# Some natural solutions to the p-value communication problem—and why they won't work*

Andrew Gelman† and John Carlin‡

26 Feb 2017

## 1. Significance testing in crisis

It is well known that even experienced scientists routinely misinterpret $p$-values in all sorts of ways, including confusion of statistical and practical significance, treating non-rejection as acceptance of the null hypothesis, and interpreting the $p$-value as some sort of replication probability or as the posterior probability that the null hypothesis is true.

A common conceptual error is that researchers take the rejection of a straw-man null as evidence in favor of their preferred alternative (Gelman, 2014). A standard mode of operation goes like this: $p < 0.05$ is taken as strong evidence against the null hypothesis, $p > 0.15$ is taken as evidence in favor of the null, and $p$ near 0.10 is taken either as weak evidence for an effect or as evidence of a weak effect.

Unfortunately, none of those inferences is generally appropriate: a low $p$-value is not necessarily strong evidence against the null (see, for example, Morris, 1987, and Gelman and Carlin 2014), a high $p$-value does not necessarily favor the null (the strength and even the direction of the evidence depends on the alternative hypotheses), and $p$-values are in general not measures of the size of any underlying effect. But these errors persist, reflecting (a) inherent difficulties in the mathematics and logic of $p$-values, and (b) the desire of researchers to draw strong conclusions from their data.

Continued evidence of these and other misconceptions and their dire consequences for science (the "replication crisis" in psychology, biology, and other applied fields), especially in light of new understanding of how common it is that abundant "researcher degrees of freedom" (Simmons, Nelson, and Simonsohn, 2011) and "gardens of forking paths" (Gelman and Loken, 2014) allow researchers to routinely obtain statistically significant and publishable results from noise, motivated the American Statistical Association to release a Statement on Statistical Significance and $p$-values in an attempt to highlight the magnitude and importance of problems with current standard practice (Wasserstein and Lazar, 2016).

At this point it would be natural for statisticians to think that this is a problem of education and communication. If we could just add a few more paragraphs to the relevant sections of our textbooks, and persuade applied practitioners to consult more with statisticians, then all would be well, or so goes this logic.

In their new paper, McShane and Gal present survey data showing that even authors of published articles in a top statistics journal are often confused about the meaning of $p$-values, especially by treating 0.05, or the range 0.05–0.15, as the location of a threshold. The underlying problem seems to be deterministic thinking. To put it another way, applied researchers and also statisticians are in the habit of demanding more certainty than their data can legitimately supply. The problem is not just that 0.05 is an arbitrary convention; rather, even a seemingly wide range of $p$-values such as 0.01–0.10 cannot serve to classify evidence in the desired way (Gelman and Stern, 2006).

*Discussion of "Statistical significance and the dichotomization of evidence," by Blakeley McShane and David Gal. To appear in the *Journal of the American Statistical Association.*

†Department of Statistics and Department of Political Science, Columbia University, New York

‡Clinical Epidemiology and Biostatistics, Murdoch Children's Research Institute, and Centre for Epidemiology & Biostatistics, University of Melbourne, Parkville, Victoria, Australia

It is shocking that these errors seem so hard-wired into statisticians' thinking, and this suggests that our profession really needs to look at how it teaches the interpretation of statistical inferences. The problem does not seem just to be technical misunderstandings; rather, statistical analysis is being asked to do something that it simply can't do, to bring out a signal from any data, no matter how noisy. We suspect that, to make progress in pedagogy, statisticians will have to give up some of the claims we have implicitly been making about the effectiveness of our methods.

## 2. Some natural solutions that won't, on their own, work

### 2.1. Listen to the statisticians, or clarity in exposition

It would be nice if the statistics profession was offering a good solution to the significance testing problem and we just needed to convey it more clearly. But, no, as McShane and Gal reveal, many statisticians misunderstand the core ideas too. It might be a good idea for other reasons to recommend that students take more statistics classes—but this won't solve the problems if textbooks point in the wrong direction and instructors don't understand what they are teaching. To put it another way, it's not that we're teaching the right thing poorly; unfortunately, we've been teaching the wrong thing all too well.

This is one of the difficulties we had with the American Statistical Association's statement on $p$-values: the statistics profession has been spending decades selling people on the idea of statistics as a tool for extracting signal from noise, and our journals and textbooks are full of triumphant examples of learning through statistical significance; so it's not clear why we as a profession should be trusted going forward, at least not until we take some responsibility for the mess we've helped to create.

### 2.2. Confidence intervals instead of hypothesis tests

A standard use of a confidence interval is to check whether it excludes zero. In this case it's a hypothesis test under another name.

Another use is to consider the interval as a statement about uncertainty in a parameter estimate. But this can give nonsensical answers, not just in weird trick problems but for real applications. For example, Griskevus et al. (2014) use data from a small survey to estimate that single women were 20 percentage points more likely to vote for Barack Obama during certain days in their monthly cycle. This estimate was statistically significant with a reported standard error of 8 percentage points; thus the classical 95% interval for the effect size was (4%, 36%), an interval that makes no sense on either end! Even an effect of 4% is implausible given what we know about the low rate of opinion change during presidential election campaigns (e.g., Gelman et al., 2016)—and it would certainly be a mistake to use this survey to rule out zero or small negative net effects.

So, although confidence intervals contain some information beyond that in $p$-values, they do not resolve the larger problems that arise from attempting to get near-certainty out of noisy estimates.

### 2.3. Bayesian interpretation of one-sided $p$-values

Consider a parameter estimate that is greater than zero and whose statistical significance is being assessed using a $p$-value. Under a symmetric continuous model such as the normal distribution, the one-sided $p$-value or tail-area probability is identical to the posterior probability that the parameter of interest is negative, given the data and a uniform prior distribution. This mathematical identity has led Greenland and Poole (2013) to suggest that "P values can be incorporated into a modern analysis framework that emphasizes measurement of fit, distance, and posterior probability in place of 'statistical significance' and accept/reject decisions." We agree with that last bit about

moving away from binary decisions but we don't think the Bayesian interpretation of the $p$-value is particularly helpful except as some sort of bound.

The problem comes with the uniform prior distribution. We tend to be most concerned with overinterpretation of statistical significance in problems where underlying effects are small and variation is high, and in these settings the use of classical inferences—or their ~~flat-prior Bayesian equivalents~~—will lead to systematic overestimation of effect sizes and over-certainty regarding their signs: high type M and type S errors, in the terminology of Gelman and Tuerlinckx (2000) and Gelman and Carlin (2014). We do *not* consider it reasonable in general to interpret a $z$-statistic of 1.96 as implying a 97.5% chance that the corresponding estimate is in the right direction.

### 2.4. Focusing on "practical significance" instead of "statistical significance"

Realistically, all statistical hypotheses are false: effects are not exactly zero, groups are not exactly identical, distributions are not really normal, measurements are not quite unbiased, and so on. Thus, with enough data it should be possible to reject any hypothesis. It's a commonplace among statisticians that a $\chi^2$ test (and, really, any $p$-value) can be viewed as a crude measure of sample size, and this can be framed as the distinction between practical and statistical significance, as can be illustrated with a hypothetical large study in which an anti-hypertension drug is found to reduce blood pressure by 0.3 mmHg with a standard error of 0.1. This estimate is clearly statistically significantly different from zero but is tiny on a substantive scale.

So, in a huge study, comparisons can be statistically significant without having any practical importance. Or, as we would prefer to put it, effects can vary: a $+0.3$ for one group in one scenario might become $-0.2$ for a different group in a different situation. Tiny effects are not only possibly trivial, they can also be unstable, so that for future purposes an estimate of $0.3 \pm 0.1$ might not even be so likely to remain positive. To put it another way, the characterization of an effect as "small" or "not of practical significance" is relative to some prior understanding of underlying variation.

That said, the distinction between practical and statistical significance does *not* resolve the difficulties with $p$-values. The problem is not so much with large samples and tiny but precisely-measured effects but rather with the opposite: large effect-size estimates that are hopelessly contaminated with noise. Consider an estimate of 30 with standard error 10, of an underlying effect that cannot realistically be much larger than 1. In this case the estimate is statistically significant and also practically significant but is essentially entirely the product of noise. This problem is central to the recent replication crisis in science (see Button et al., 2013, and Loken and Gelman, 2017) but is not at all touched by concerns of practical significance.

### 2.5. Bayes factors

Another direction for reform is to preserve the idea of hypothesis testing but to abandon tail-area probabilities ($p$-values) and instead summarize inference by the posterior probabilities of the null and alternative models, a method associated with Jeffreys (1961) and discussed recently by Rouder et al. (2009). The difficulty of this approach is that the marginal likelihoods of the separate models (and thus the Bayes factor and the corresponding posterior probabilities) depend crucially on aspects of the prior distribution that are typically assigned in a completely arbitrary manner by users. For example, consider a problem where a parameter has been assigned a normal prior distribution with center 0 and scale 10, and where its estimate is likely to be in the range $(-1, 1)$. The chosen prior is then essentially flat, as would also be the case if the scale were increased to 100 or 1000. But such a change would divide the Bayes factor by 10 or 100.

Beyond this technical criticism, which is explored further by Gelman and Rubin (1995) and Gelman et al. (2013, chapter 8), the use of Bayes factors for hypothesis testing is also subject to

many of the problems of $p$-values when used for that same purpose and which are discussed by McShane and Gal: the temptation to discretize continuous evidence and to declare victory from the rejection of a point null hypothesis that in most cases cannot possibly be true.

## 3. Where next?

Our own preferred replacement for hypothesis testing and $p$-values is model expansion and Bayesian inference, addressing concerns of multiple comparisons using hierarchical modeling (Gelman, Hill, and Yajima, 2013) or through non-Bayesian regularization techniques such as lasso (Lockhart et al., 2013). The general idea is to use Bayesian or regularized inference as a replacement of hypothesis tests but in the manner of Kruschke (2013), through estimation of continuous parameters rather than by trying to assess the probability of a point null hypothesis. And, as we discuss in Sections 2.2–2.4 above, informative priors can be crucial in getting this to work. Indeed, in many contexts it is the prior information rather than the Bayesian machinery that is the most important. Non-Bayesian methods can also incorporate prior information in the form of postulated effect sizes in post-data design calculations (Gelman and Carlin, 2014).

In short, we'd prefer to avoid hypothesis testing entirely and just perform inference using larger, more informative models.

To stop there, though, would be to deny one of the central goals of statistical science. As Morey et al. (2012) write, "Scientific research is often driven by theories that unify diverse observations and make clear predictions. ... Testing a theory requires testing hypotheses that are consequences of the theory, but unfortunately, this is not as simple as looking at the data to see whether they are consistent with the theory." To put it in other words, there is a demand for hypothesis testing. We can shout till our throats are sore that rejection of the null should not imply the acceptance of the alternative, but acceptance of the alternative is what many people want to hear. There is a larger problem of statistical pedagogy associating very specific statistical "hypotheses" with scientific hypotheses and theories, which are nearly always open-ended.

As we wrote in response to the ASA's much-publicized statement from last year, we think the solution is not to reform $p$-values or to replace them with some other statistical summary or threshold, but rather to move toward a greater acceptance of uncertainty and embracing of variation (Carlin, 2016, Gelman, 2016).

We will end not on this grand vision but on an emphasis on some small steps that we hope can make a difference. If we do the little things right, those of us who write textbooks can then convey some of this sensibility into our writings.

To start with, we recommend saying No to binary conclusions in our collaboration and consulting projects: resist giving clean answers when that is not warranted by the data. Instead, do the work to present statistical conclusions with uncertainty rather than as dichotomies. Also, remember that most effects can't be zero (at least in social science and public health), and that an "effect" is usually a mean in a population (or something similar such as a regression coefficient)—a fact that seems to be lost from consciousness when researchers slip into binary statements about there being "an effect" or "no effect" as if they are writing about constants of nature. Again, it will be difficult to resolve the many problems with $p$-values and "statistical significance" without addressing the mistaken goal of certainty which such methods have been used to pursue.

## References

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of

neuroscience. *Nature Reviews Neuroscience* **14**, 365-–376.

Carlin, J. B. (2016). Is reform possible without a paradigm shift? *American Statistician* online. http://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108

Gelman, A. (2012). P-values and statistical practice. *Epidemiology* **24**, 69–72.

Gelman, A. (2014). Confirmationist and falsificationist paradigms of science. Statistical Modeling, Causal Inference, and Social Science blog, 5 Sept. http://andrewgelman.com/2014/09/05/confirmationist-falsificationist-paradigms-science/

Gelman, A. (2016). The problems with p-values are not just with p-values. *American Statistician* online. http://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108

Gelman, A., and Carlin, J. B. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science* **9**, 641–651.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*, third edition. London: Chapman and Hall.

Gelman, A., Goel, S., Rivers, D., and Rothschild, D. (2016). The mythical swing voter. *Quarterly Journal of Political Science* **11**, 103-–130.

Gelman, A., Hill, J., and Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness* **5**, 189–211.

Gelman, A., and Loken, E. (2014). The statistical crisis in science. *American Scientist* **102**, 460–465.

Gelman, A., and Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology 1995*, 165–173.

Greenland, S., and Poole, C. (2013). Living with $P$-values: Resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology* **24**, 62–68.

Jeffreys, H. (1961). *Theory of Probability*, third edition. Oxford University Press.

Kruschke, J. K. (2013). Bayesian estimation supersedes the $t$ test. *Journal of Experimental Psychology: General* **142**, 573–603.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2013). A significance test for the lasso. Technical report, Department of Statistics, Stanford University.

Loken, E., and Gelman, A. (2017). Measurement error and the replication crisis. *Science* **355**, 584–585.

Morey, R. D., Rouder, J. N., Verhagen, J., and Wagenmakers, E. J. (2014). Why hypothesis tests are essential for psychological science. *Psychological Science* **25**, 1289–1290.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., and Wagenmakers, E. J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review* **23**, 103–123.

Morris, C. N. (1987). Comment. *Journal of the American Statistical Association* **82**, 131–133.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian $t$ tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review* **16**, 225–237.

Simmons, J., Nelson, L., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science* **22**, 1359–1366.

Tversky, A., and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin* **76**, 105—110.

Wasserstein, R. L., and Lazar, N. A. (2016). The ASA's Statement on $p$-values: Context, process, and purpose. *American Statistician* **70**, 129–133.