The ASA p-value statement 10 years on
An event of statistical significance?

It’s been a decade since the ASA’s famous statement on p-values. **Robert Matthews** evaluates the aftermath and considers how the debate has evolved

Ten years ago this month, the American Statistical Association (ASA) took the unprecedented step of issuing a statement on one of the most controversial issues in statistics: the use and abuse of p-values.

Popularised in the 1920s by the hugely influential English statistician Ronald Fisher, p-values lie at the heart of “significance testing”, widely used by researchers to claim to have found something interesting lurking in data. Yet despite their ubiquity in research journals, p-values have also long been criticised as misunderstood, misleading and open to abuse.

The problem lies in their definition. p-values typically give the chances of getting an effect at least as impressive as that seen, *assuming* it’s actually just a fluke. If these chances are sufficiently low – less than 0.05 is the traditional standard – the finding is then deemed “statistically significant”. For many researchers, this has been taken as implying that their finding is not a fluke, and worth taking seriously. But this overlooks the fact that p-values are calculated on the *assumption* the result is a fluke. As such, they cannot also be used to decide if this assumption is valid – still less whether the theory being tested is correct. Worse, p-values around 0.05 are at best feeble evidence against fluke results – especially if the result is surprising.

Statisticians have been warning about this for decades, without obvious success. Cynics argue that this is because p-values make it easier to get publishable results. For years another barrier was that, apart from anecdotal evidence of startling claims being prone to collapse over time, there was no systematic evidence of problems with p-values. That changed in the early 2010s with the emergence of the “replication crisis”, where systematic attempts to replicate research claims failed with startling frequency – over 50% in some cases.[1]

Among the many potential causes cited was the abuse of p-values, and practices such as “p-hacking”, where researchers massage the data and analysis until the p-values reach statistical significance.

All this formed part of the background to the ASA’s statement published in March 2016 in *The American Statistician* (*TAS*).[2] At its core were six principles for the proper use and interpretation of p-values (see box), with an editorial explaining the statement’s origins and purpose by Ron Wasserstein, the ASA executive director. The myriad misinterpretations of p-values and related concepts were also covered in a supplementary article co-authored by some of the many leading statisticians who had spoken out in the past.[3] Of the 25 issues identified, one was singled out for particular criticism: the use of the p-value to set a cut-off between “significant” and “non-significant” findings.

The landmark set of articles concluded with commentaries by 22 statisticians, giving their reflections on the past, present and future of the p-value debate (bit.ly/4jvlYIN). Given the apparent lack of progress over the decades, some expressed scepticism about anything changing any time soon: “This much should be clear”, wrote Steven Goodman of Stanford University (bit.ly/3LuKgFY): “What follows this statement is as or more important than the statement itself.”

So now, a decade on from the ASA’s statement, what has become of the p-value debate? As a former journalist turned statistician who has covered the controversy for 30 years, I have reported on the aftermath several times for *Significance*. The first of these “progress reports” appeared in 2017, a year after publication of the ASA’s statement.[4] It described the widespread coverage the statement had attracted in the media, but also the absence of any obvious signs of impact. While it was clearly still early days, I argued that the statement had put too much emphasis on what researchers should *not* do. If progress was to be made, they needed to know what they should be using instead of p-values and significance testing.

That first retrospective included a response by Wasserstein, who revealed that the ASA had this covered. It was organising an international symposium which would address this and many other aspects of moving towards “a world beyond $p < 0.05$”. Held in Bethesda, Maryland, in October 2017, this meeting – dubbed “the Woodstock of Inference” – attracted huge interest; many of the sessions were standing-room only. These also showed there are many ways of extracting much more out of p-values than mere statistically significance; these were explored in some of the 40-plus invited papers published in a special issue of *TAS* in 2019.[5]

An editorial by Wasserstein and two co-authors stressed that the aim was to give researchers ways of moving beyond the $p <$ 0.05 criterion, and declared it was time to go beyond the original 2016 ASA statement and to stop using concepts like statistical significance and threshold values.[6]

The symposium and special issue of *TAS* formed the centrepiece of the second *Significance* progress report on the p-value debate published in 2021, five years after the ASA’s statement.[7] This also covered the re-emergence of an idea that had been circulating since at least the 1990s: tightening the threshold for claiming statistical significance to $p = 0.005$.

Published in *Nature Human Behaviour* in 2018,[8] its authors argued that this “simple step” would help deal with the cause of the renewed concern over p-values: the replication crisis. Using theoretical and empirical evidence, the authors estimated that novel claims meeting the new $p < 0.005$ criterion would be around twice as likely to replicate as those based on the usual 0.05 standard.

While stressing that the proposal left many issues – like p-hacking – unresolved and risked delaying the use of most sophisticated approaches, the authors argued it was “an actionable step that will immediately improve reproducibility”. Even so, critics argued that the new threshold was as arbitrary as the original, and meant that researchers would either have to increase study sizes substantially or face a higher risk of missing genuine effects.

The proposal also seemed

Robert Matthews is a visiting professor in the department of mathematics at Aston University, Birmingham, UK. He is also a former member of the *Significance* editorial board.

to defy the call of both many contributors to the original 2016 ASA publications and the *TAS* editorial of 2019: that the use of *any p*-value threshold and terms such as "statistical significance" should cease. By the end of 2019, the *p*-value debate seemed headed for deadlock. Those arguing for evolutionary change via simple tweaks were at loggerheads with those seeking radical change using more sophisticated methods.

At the same time, the responses of research journals to the debate – probably the most potent driver of any change in practice – were similarly mixed. Top-tier medical journals had already begun moving away from simple *p*-value summaries in favour of more informative confidence intervals, and some referenced the ASA statement to bolster their policies. In contrast, a 2019 editorial on the debate in *Nature* – arguably the most prestigious science journal – said "There are reasonable arguments on all sides". Like many others, the journal's editors seemed content to sit on the fence.

In his aforementioned commentary on the ASA's 2016 statement, Goodman had warned that some authority needed to take responsibility for promoting the ASA's original 2016 principles, "otherwise nothing will change". Those hoping that pre-eminent national scientific bodies would take up this challenge were to be disappointed. In December 2019, the US National Academies of Sciences, Engineering and Medicine published a report on the replication crisis[9] which called on academic institutions to give training "in the proper use of statistical analysis". It also suggested researchers learn to use inferential methods "properly"; ironically, the original report had to be amended to fix a garbled definition of *p*-values (bit.ly/4q7euhu).

Finally, in 2021, the ASA issued another statement (bit.ly/44UxtmQ), this time from a Presidential Task Force whose focus was not promoting the 2016 principles but addressing concerns that the 2019 *TAS* editorial might be seen as official ASA policy. In place of the original 2016 principles, the Task Force gave broad pointers to good statistical practice, and – like the National Academies' report – concluded that *p*-values and significance tests are valuable "when properly applied and interpreted".

For those who saw *im*proper use and *mis*interpretation as the key issue in the *p*-value debate, this seemed to miss the point. Yet by this time, a far more pressing challenge was confronting statisticians world-wide: the Covid pandemic (bit.ly/497LQH3).

"More basic concerns in statistical science became particularly apparent during Covid, for example a really serious acknowledgement that all models are wrong, and every statistical claim relies on judgement," recalls Sir David Spiegelhalter, president of the RSS at the time of the 2016 ASA statement and co-chair of the society's pandemic task force. Responding to the first *Significance* progress report in 2017, Sir David had been sceptical that significance tests could – or should – be eliminated: "*p*-values are just too familiar and useful." His views remain the same: "I suppose, to be honest, that I don't think it's such an important issue."

This pragmatic view is reflected in the fact that, of all the recently proposed methods for dealing with the *p*-value problem, by far the most widely cited is the tightening of the *p*-value threshold to p = 0.005. According to lead author of the 2018 paper Daniel Benjamin, now at UCLA, the basic message of the paper has now been understood by many researchers: "Statistical standards of evidence for claiming new discoveries in many fields of science are too low." He adds that there is now good evidence that findings meeting the tighter criterion are substantially more replicable than those based on p = 0.05. "We now use 0.005 as a default *p*-value threshold for making or interpreting claims of novel discoveries. We continue to encourage other researchers to join us."

A decade after organising the original ASA statement, what does Wasserstein see as its real significance? He points to the 8,000-plus citations and 850,000 views: "The statement appears to have been helpful to researchers and has, to some extent, even affected entire disciplines," he says. "Journal editors have been more mindful of good practice regarding *p*-values because authors and reviewers have brought the statement to their attention."

This may not be the radical change many wanted to see. Yet for the statistical community – the collective noun for which has been said to be "a quarrel" – evolution not revolution was perhaps always the most likely outcome. ■

References

1. Johnson, V. E., Payne, R. D., Wang, T., Asher, A. and Mandal, S. (2017) On the reproducibility of psychological science. *Journal of the American Statistical Association*, **112**(517), 1–10.

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

3. Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N. and Altman, D. G. (2016) Statistical tests, *P*-values, confidence intervals, and power: A guide to misinterpretations. *The American Statistician*, online supplement. bit.ly/2kX5obX

4. Matthews, R., Wasserstein, R. and Spiegelhalter, D. (2017) The ASA's *p*-value statement, one year on, *Significance*, **14**(2), 38–41.

5. Wasserstein, R. L., Schirm, A. L. and Lazar, N. A. (eds) (2019) Statistical inference in the 21st century: A world beyond $p < 0.05$. *The American Statistician*, **73**(supplement 1).

6. Wasserstein, R. L., Schirm, A. L. and Lazar, N. A. (2019) Moving to a world beyond "$p < 0.05$". *The American Statistician*, **73**(supplement 1), 1–19.

7. Matthews, R. (2021) The *p*-value statement, five years on, *Significance*, **18**(2), 16–19.

8. Benjamin, D. J., Berger, J. O., Johannesson, M. *et al.* (2018) Redefine statistical significance. *Nature Human Behaviour*, **2**(1), 6–10.

9. National Academies of Sciences, Engineering, and Medicine (2019) *Reproducibility and Replicability in Science.* Washington, DC: National Academies Press.

The ASA Statement's six principles of the proper use and interpretation of p-values (2016)

1. ***p*-values can indicate how incompatible the data are with a specified statistical model.**
2. ***p*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.**
3. **Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.**
4. **Proper inference requires full reporting and transparency.**
5. **A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.**
6. **By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.**