The *p*-value statement, five years on

The American Statistical Association's 2016 *p*-value statement generated debates and disagreements, editorials and symposia, and a plethora of ideas for how science could be changed for the better. Now, five years on, **Robert Matthews** asks what, if anything, has the statement achieved?

Five years ago, in March 2016, the American Statistical Association (ASA) published its landmark statement on the most-widely used – and abused – method for extracting insight from data.[1] Known as null-hypothesis significance testing (NHST), for almost a century it has been the go-to method for researchers seeking to show they have made a discovery.

And that, as the ASA statement made clear, is precisely the problem. The key concepts of NHST – and, in particular, *p*-values – cannot do what researchers ask of them. Despite the impression created by countless research papers, lecture courses and textbooks, *p*-values below 0.05 do not "prove" the reality of anything. Nor, come to that, do *p*-values above 0.05 disprove anything. As the ASA's statement pointed out, statisticians have been trying to make this clear for decades without success. By bringing the issue to public attention, the board of the world's largest professional association of statisticians hoped to "draw renewed and vigorous attention to changing the practice of science with regards to the use of statistical inference".

Five years on, the ASA has clearly achieved that aim. Its statement has been viewed almost half a million times and received over 3,600 citations, and it has spawned countless articles in academic journals and even the popular media. Yet when it first appeared, many – including this author[2] – doubted it would do more than spark yet more debate about how to "steer research into a 'post $p < 0.05$ era'" (bit.ly/2aQdmln).

So, five years on, what has the statement achieved? Both a huge amount, and very little.

Robert Matthews is a visiting professor in the Department of Mathematics, Aston University, Birmingham, UK. As a science writer, he has been reporting on the role of NHST in undermining the reliability of research since the late 1990s.

The Battle of the Commentaries

One major criticism of the March 2016 statement was its focus on the many ways NHST can be – and is – abused by the research community, while giving scant guidance on remedies. This reckoned without the determination of ASA executive director Ron Wasserstein and his colleagues to maintain the momentum. An international symposium was organised in October 2017 (bit.ly/3q4Io6i), along with a special open-access issue of *The American Statistician* (*TAS*) dedicated to practical ways of moving beyond $p < 0.05$.[3] Both generated huge interest; sessions at the symposium – dubbed "The Woodstock of Inference" – were often standing-room only, and the special issue of *TAS* included over 40 papers. It also carried an editorial which went beyond the ASA's original statement, declaring "it is time to stop using the term 'statistically significant' entirely", along with variants like "non-significant" and "$p < 0.05$".

All this set the stage for arguably the biggest achievement of the ASA's statement over the last five years: its encouragement of multiple strategies for assailing the edifice of NHST.

First to emerge was an attempt at evolution rather than revolution. In September 2017, *Nature Human Behaviour* carried a commentary calling for the *p*-value threshold for declaring new findings statistically significant to be tightened from 0.05 to 0.005.[4] The authors – among them many high-profile statisticians – argued that this "simple step" would help combat the problem which had rekindled the *p*-value debate in the first place: the "replication crisis". This centred on studies revealing a startlingly high proportion of research claims failing to be replicated in fields ranging from psychology and medicine to economics. There are plenty of explanations for such failures, from illicit trawling of data for "significant" *p*-values to incompetence and fraud. The commentary argued that a key part of the problem is that the standard threshold for statistical significance is just too lenient. Tightening it by an order of magnitude would, the authors claimed, "reduce the false positive rate to levels we judge reasonable".

Their argument was backed by theoretical arguments and empirical evidence suggesting that findings with $p < 0.005$ were roughly twice as likely to replicate as those merely meeting the usual standard. While conceding it was no panacea, the authors argued that adopting 0.005 was at least "an actionable step that will immediately improve reproducibility", with findings meeting only the usual standard now being termed "suggestive".

The response reflected renowned statistician John Tukey's remark that the collective noun for the profession should be a "quarrel" of statisticians. Even in pre-print form, the commentary came under attack. "Very disappointed such a large group of smart people would give such horribly bad advice", tweeted methodologist Daniël Lakens of Eindhoven University of Technology.

Lakens organised a multi-author rejoinder, which appeared in the same journal six months later.[5] Its authors argued the proposed $p = 0.005$ threshold was no less arbitrary than $p = 0.05$, and that the supposed increase in evidential weight rested on questionable assumptions. As for the empirical support for tightening the standard, Lakens *et al.* said it failed the very standard set by its advocates, being itself merely "suggestive". They also warned that insisting on $p = 0.005$ could result in fewer replication studies, as sample sizes – and costs – would have to increase to give a reasonable chance of meeting the tougher standard. This in turn could lead to increasing use of larger but potentially biased data sets, like online survey results.

In short, Lakens and his co-authors wanted a broader and deeper assault on NHST, with everything from prior evidence and study design to target effect size and precision clearly stated and justified. As for "redefining" statistical significance, Lakens *et al.* presaged the *TAS* editorial by recommending the term be expunged.

Trench warfare

The Battle of the Commentaries kept the NHST debate burning well into 2019, even in the popular media. "What a nerdy debate about p-values shows about science – and how to fix it", declared the US news website *Vox* (bit.ly/2MzLWQ6). Yet 2019 now appears to have been the year when the mass assaults on NHST settled into trench warfare.

That spring saw the publication of the special issue of *TAS*, along with its many proposals for moving beyond *p*-values.[3] It also saw *Nature* – arguably the world's most prestigious research journal – publish a call by three high-profile statisticians for the concept of statistical significance to be ditched.[6] This time, however, the focus was on the concept of non-significance, and how it routinely fools researchers into dismissing potentially genuine effects as "null results" (see "Nothing to see here?", page 18).

Citing journal surveys where over half the papers wrongly interpreted non-significant findings as implying no effect, the authors declared: "We're frankly sick of seeing such nonsensical 'proofs of the null'" – a widely shared sentiment, judging by the 800-plus academic co-signatories of the article from over 50 countries. Any hope that *Nature* would institute major changes itself were quickly dashed, however. "There are reasonable arguments on all sides", intoned an editorial in the same issue. "*Nature* is not seeking to change how it considers statistical analysis in evaluation of papers at this time".[7]

Worse was to follow in December 2019, with publication of the US National Academy of Sciences (NAS) Consensus Study Report on the replicability crisis.[8] While acknowledging there were problems with NHST, the report blandly recommended that academic institutions "should include training in the proper use of statistical analysis and inference" and that "[r]esearchers who use statistical inference analyses should learn to use them properly." Ironically, the report itself defined *p*-values as a "measure of the likelihood that an obtained value occurred by chance" – a particularly mangled version of the usual misconception. The NAS later amended the wording to reflect the correct definition (see "Nothing to see here?") though its anodyne recommendations remained.

In publishing the report, the United States' leading scientific academy had at least recognised the problematic use of NHST by researchers. In contrast, the UK's Royal Society – the world's oldest national scientific institution – has yet to make a single substantive statement on arguably the most pressing threat to the scientific enterprise.

With prestigious academies showing no interest in substantive change and most leading journals doing little more than tweak ▶

their guidance to authors, there has been little incentive for working researchers to move towards a post $p < 0.05$ world. Instead, they have stuck to giving journal editors and referees what they so often demand: "proof" of novel claims based on significance tests.

Why change needs to come

The reality is that, in terms of changing research practice, the ASA statement has achieved little. Yet the need for such change has never been greater. The scale and importance of what remains to be done is exemplified by a study published in the prestigious *Journal of the American Medical Association* (*JAMA*) at the height of the NHST debate.

The ANDROMEDA-SHOCK randomised controlled trial (RCT) was set up to compare strategies for treating patients with septic shock – a life-threatening drop in blood pressure triggered by infection.[9] The researchers wanted to know if the risk of death could be reduced by treating patients on the basis of so-called capillary refill time (CRT) rather than blood lactate levels. CRT measurement is relatively simple, quick and needs fewer resources, so a positive result would be good news for patients, especially those treated in low-tech health-care systems.

Over 400 patients were recruited into the trial, each randomly assigned to receive either the CRT or the lactate-based strategy, and monitored for 28 days. The results suggested CRT was indeed better: the mortality rate among the CRT patients was 8.5% lower than for the lactate approach, while the so-called hazard ratio was 0.75 – a 25% improvement. In designing the trial, however, the researchers had assumed a relatively large 15% mortality reduction. That optimistic assumption led to the trial being underpowered, producing positive results which nevertheless failed to meet the conventional $p = 0.05$ threshold. The findings were thus "non-significant", and led the researchers to declare that the CRT strategy "did not reduce all-cause 28-day mortality".

Statisticians took to social media to lament what they saw as yet another case of a top journal failing to spot the blunder of non-significance being interpreted as no effect. It then emerged that the researchers also believed CRT was better, and had sought to "go stronger on the conclusion" – only to be told by the journal's referees and editor "to temper our enthusiasm and stick to the cold stats" (bit.ly/3r3dfvL).

The journal declines to comment beyond stating that "*JAMA* editors have detailed their policies regarding interpretation of RCTs". But for many statisticians, whatever those policies might be, the stated conclusion of ANDROMEDA-SHOCK is just flat wrong.

The controversy prompted a reanalysis of the findings using Bayesian methods,[10] which can help extract insights from the inferential tangle of NHST. In the case of ANDROMEDA-SHOCK, the analysis set the findings in the context of different levels of prior insight into the effectiveness of CRT. This showed that despite being "non-significant", the chances of the CRT strategy outperforming lactate in cutting 28-day mortality exceeded 90% under all the scenarios examined.

Despite its usefulness – acknowledged in the ASA's 2016 statement – such Bayesian analysis remains rare in leading journals. Part of the reason lies in the troubled history of Bayesian methods and their reputation for complexity. As the special issue of *TAS* showed, however, it is not necessary to use full-blown Bayesian methods to go beyond significance and non-significance. Techniques needing nothing more than a calculator can "unpack" standard confidence intervals and *p*-values, revealing additional insights (see "Making *p*-values work harder").

There is now growing acceptance among medical specialists that the conclusion of the ANDROMEDA-SHOCK trial was misleading. Even so, it continues to cast a shadow over how best to treat patients with septic shock – a condition which has since become part of a global threat to health: Covid-19.

An inferential virus

March 2021 marked the first anniversary of the World Health Organization declaring the spread of Covid-19 to be a pandemic – one that has (as of 1 March) claimed over 2.5 million lives. It has also been a year when the cost of failing to deal with the shortcomings of NHST has never been starker.

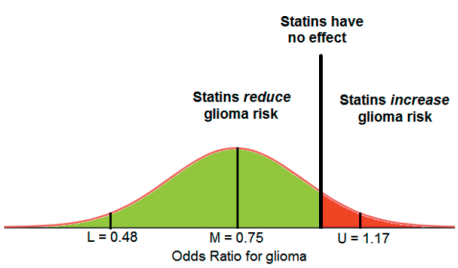
Nothing to see here?

How non-significance can be anything but

Much of the concern about *p*-values lies in their misinterpretation. They represent the probability of getting at least as large an effect as that seen, *assuming* it is just a fluke. All too often, however, this convoluted definition is twisted into something simpler, more useful but quite different: the probability that an effect *really is* just a fluke. And this has led to *p*-values below 0.05 being widely – and wrongly – seen as "proving" the reality of an effect. "Non-significant" *p*-values – those exceeding 0.05 – are equally widely and wrongly seen as "proof" that an effect does not exist. Such misconceptions can have bizarre consequences, such as studies being said to refute others when they are actually consistent.

A case in point concerns evidence that statins – widely used to control cholesterol – may cut the risk of developing brain tumours termed "gliomas" by around 25%. In 2016 a large study seemed to confirm this risk reduction.[12] However, the level of uncertainty ranged from a 52% reduction all the way up to a 17% increase in risk, implying a *p*-value of 0.2. This made the finding "non-significant", leading the researchers to conclude their findings did not support the previous claims. Simply sketching out the probability distribution tells a different story, however (see illustration).

The most likely value is a 25% risk reduction – in line with the previous studies – and the area under the curve corresponding to the "non-significant" *p*-value (in red) is clearly much smaller than the area supporting a protective effect from statins (green). As Valentin Amrhein *et al.* lamented in their *Nature* article: "How do statistics so often lead scientists to deny differences that those not educated in statistics can plainly see?"[6]



Making *p*-values work harder

Following the ASA statement in 2016, there was a call for methods that could help researchers move beyond $p < 0.05$ when reporting their findings. In March 2019, a special issue of *The American Statistician* (*TAS*) published a variety of techniques for turning *p*-values from inferential traps into sources of insight.[3]

Statisticians have long warned against using *p*-values to decide whether an effect is real or not. Given the confusion over their meaning, many have called for *p*-values to be banned. However, others argue that, correctly interpreted, *p*-values are compact but versatile sources of insight. Among them is Sander Greenland of the University of California, Los Angeles. In the *TAS*, he suggests that *p*-values are best seen as measures of compatibility with various hypotheses – not just no effect – under specific assumptions.[13] Applied to the ANDROMEDA-SHOCK trial (see main text), Greenland's approach shows that the "non-significant" finding is actually *more* compatible with the existence of benefit from CRT than with no effect.

Another issue with *p*-values is that their interpretation depends on the purpose and design of the experiment that generates them. In the same issue of *TAS*, biostatistician Rebecca Betensky of New York University puts forward a way of setting *p*-values in the context of the effect size being investigated and the sample size used.[14] Betensky's method shows how to take these constraints into account when interpreting *p*-values. In the case of the ANDROMEDA-SHOCK trial, the method again reveals that despite the "non-significance" of the finding, this does not imply CRT is no better than the lactate method. The "non-significant" finding could even have turned into evidence of genuine benefit had the target level of improvement been less optimistic. As Betensky puts it: "Context is everything".

Other researchers are also putting forward ways of getting more from *p*-values. Among them is Leonhard Held of the University of Zurich, who has developed methods linking *p*-values to the "inherent credibility" of findings and the probability of a successful replication.[15] Held's methods show that a replication of ANDROMEDA-SHOCK has a 90% probability of showing CRT is better, despite the original finding being "non-significant".

In the scramble for clues as to how best to tackle the pandemic, researchers have turned to the published literature, only to find a morass of poorly designed studies. Often too small to give clear answers by themselves, these findings have been combined to pool their evidential weight. Even that has sometimes led to positive but non-significant findings – leading reviewers to fall into the trap of declaring there is no evidence for any benefit. A year into the pandemic, the role of such simple countermeasures as antiseptic gargling is still blighted by mangled interpretations of early studies.[11]

The world is now looking to vaccines to end the Covid pandemic and find a *modus vivendi* for living with a virus that is probably here to stay. If the five years since the ASA's statement have shown anything, it is that the inferential virus of NHST is not going away any time soon either. But over those five years, a potential way forward has emerged, a way of "inoculating" researchers against the most pernicious effect of NHST: the delusory belief that statistical significance constitutes proof.

Both the ASA and Royal Statistical Society (bit.ly/3uSlGRo) have encouraged the development of statistical vaccines, in the form of simple methods for extracting more insight from *p*-values and confidence intervals with less risk of misinterpretation (again, see "Making *p*-values work harder"). Now these techniques need to be actively promoted with the explicit aim of showing researchers it is no longer necessary – or acceptable – to simply take findings, apply a discredited procedure based on $p = 0.05$ and claim a discovery.

The biggest barrier to bringing about this transformation may be the statistical community itself. History suggests that no inferential technique yet devised has escaped being condemned as fatally flawed by some element of the "quarrel". Yet ending the pandemic of unreliable research driven by NHST requires pragmatic acceptance that all inferential methods can mislead, but some are far more misleading than others.

Unless researchers get the help they need from the statistical community soon, the threat posed by NHST could prove fatal to the scientific enterprise. ■

Disclosure statement

The author declares no conflicts of interest.

References

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

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

3. Wasserstein, R. L., Schirm, A. L. and Lazar, N. A. (eds) (2019) Statistical inference in the 21st century: A world beyond $p < 0.05$. *American Statistician*, **73**(sup1).

4. Benjamin, D. J., *et al.* (2018) Redefine statistical significance. *Nature Human Behaviour*, **2**(1), 6–10.

5. Lakens, D., *et al.* (2018) Justify your alpha. *Nature Human Behaviour*, **2**(3), 168–171.

6. Amrhein, V., Greenland, S. and McShane, B. (2019) Retire statistical significance. *Nature*, **567**, 305–307.

7. Editorial (2019) It's time to talk about ditching statistical significance. *Nature*, **567**, 283.

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

9. Hernández, G., *et al.* (2019) Effect of a resuscitation strategy targeting peripheral perfusion status vs serum lactate levels on 28-day mortality among patients with septic shock: The ANDROMEDA-SHOCK randomized clinical trial. *Journal of the American Medical Association*, **321**(7), 654–664.

10. Zampieri, F. G., *et al.* (2020) Effects of a resuscitation strategy targeting peripheral perfusion status versus serum lactate levels among patients with septic shock. A Bayesian reanalysis of the ANDROMEDA-SHOCK trial. *American Journal of Respiratory and Critical Care Medicine*, **201**(4), 423–429.

11. Matthews, R. A. J. (2020) A simple, low-cost potential means of protecting healthcare staff is being overlooked. *British Medical Journal*, **369**, m1324.

12. Seliger, C., *et al.* (2016) Statin use and risk of glioma: Population-based case–control analysis. *European Journal of Epidemiology*, **31**, 947–952.

13. Greenland, S. (2019) Valid *P*-values behave exactly as they should: Some misleading criticisms of *P*-values and their resolution with *S*-values. *American Statistician*, **73**(sup1), 106–114.

14. Betensky, R. A. (2019) The *p*-value requires context, not a threshold. *American Statistician*, **73**(sup1), 115–117.

15. Held, L. (2019) The assessment of intrinsic credibility and a new argument for $p < 0.005$. *Royal Society Open Science*, **6**(3), 181534.