https://doi.org/10.1038/s41562-023-01586-w

Points of significance

The majority of empirical articles that we publish use null-hypothesis significance testing. In most cases, researchers rely on *P* values to establish the scientific or practical significance of their findings. However, statistical significance alone provides very little information that is useful for making inferences about scientific or policy significance. For this reason, we require authors to provide much more information than just *P* values – in this Editorial, we explain our requirements.

Statistical significance and *P* values have been much discussed over the past decade. In 2016, the American Statistical Association published a statement on *P* values, aiming to dispel some of the misconceptions that surround their use and interpretation[1]. Despite heightened attention to the misuse of *P* values, we frequently encounter research that demonstrates the types of misunderstandings that the American Statistical Association statement tried to allay.

In most empirical studies using null-hypothesis significance testing (NHST) that we receive, authors report only the statistical test, degrees of freedom, test value and *P* value. In some cases, we see only *P* values and nothing else. This extremely limited information can be misleading[2] and in studies with very large sample sizes it is meaningless (as overpowered studies or studies with very large samples can identify statistically significant but trivial effects). We therefore require that authors also report effect sizes and confidence intervals. Reporting of NHST statistics should typically take the following form: statistic (degrees of freedom) = value; *P* = value; effect size statistic = value; and per cent confidence intervals = values.

The *P* value threshold of 0.05 for declaring significance is an arbitrary one that is established by convention. However, if authors choose to use NHST, we ask that they abide by the convention (except if they preregistered a different alpha level for their study, providing a robust justification for their choice[3]). Statements such as 'marginally significant' and 'just missed statistical significance' for *P* values above the threshold of 0.05 that are followed by theoretical interpretations as if the null hypothesis had been rejected are misleading. *P* values that exceed the conventional or prespecified threshold are simply not statistically significant and we ask that authors report them as such.

One of the most common issues that we encounter in submitted manuscripts is inferences about differences between studies or conditions, where the authors compare statistical significance levels without using formal statistical tests of the difference itself. In a 2006 article, Gelman and Stern provided a compelling explanation as to why "the difference between significant and not significant is not itself statistically significant"[4]. Using significance levels to compare effect estimates is not appropriate and we ask authors to provide statistical evidence of any argued difference.

If authors carry out multiple comparisons, we expect that they will use a form of adjustment or correction (for example, Bonferroni, Benjamini–Hochberg, family-wise error rate or false-discovery rate) that is appropriate for their data and the number of comparisons they are performing. This correction is an essential part of the analysis (not merely a robustness check) and all interpretations of results should be based on the corrected *P* values.

We select studies for peer review and publication based on the importance of the research question, the breadth of its potential relevance to a multidisciplinary audience and the substantiveness of the evidence, not on the basis of their results. This means that we publish studies where the main results are null. For studies reporting statistically null results, we ask authors not to interpret the absence of evidence as evidence of absence. There is no statistical test that can demonstrate the absence of an effect. Statements such as 'there is no association between X and Y' or 'X has no effect on Y' are inaccurate, and are best revised to read '[no or little] credible evidence of an association between X and Y' or '[no or little] credible evidence that X affects Y.'

Regardless of whether the main or ancillary results are null, if these results are interpreted in the article we ask that authors use an appropriate statistical method for interpreting them (for example, Bayes factors[5] or equivalence tests[6]).

Power is fundamental for all studies, regardless of the direction of the results. Null results in underpowered studies are uninterpretable. If researchers did not use a formal method to prespecify their sample size and the main results of their study are null, we ask that they perform a power sensitivity analysis[7]. This should demonstrate the power of their statistical test across a range of possible effect sizes that includes the smallest theoretically or practically meaningful effect size.

There are numerous calls to retire statistical significance or entirely move away from NHST. Until that happens, however, it is important to make sure that published research using NHST makes statistically valid inferences that are appropriately interpreted.

Published online: 24 March 2023

References

1. Wasserstein, R. L. & Lazar, N. *Am. Stat.* **70**, 129–133 (2016).
2. Amrhein, V., Greenland, S. & McShane, B. *Nature* **567**, 305–307 (2019).
3. Lakens, D. et al. *Nat. Hum. Behav.* **2**, 168–171 (2018).
4. Gelman, A. & Stern, H. *Am. Stat.* **60**, 328–331 (2006).
5. Dienes, Z. *Front. Psychol.* **5**, 781 (2014).
6. Lakens, D. *Soc. Psychol. Personal. Sci.* **8**, 355–362 (2017).
7. Lakens, D. *Collabra Psychol.* **8**, 33267 (2022).