meta-analysis is a way to average effect sizes from different studies, a method that produces a more precise overall estimate (a narrower confidence interval). In medical research, meta-analytic studies of randomized controlled clinical trials are considered the strongest form of medical evidence: they are used to reconcile discordant results, produce precise estimates of treatment effects, identify knowledge gaps, guide clinical practice and inform further investigation. Meta-analyses are published in numerous medical journals (*e.g.*, *The Lancet, The BMJ*) as a quantitative alternative to the conventional 'he said/she said' narrative review. Meta-analytic studies are also now being used in preclinical research (Vesterinen *et al.* 2014), for example a recent study showing that the animal model literature on stroke overstates efficacy (Sena *et al.* 2010).

Estimation statistics' third important benefit is its use of model construction to quantify trends in either primary or published data. Models can be basic or more advanced, such as multivariate meta-regression, a method that accounts for sources of experimental heterogeneity in complex data. Like clinical data, basic research results are well-suited to the use of multivariate models to analyze both primary and pooled published data from complex experimental designs, especially when the data have high integrity (Yildizoglu *et al.* 2015).

Estimation statistics use remains rare in basic research. This situation may improve with increasing awareness of the limitations of significance testing. Estimation is a comprehensive analysis framework that offers a better way to interpret data, quantitative literature review methods and techniques to analyze heterogeneous data; it can be readily adopted as a replacement for significance testing.

This letter was previously published in NATURE METHODS.

QUESTIONS FOR THE AUTHORS

Why don't you criticize significance testing more?

Criticism of *P* values has a rich, long history dating back to before John Tukey's 1969 mockery of the technique (Tukey 1969; Morrison and Henkel 2006). The aim here was not to attempt to reiterate all these arguments, but to identify and define estimation statistics at a time when this is urgently needed. We hope to 'brand' estimation statistics and to define this framework as including the reporting of effect sizes as well as systematic review, meta-analysis and mixed models.

Meta-analysis can produce a *P* value as well, why do you consider meta-analysis to be an estimation method?

Meta-analysis is entirely dependent on estimation: a meta-analyst extracts effect sizes and their precision, and then performs weighted averaging to calculate a synthetic effect size. Individual *P* values are irrelevant to this process. It is true that at the end of this process one can calculate a *P* value for that meta-analytic effect size, but in the view of the estimation school this is beside the point. The main benefit of meta-analysis is the acquisition of a precise estimate, which can be used to inform cost-benefit analysis (for clinical interventions) or quantitative models of the biological mechanism (for biological experiments).

How will using estimation statistics solve the problem of poor reproducibility in the literature?

The estimation school views the false dichotomization of significance testing into true/false states as one of the major problems with this method. Presenting results in dichotomous terms can accentuate discordance while failing to deal with the continuous nature of data. Establishing estimation statistics as an identifiable methodological framework is one step on the way to grappling with these other issues.

How might estimation help quantitative modeling of biological systems?

Consider the computational biologist of the future, extracting data from today's literature so as to define parameters in a mathematical model of a molecular pathway. She is not aiming to extract or use *P* values in the model, but is hoping to find estimates of each molecule's role in the pathway. If she is performing meta-analysis, she will be using the effect sizes. She will not be calculating the probability of observing a particular effect size or greater under the assumption that the molecule plays absolutely no role in the pathway, and whether that probability is above or below some arbitrary threshold.

Estimation might have relevance to clinical research, which measures real-world patient outcomes with practical relevance. How would this apply to basic research in which metrics are often abstract or complex?

One of the important benefits of using estimation for interpretation it that it encourages researchers to think about the quantitative and theoretical importance of experimental findings. Currently, almost all research articles already contain metrics that could be readily employed as effect sizes, but are often ignored in the rush to calculate *P* values. For example, in the memory genetics field, assay metrics are routinely reported as quantitative values in bar charts, *e.g.* mutant and wild type memory scores.

Ideally, researchers would report, discuss and interpret an intervention's effect size (*e.g.* wild type score minus mutant score). Instead, researchers report and discuss the probability of seeing such a large contrast or greater assuming that there is really no difference between the two and whether that probability falls above or below an arbitrary threshold. Relating the memory contrast values to a quantitative model of memory in many cases never occurs because significance testing tends to limit interpretation to simply whether a mutation has an effect or not (false dichotomization).

If currently used metrics and effect sizes are not optimal, then this needs to be discussed and improved, but the use of yes/no tests impedes this. Without a shift to estimation, discussions of the theoretical or practical relevance of experimental effect sizes are unlikely to be initiated in the many fields currently using significance testing.

Why am I only hearing about estimation statistics now?

Previous branding attempts have so far failed to catch on. The importance of branding at this moment is apparent as the refutation of *P* values has left confusion in the minds of many scientists about what can replace significance testing. One editorial maintains that biologists should retain significance testing, but adopt a greatly smaller alpha (Horton 2015). Another article mourns the loss of *P* values and proposes that the alternative is Bayesian statistics, a methodology that would require substantial retraining for most (Nuzzo 2015). We believe that estimation provides the ideal upgrade for the majority of researchers' analysis needs.

Why can't *P* values be used in combination with estimation statistics?

We suggest that scientists use estimation for data analysis, interpretation and discussion. If—despite the overwhelming case against significance testing—reviewers and/or editors demand *P* values from you, these can be provided *pro forma* but either not referred to, or referred to as little as possible in the text, thus minimizing the damage.

REFERENCES

Altman, D., D. Machin, T. Bryant, and S. Gardner. 2000. *Statistics with Confidence: Confidence Interval and Statistical Guidelines.* Bristol: BMJ Books.

Cohen, Jacob. 1994. "The Earth Is Round (p < .05)." *The American Psychologist* 49 (12): 997–1003.

Cumming, Geoff. 2012. *Understanding the New Statistics Effect Sizes, Confidence Intervals, and Meta-Analysis.* New York: Routledge.

Halsey, Lewis G., Douglas Curran-Everett, Sarah L. Vowler, and Gordon B. Drummond. 2015. "The Fickle *P* Value Generates Irreproducible Results." *Nature Methods* 12 (3): 179–85.

Horton, Richard. 2015. "Offline: What Is Medicine's 5 Sigma?" *The Lancet* 385 (9976). Elsevier: 1380.

Morrison, Denton E., and Ramon E. Henkel. 2006. *The Significance Test Controversy: A Reader.* AldineTransaction.

Nuzzo, Regina. 2015. "Scientists Perturbed by Loss of Stat Tools to Sift Research Fudge from Fact." *Scientific American*, April 16.

Sena, Emily S., H. Bart van der Worp, Philip M. W. Bath, David W. Howells, and Malcolm R. Macleod. 2010. "Publication Bias in Reports of Animal Stroke Studies Leads to Major Overstatement of Efficacy." *PLoS Biology* 8 (3): e1000344.

Tukey, John W. 1969. "Analyzing Data: Sanctification or Detective Work?" *The American Psychologist* 24 (2). American Psychological Association: 83.

Vesterinen, H. M., E. S. Sena, K. J. Egan, T. C. Hirst, L. Churolov, G. L. Currie, A. Antonic, D. W. Howells, and M. R. Macleod. 2014. "Meta-Analysis of Data from Animal Studies: A Practical Guide." *Journal of Neuroscience Methods* 221 (January): 92–102.

Yildizoglu, Tugce, Jan-Marek Weislogel, Farhan Mohammad, Edwin S-Y Chan, Pryseley N. Assam, and Adam Claridge-Chang. 2015. "Estimating Information Processing in a Memory System: The Utility of Meta-Analytic Methods for Genetics." *PLoS Genetics* 11 (12): e1005718.