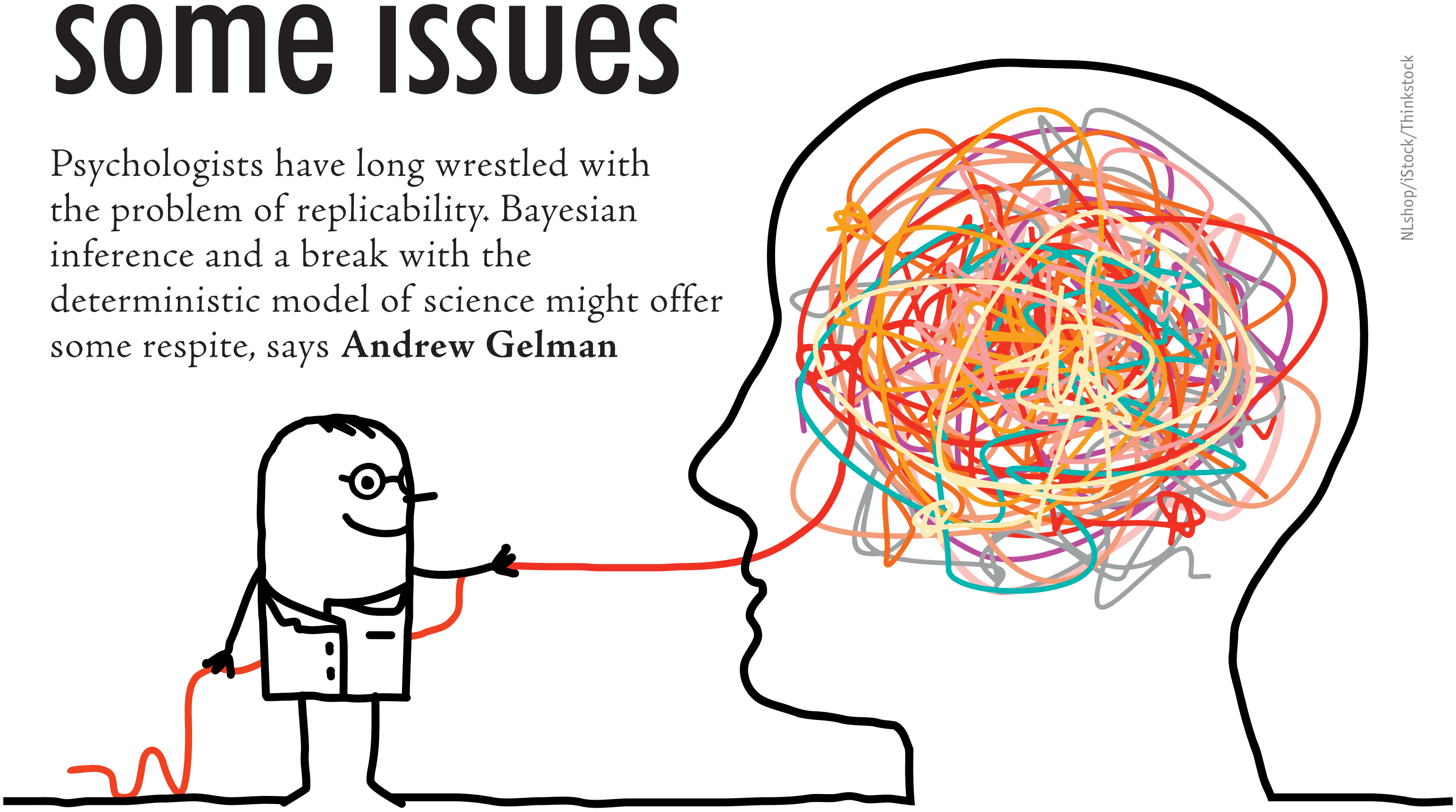
Working through some issues

Psychologists have long wrestled with the problem of replicability. Bayesian inference and a break with the deterministic model of science might offer some respite, says **Andrew Gelman**



NLshop/iStock/Thinkstock

Top journals in psychology routinely publish ridiculous, scientifically implausible claims, justified based on "p < 0.05". Recent examples of such silliness include claimed evidence of extra-sensory perception (published in the *Journal of Personality and Social Psychology*), claims that women at certain stages of their menstrual cycle were three times more likely to wear red or pink clothing and 20 percentage points more likely to vote for the Democratic or Republican candidate for president (published in *Psychological Science*), and a claim that people react differently to hurricanes with male and female names (published in the *Proceedings of the National Academy of Sciences*).

All these studies had serious flaws, to the extent that I (and others) found the claims to be completely unconvincing from a statistical standpoint, matching their general implausibility on substantive grounds.

It is easy to dismiss these particular studies, one at a time. But, to the extent that they are being conducted using standard statistical methods, this calls into question all sorts of more plausible, but not necessarily true, claims – claims that are supported by this same sort of evidence. To put it another way: we can all laugh at studies of ESP, or ovulation and voting, but what about MRI studies of political attitudes, or embodied cognition, or stereotype threat, or, for that matter, the latest potential cancer cure? If we cannot trust p-values, does experimental science involving human variation just have to start over?

Figure 1 (page 34) demonstrates what can happen with classical hypothesis testing. A study is performed in which the underlying parameter of interest (typically a causal effect or some other sort of comparison in the general population) is relatively small, and measurements are noisy and biased (not uncommon in a psychology setting in which the underlying constructs are often not clearly defined).

The particular example we were considering when constructing this graph is a published study claiming that, in the 2012 US presidential election, "Ovulation led single women to become more liberal, less religious, and more likely to vote for Barack Obama. In contrast, ovulation led married women to become more

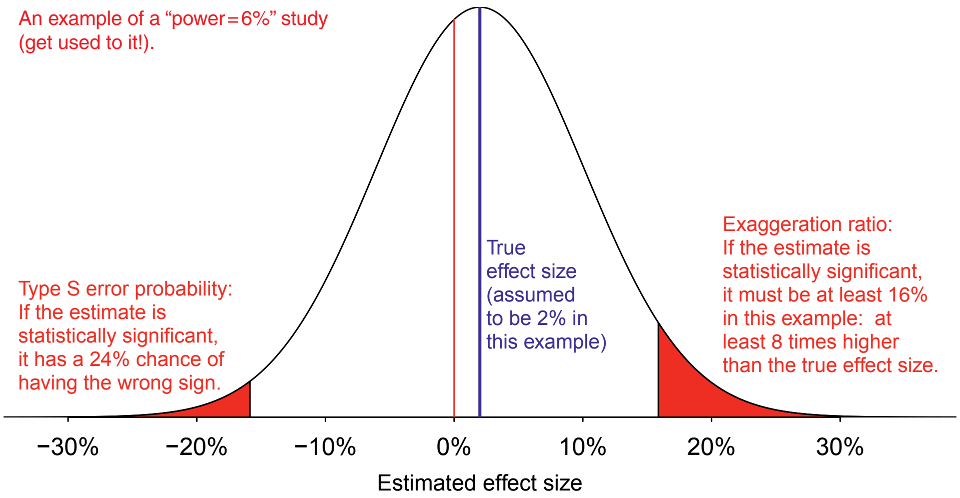


Figure 1. A study has low power when the population difference or effect size is small, while variation and measurement are also small. In low-power studies, the "Type S (sign) error rate" – the probability that the observed difference is in the opposite direction to the true effect or population difference – can be high, even if the estimate is statistically significant. And the "exaggeration ratio" – the factor by which the observed estimate exceeds the true parameter value being estimated – can be huge. The particular numbers in this graph come from a study of a difference in political attitude, comparing women at different times in their menstrual cycles, for which we know, based on substantive grounds, that the true population effect size could be at most 2 percentage points. The bell-shaped curve represents the distribution of estimates that could occur in a study with this precision. The shaded red areas indicate the probability of obtaining a statistically significant effect (the "power", which in this case is 6%). Given the precision of this particular study, for an estimate to be statistically significant it would have to be at least 16 percentage points (that's two standard errors away from zero), hence at least eight times larger than any true effect. And the probability that an estimate in this example is the wrong sign, if it is statistically significant, is 24% – the proportion of the red shaded areas on the negative side of the graph

conservative, more religious, and more likely to vote for Mitt Romney."

This dramatic set of claims was supported by a statistically significant comparison: an interaction effect estimated at about 20 percentage points that was more than two standard errors away from zero (a standard error being 8.1 percentage points in this example). Based on pre-election survey data, however, we believe that very few people changed their vote intentions during this campaign. A more plausible size of this menstrual-cycle effect would be 2 percentage points or less.

Hence, in Figure 1, the blue line indicating true effect size is at 2 percentage points, which is at the high end of any plausible effect here, and the bell-shaped curve shows the distribution of possible differences in the data that could be observed given this assumed effect size. Due to the high level of variation between people, the distribution is broad, indicating a wide range of possible data that could arise in such a study. The areas shaded red under the curve indicate the probability that the observed difference is "statistically significant"—that is, more than two standard errors away from zero. As the diagram indicates, a statistically significant finding here actually has a high probability of being in the wrong direction (a "Type S (sign)" error) and in any case will be at least 16 percentage points – that is, at least eight times higher than the assumed true effect of 2. In this sort of problem, classical hypothesis testing is a recipe for exaggeration.

When applied to the scientific process more generally, the result of all these hypothesis tests is a flow of noisy claims which bear only a weak relation to reality, but which attain statistical significance, which is, conventionally, a necessary and sufficient condition for publication, if said result is paired with any story that is qualitatively justified by a substantive theory.

Various researchers in psychology and medicine have made the following linked points: statistical significance cannot generally be taken at face value;[1] a scientific publication system based on null hypothesis significance tests leads to large-scale errors in reporting; and these problems are particularly severe in the context of low signal and high noise.[2]

Psychology is particularly subject to such problems, for several reasons:

- The objects of study (mental states, personality traits, cognitive and social abilities) are inherently latent and can typically not be precisely defined.
- Theories are correspondingly vague (in comparison with physics or chemistry, say, or even medicine), in that the magnitude and even the direction of effects cannot always be predicted based on theoretical grounds.
- Variation between people is typically large, as is variation across repeated measurements within people; indeed, analysis of this variation is often a central research goal.
- The stakes are low so that it is easy to quickly do a small study and write up the conclusions. Unlike in medical research, there is no hurdle to performing a publishable study. This is not to say that psychology research is trivial; our point here is just that, compared to much medical research, typical studies in psychology have low, if any, risks to the participants, so the barriers to performing and publishing a study are minimal.

The resulting proliferation of studies with small effect sizes and high noise, along with a willingness of high-profile, prestigious journals such as *Psychological Science* and the *Proceedings of the National Academy of Sciences* to publish surprising, newsworthy findings based on statistically significant comparisons, has led us to a crisis in scientific replication.

Based on the considerations discussed above, I would say that the most important way that statistics can help solve the replication crisis is to recognise the fundamental nature of the problem: if effects are small and measurements are biased and noisy, there is no way out, other than to put effort into taking measurements that are more valid and reliable, most notably in psychology studies by using more carefully designed instruments and performing within-person comparison where possible to reduce variance.

Once better data have been collected, how can statistical inference help? Given the problems with classical significance testing, there should be something better. Some have suggested replacing hypothesis tests with confidence intervals, but this by itself will not solve any problems: checking whether a 95% interval excludes zero is mathematically equivalent to checking whether $p < 0.05$. And, just as statistically significant results can be huge overestimates, confidence intervals can similarly contain wildly implausible effect sizes, estimates that happen to be consistent with the data at hand but make no sense in the context of subject-matter understanding.

One direction for statistical analysis that appeals to me is Bayesian inference, an approach in which data are combined with prior information (in this case, the prior expectation that newly studied effects tend to be small, which leads us to downwardly adjust large estimated effects in light of the high probability that they could be coming largely from noise). I do see a potential Bayesian solution using informative priors and models of varying treatment effects,[3] but these steps will not be easy because they move away from the usual statistical paradigm in which each scientific study stands alone.

To resolve the replication crisis in science we may need to consider each individual study in the context of an implicit meta-analysis. And we need to move away from a simplistic deterministic model of science with its paradigm of testing and sharp decisions: accept/reject the null hypothesis and do/don't publish the paper. To say that a claim should be replicated is not to criticise the original study; rather, replication is central to science, and statistical methods should recognise this. We should not get stuck in the mode in which a "data set" is analysed in isolation, without consideration of other studies or relevant scientific knowledge. We must embrace variation and accept uncertainty.

References

1. Simmons, J., Nelson, L. and Simonsohn, U. (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, **22**, 1359–1366.
2. Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B., 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**, 1–12.
3. Gelman, A. (2014). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management*, **41**, 632–643.

Andrew Gelman is a professor of statistics and political science and director of the Applied Statistics Center at Columbia University. He blogs at andrewgelman.com

Psychology and statistics, continued...

Readers respond to the BASP ban on *p*-values

Andrew Gelman is not alone in questioning the way *p*-values are used in psychology. Earlier this year the journal *Basic and Applied Social Psychology* (BASP) decided to ban the null hypothesis significance testing procedure (NHSTP) from the papers it published. The decision was motivated by "the logical invalidity of the NHSTP", said BASP editor David Trafimow – his argument being that *p*-values say nothing about the probability of a null hypothesis being false, so they should not be used to reject it. Our April report of the story prompted a handful of responses from readers, two of which are published below.

Though I agree with the criticisms and cautions that have been raised about the NHSTP, I feel, as others do, that an outright ban on its use is unwarranted, and I hope the editors of BASP reconsider their decision.

In my opinion, the controversy regarding the NHSTP partly misplaces the locus of the problem. I believe a big difficulty with the current use of the NHSTP is the exaggerated practical implications that have come to be attached to its results. It seems to me that the debate on the NHSTP is implicitly fuelled by the excessive weight given to whether a study's primary results are statistically significant in determining whether the study gets reported in the literature. A study with non-significant findings is often considered a "failure" not