P values are just the tip of the iceberg

Ridding science of shoddy statistics will require scrutiny of every step, not merely the last one, say **Jeffrey T. Leek** and **Roger D. Peng**.

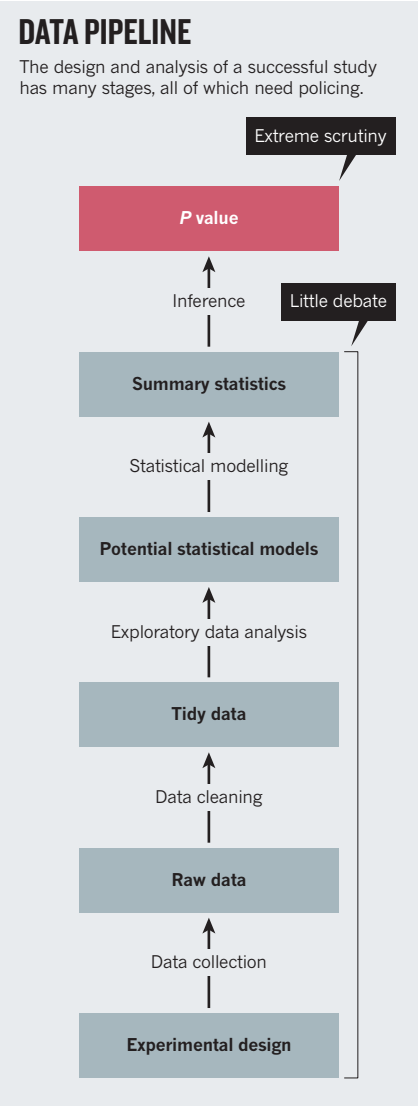
There is no statistic more maligned than the *P* value. Hundreds of papers and blogposts have been written about what some statisticians deride as 'null hypothesis significance testing' (NHST; see, for example, go.nature.com/pfvgqe). NHST deems whether the results of a data analysis are important on the basis of whether a summary statistic (such as a *P* value) has crossed a threshold. Given the discourse, it is no surprise that some hailed as a victory the banning of NHST methods (and all of statistical inference) in the journal *Basic and Applied Social Psychology* in February[1].

Such a ban will in fact have scant effect on the quality of published science. There are many stages to the design and analysis of a successful study (see 'Data pipeline'). The last of these steps is the calculation of an inferential statistic such as a *P* value, and the application of a 'decision rule' to it (for example, $P < 0.05$). In practice, decisions that are made earlier in data analysis have a much greater impact on results — from experimental design to batch effects, lack of adjustment for confounding factors, or simple measurement error. Arbitrary levels of statistical significance can be achieved by changing the ways in which data are cleaned, summarized or modelled[2].

P values are an easy target: being widely used, they are widely abused. But, in practice, deregulating statistical significance opens the door to even more ways to game statistics — intentionally or unintentionally — to get a result. Replacing *P* values with Bayes factors or another statistic is ultimately about choosing a different trade-off of true positives and false positives. Arguing about the *P* value is like focusing on a single misspelling, rather than on the faulty logic of a sentence.

Better education is a start. Just as anyone who does DNA sequencing or remote-sensing has to be trained to use a machine, so too anyone who analyses data must be trained in the relevant software and concepts. Even investigators who supervise data analysis should be required by their funding agencies and institutions to complete training in understanding the outputs and potential problems with an analysis.

There are online courses specifically designed to address this crisis. For example, the Data Science Specialization, offered by Johns Hopkins University in Baltimore, Maryland, and Data Carpentry, can easily be integrated into training and research. It is increasingly possible to learn to use the computing tools relevant to specific disciplines — training in Bioconductor, Galaxy and Python is included in Johns Hopkins' Genomic Data Science Specialization, for instance.



But education is not enough. Data analysis is taught through an apprenticeship model, and different disciplines develop their own analysis subcultures. Decisions are based on cultural conventions in specific communities rather than on empirical evidence. For example, economists call data measured over time 'panel data', to which they frequently apply mixed-effects models. Biomedical scientists refer to the same type of data structure as 'longitudinal data', and often go at it with generalized estimating equations.

Statistical research largely focuses on mathematical statistics, to the exclusion of the behaviour and processes involved in data analysis. To solve this deeper problem, we must study how people perform data analysis in the real world. What sets them up for success, and what for failure? Controlled experiments have been done in visualization[3] and risk interpretation[4] to evaluate how humans perceive and interact with data and statistics. More recently, we and others have been studying the entire analysis pipeline. We found, for example, that recently trained data analysts do not know how to infer *P* values from plots of data[5], but they can learn to do so with practice.

The ultimate goal is evidence-based data analysis[6]. This is analogous to evidence-based medicine, in which physicians are encouraged to use only treatments for which efficacy has been proved in controlled trials. Statisticians and the people they teach and collaborate with need to stop arguing about *P* values, and prevent the rest of the iceberg from sinking science. ■

Jeffrey T. Leek *and* **Roger D. Peng** *are associate professors of biostatistics at the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, USA.*
e-mail: jleek@jhsph.edu

1. Trafimow, D. & Marks, M. *Basic Appl. Soc. Psych.* **37,** 1–2 (2015).
2. Simmons, J. P., Nelson, L. D. & Simonsohn, U. *Psychol. Sci.* **22,** 1359–1366 (2011).
3. Cleveland, W. S. & McGill, R. *Science* **229,** 828–833 (1985).
4. Kahneman, D. & Tversky, A. *Econometrica* **47,** 263–291 (1979).
5. Fisher, A., Anderson, G. B., Peng, R. & Leek, J. *PeerJ* **2,** e589 (2014).
6. Leek, J. T. & Peng, R. D. *Proc. Natl Acad. Sci. USA* **112,** 1645–1646 (2015).