The ASA's *p*-value statement, one year on

Its aim was to stop the misuse of statistical significance testing. But **Robert Matthews** argues that little has changed in the 12 months since the ASA's intervention

A little over a year ago now, in March 2016, the American Statistical Association (ASA) took the unprecedented step of issuing a public warning about a statistical method. Published in *The American Statistician*,[1] it came with statements from leading statisticians suggesting the method was damaging science, harming people – and even causing avoidable deaths.

The notion of a lethal statistical method may seem outlandish, but no statistician would have been surprised by either the allegations or the identity of the accused: statistical significance testing and its notorious reification, the *p*-value.

From clinical trials to epidemiology, educational research to economics, *p*-values have long been used to back claims for the discovery of real effects amid noisy data. By serving as the acid test of "statistical significance", they have underpinned decisions made by everyone from family doctors to governments. Yet according to the ASA's statement, *p*-values and significance testing are routinely misunderstood and misused, resulting in "insights" which are more likely to be meaningless flukes.

Not surprisingly, the ASA's statement attracted widespread media coverage. Many journalists spotted the obvious connection between the unreliability of *p*-values and one of the biggest scientific controversies of our time: the replication crisis, in which widely cited research claims fade away on reinvestigation.

Given the implications for the trillion-dollar global scientific enterprise, one might have expected the ASA's statement to prompt urgent meetings by academic bodies like the US National Academy of Sciences, and revised guidance to authors by leading research journals. Perhaps even a science minister or two might have seen fit to demand action over the presumably egregious waste of time, effort and taxpayers' money caused by the use of significance testing.

Yet a year on, it is not clear that the ASA's statement has had any substantive effect at all. A quick check of the latest issues of leading journals like *The Lancet* or *Proceedings of the National Academy of Sciences* shows it's business as usual – even in papers submitted after the ASA's statement. Claims are backed by the sine qua non of statistical significance "$p < 0.05$", plus a smattering of the usual symptoms of statistical cluelessness like "$p = 0.00315$" and "$p < 0.02$". Of course, one could argue it's still early days. But in the press release accompanying the statement (bit.ly/2mw2mXF), the ASA's then-president Jessica Utts pointed out what all statisticians know: that calls for action over the misuse of *p*-values have been made many times before.

Liudmyla_M/Bigstock.com

Robert Matthews is visiting professor in the Department of Mathematics, Aston University, Birmingham, and a member of the editorial board of *Significance*. His latest book is *Chancing It: The Laws of Chance and How They Can Work for You*

As she put it, "statisticians and other scientists have been writing on the topic for decades".

To a journalist, this is a worrying symptom of a non-story. It raises suspicions that the issue cannot be that serious, as surely the scientific community would have done something about it by now. Perhaps statisticians are just trying to raise their profile using some nitpicking argument that does not really matter.

Statisticians know differently, of course. The misuse of statistical significance testing is a very big deal (see box). A year on from the statement, however, it seems to have slipped off the agenda – again. What will it take to get substantive action?

Personal history

This is a question I have been wrestling with for 20 years, since first investigating the *p*-value issue for a national newspaper in the UK in the late 1990s. As science correspondent of the *Sunday Telegraph*, my bread-and-butter work lay in reporting on research published in refereed academic journals. But over the years I became intrigued by how many "statistically significant" findings seemed to fade away, contradicted by subsequent studies.

As a science graduate, I knew something about *p*-values, and it seemed the fade-out rate was higher than the 1-in-20 limit I thought $p < 0.05$ implied. In nutritional studies and epidemiology in particular, the flip-flopping of findings was striking. Reading around the subject, I learned there were many reasons for the lack of convergence on a single "truth": small study size, confounding, bias. But then the same flip-flopping began appearing in large randomised controlled trials of life-saving drugs.

Sensing a major science story, I started reading more technical papers in search of clues as to what might be to blame. The breakthrough came when I read the classic 1987 paper by Berger and Sellke[2] in the *Journal of the American Statistical Association*, which quantifies the impact of regarding a *p*-value as a measure of weight of evidence. Their findings were – to me at least – shocking. Put simply, they showed that even under broad assumptions about the underlying model for the data, a *p*-value of 0.05 could exaggerate the true "significance" of a finding by an order of magnitude.

My first story on the connection between *p*-values and unreliable science appeared on the front page of the *Sunday Telegraph*'s Review section in September 1998. Rightly concerned that readers might find a 2700-word critique of significance testing somewhat daunting, the editor insisted on an eye-catching if somewhat inappropriate headline: "The Great Health Hoax" (bit.ly/2lGSbBi).

By then, however, I already suspected that even the most tabloid headline would not affect the impact of the story. During my research, I had contacted various academic institutions, including the Royal Statistical Society (RSS), to get their view on the problem with *p*-values. The responses were strikingly similar: yes, we know about it, yes, we have considered taking action – but no, we are not going to do so, because it would cause too much upheaval. A contact at the RSS told me off the record that the fear was that it would rekindle the bitter postwar rows between the Bayesian and frequentist communities.

With no backing from the professional bodies, and with only anecdotal evidence of a problem that even many scientists struggle to grasp, there was no follow-up from the rest of the media. Determined to press on, I tried to drum up Parliamentary interest, and met the chairman of an influential House of Commons select committee. He listened intently, and quickly grasped the nature of the issue. Then came the home truths. Without hard evidence of the size of the problem – either in terms of wasted taxpayers' money or lives lost, preferably both – his committee could not justify investigating further.

Over the next few months, I wrote about the *p*-value issue for some specialist media outlets, to no greater effect. Even *New Scientist* struggled to see the story as anything other than statistical pedantry. So I gave up, wrote up some thoughts I had had for academic journals,[3] and moved on – always expecting the controversy would flare again one day.

Back in the headlines

The first glimmerings emerged a few years ago, in the form of reports of failed replication studies in *Science* and *Nature*. Instead of my anecdotal evidence, these were systematic attempts to replicate highly cited research – and they were encountering disturbingly high failure rates. Then in 2015 came the decision by the editors of *Basic and Applied Social Psychology* to ban the use of *p*-values by submitting authors.

Journalists saw the connection – as did the ASA Board, which asked executive director Ronald Wasserstein to assemble a panel of experts to work towards a policy statement. By all accounts, it proved much harder than expected. A year on, the question is: was it worth it?

The *p*-value problem in a nutshell

A simple means of investigating claims of extra-sensory perception (ESP) is to ask people to guess the identity of so-called Zener cards, each bearing one of five different symbols. As random guessing thus has a 20% chance of success, someone achieving a hit rate of, say, 32 correct guesses in 100 attempts has clearly done something unusual. But is it convincing evidence of ESP?

An obvious way to measure this would be to work out the probability

Prob(ESP is at work, given 32 hits out of 100)

with a high value suggesting impressive evidence. Unfortunately, the laws of probability show that estimating this probability demands a number of controversial assumptions – not least about the inherent plausibility of ESP. But statistical textbooks offer something apparently very similar: a measure of the "statistical significance" of the result called a *p*-value, defined as

Prob(at least 32 hits, given mere fluke was the cause)

Using the appropriate formula, this hit-rate implies a *p*-value of 0.003. That is, there is barely a 1 in 300 chance of getting at least as impressive a result, if mere fluke were the true cause. Better yet, this value easily passes the time-honoured $p < 0.05$ test for "statistical significance". Indeed, it is so low it seems clear that the probability ESP is at work must be correspondingly high. But that is where *p*-values spring their trap. As they are calculated *assuming* fluke is the real cause, they cannot simply be flipped around to give a measure that this assumption is correct. Worse, when the conversion is done properly, *p*-values typically prove to be radical *under*estimates of the chances that fluke accounts for the findings – thus exaggerating their apparent "significance".

In short, *p*-values give a convoluted answer to a question of very restricted interest – so restricted, in fact, that it is easy to be tricked into thinking they *must* mean something more significant than they do.

The commentaries accompanying the statement (bit.ly/2kX50bX) show that even at the time some participants feared there would be no lasting impact. Unsurprisingly they included distinguished veterans of previous failed attempts to tackle the *p*-value issue such as James Berger, Donald Berry and Kenneth Rothman. Steven Goodman put his concerns bluntly: "We need to formulate a vision of what success looks like, and how we will get there. If not, we can start drafting the language of the 2116 ASA statement tomorrow."

As far as I could tell, the vision was one of researchers recognising the vital importance of appropriate study design, data acquisition and the use and interpretation of inferential methods. Yet try as I might, I could see no answer to the biggest question raised by the ASA's statement. Statisticians are clearly disturbed by what passes for inference in most scientific disciplines. But what, exactly, does the profession propose researchers do instead?

On this specific issue, the ASA statement is strikingly – and, I would argue, fatally – vague. There is an anaemic explanation of how "[S]ome statisticians prefer to supplement, or even replace *p*-values" with alternatives such as estimation, Bayes factors and false discovery rates. Goodman's frustration with this resonated with mine: "Exactly how [are] scientists supposed to do that?", he asks in his commentary. "If we are to make such recommendations, we need to figure out what to tell or teach people."

A way forward

The statement is clearly not the place to go into prescriptive detail. Even so, the commentaries give clues to the cause of the vagueness of the ASA guidance: the lack of consensus among statisticians about how best, as Wasserstein has put it, to "steer research into a post $p < 0.05$ era". While the Bayesian versus frequentist controversy I encountered 20 years ago has faded, it has been replaced by factionalism. Everything from simply lowering *p*-value thresholds to full Bayesian analyses is being advocated, via half-way houses such as Bayesian conversion of *p*-values and the use of uninformative priors. Yet while the profession ruminates over their relative merits, the workaday researcher's question goes unanswered: how do I turn my data into insight?

For decades, the answer was clear: reach for *Statistical Methods for Research Workers*. Ronald Fisher wrote his hugely successful text for the express purpose of putting inferential tools "into the hands of research workers".[4] By giving them simple recipes that answered their routine statistical challenges, he ensured that few felt minded to ask awkward questions about the underlying concepts. Those who did had their doubts swept away by the peremptory one-paragraph dismissal of Bayesian methods in the introduction to the book.

Almost a century after its publication, its subliminal message has seeped into software, lecture courses and countless "statistics for scientists" texts: inference is simply a game – and anyone can play.

The replication crisis has given the statistical community a priceless opportunity to correct this misconception. It opens the way for the ASA and RSS to reach out to their counterparts in the sciences and, working together, provide what Fisher did: authoritative guidance on dealing with standard inferential problems encountered in each discipline.

There should, however, be a key difference: the guides should illustrate the use of the various alternatives to significance testing – while pointing out their limitations, and giving advice on when specialist help is required.

Given its proven threat to reliable inference, I would argue that significance testing has no place in such guidance, except to illustrate its pitfalls. Others, doubtless, will vehemently disagree. Yet Berry was surely right in stating in his commentary that: "Our collective credibility in the science community is at risk." The way forward, I believe, lies in conceding that there is no single, perfect way to turn data into insight. Instead, the goal of the statistical community should be to help researchers choose better options.

Only this can end the parody of inference that has blighted research for so long. ■

References

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

2. Berger, J. O. and Sellke, T. (1987) Testing a point null hypothesis: The irreconcilability of *p* values and evidence. *Journal of the American Statistical Association*, **82**(397), 112–122.

3. Matthews, R. A. J. (2001) Methods for assessing the credibility of clinical trial outcomes. *Drug Information Journal*, **35**(4), 1469–1478.

4. Fisher, R. A. (1925) *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

"The game is still afoot"

By Ron Wasserstein

We concede. There is no single, perfect way to turn data into insight. The only surprise is that anyone believes there is!

Science is complex. Inference is hard work. It has been extraordinarily costly to science that the shared understanding of generations of researchers has been that a *p*-value, or any other single index, could provide a simple, clear, objective answer to the question "What does this data tell us?"

Thus, the leadership of the ASA was keen to join in the battle that Robert Matthews describes and that he and many, many others have long fought. We were keen to join because it is a battle that must be won. We engaged in this struggle with complete clarity that change would not be easy. We agree with Matthews that more scientific organisations should be taking up the cause, that there should be a sense of urgency that is largely lacking in most quarters. However, the game is still afoot.

In the year since the *p*-value statement was released, the document has been downloaded over 175 000 times and has already been cited several hundred times. These citations are largely in other research domains, not in statistical journals. Major US organisations such as the American Medical Association, the American College of Cardiology and the American Society for Clinical Oncology have taken the statement seriously enough to prominently address it in journals, on websites, and in newsletters (see bit.ly/2m3dGoZ, bit.ly/2mvebwa, and bit.ly/2nuG2P2 for examples). The statement was mentioned many times at last month's Sackler Symposium on Reproducibility of Research at the National Academy of Sciences in Washington, DC. Attention is being paid.

We also agree with Matthews that the ASA statement was specific on what not to do and vague on what to do. The statement did not go as far as it should go, but it went as far as it could go. Matthews is right about a lack of consensus among statisticians about how best to navigate in the post $p < 0.05$ era. This is unsurprising for many reasons, but especially because it is unreasonable to expect one fundamental approach to solve every inferential issue. A lack of consensus, however, does not imply that

there are not plenty of powerful approaches to improved inference.

To address the statement's shortcomings, the ASA is convening a Symposium on Statistical Inference this October. The tagline for the symposium is "Scientific Method for the 21st Century: A World Beyond $p < 0.05$". Discussions will centre on specific approaches for improving statistical practice as it intersects with three broad components of research activities: conducting research; using research; and sponsoring, disseminating, and replicating research. The vision of the symposium is to push change forward, change that leads to lasting improvements in research, in communicating and understanding uncertainty, and ultimately in decision-making.

We cannot accomplish this simply by having presentations at a conference. Instead, we envision teams of symposium delegates developing papers, briefs, practice guides, and statements on a wide variety of topics to help researchers, research sponsors, journal editors and referees, regulators, educators, the media, and policy- and other decision-makers.

If the symposium is successful in doing *some* of this, research will benefit. Yet if the symposium is successful at *all* of this, we will not really have achieved success until we have not only identified for researchers a rich variety of inferential methods and the situations in which they should be applied, but also ensured that these methods are being taught wherever researchers are being trained. ■

■ **Ron Wasserstein** is executive director of the American Statistical Association

"Too familiar to ditch"

By David Spiegelhalter

I have a confession to make. I like *p*-values. I like skimming regression output or large tables for those twinkling stars (and mentally checking if the proportion is any more than I would expect from chance alone). And I also like a single carefully adjusted *p*-value that helps summarise an entire experimental programme, such as the "five sigma" ($p = 1$ in 3.5 million) attached to the Higgs boson. As the first point of the 2016 ASA statement says, *p*-values can be useful summaries of the compatibility between data and hypotheses.

Concern about *p*-values is being driven by claims of a "reproducibility crisis". But how much are *p*-values to blame for this situation? Among the fine commentaries accompanying the ASA statement, many point out that the problem lies not so much with *p*-values in themselves as with the willingness of researchers to lurch casually from descriptions of data taken from poorly designed studies, to confident generalisable inferences. The ASA critique is great, but what is to be done about this issue that any half-decent statistician knows so well?

It should be possible to establish firm general principles which focus on what is right rather than what is wrong

Robert Matthews appropriately calls for "authoritative guidance on dealing with standard inferential problems encountered in each discipline", although I do wonder how this guidance is to be produced when there are so many different opinions among "authorities". He then argues that "significance testing has no place in such guidance, except to illustrate its pitfalls", and if by this he means all use of *p*-values, then I am afraid I must disagree. *p*-values are just too familiar and useful to ditch (even if it were possible).

But we can agree on scepticism about formal or informal rules that mechanically dichotomise findings into "significant" and "non-significant", and which can apply equally to rigid interpretation of intervals. Fortunately, Neyman and Pearson's decision-theoretic idea of "accepting the null" has just about been consigned to the overflowing dustbin of inappropriate scientific ideas, even if it lingers on in the misinterpretation of a "non-significant" result. Could we ditch "significant" as a similar anachronism? Sadly I think not, due to the habit of use and the lack of an alternative (apart from anything else, it would mean renaming this magazine). So, what are we left with? I have some personal opinions.

While there is not one universal solution, I believe it should be possible to establish firm general principles which focus on what is right rather than what is wrong. Then more specific guidance for different disciplines, to be enshrined in revised statistical education and statistical guidelines for journals and other outlets.

The crucial issue, identified by Berry, Gelman, Few and other commentators on the ASA statement, is to try and clearly separate (a) data description, (b) what it might be reasonable to believe in the light of this new evidence, and (c) categorical decisions and recommendations. *p*-values can have a role, although not be the sole determinant, at all stages. In particular, when describing data at stage (a) it may be fine to litter a results section with exploratory *p*-values, but these should not appear in the conclusions or abstract unless clearly labelled as such – perhaps by a specific notation p_{exp}.

A *p*-value should only be considered part of a confirmatory analysis at stage (b), and perhaps given the notation p_{con}, if the analysis has been pre-specified, all results reported, and *p*-values adjusted for multiple comparisons, and so on. Any p_{con}-values should be supplemented by informal, and even formal, Bayesian analysis that takes into account what else is known, the context, and in particular whether the null hypothesis or values close to it has any particular salience or plausibility, in which case Bayes factor arguments can be used to show the weakness of $p_{con} < 0.05$ and the need for higher thresholds.

But even if some agreement could be reached on a "positive" statement, then there is the problem of promulgating and enforcing it. At this point I get rather authoritarian. I believe that drawing unjustified conclusions based on selected exploratory *p*-values should be considered as scientific misconduct and lead to retraction or correction of papers. This requires both encouragement and training, but also publicly calling out journals, press offices and authors.

A colleague once told me of being confronted by a doctor at 4 pm on a Friday with "Could you just '*t* and *p*' this data for Monday?" While it would be wonderful if every analysis was going to be informed by someone skilled in statistical methodology, whether a nominal "statistician" or not, the rise of data science means that even more practitioners will be without a full professional training, and continue to do their *t*-ing and *p*-ing. We must do our best to help them. ■

■ **David Spiegelhalter** is chair of the Winton Centre for Risk and Evidence Communication at the University of Cambridge. He is currently president of the Royal Statistical Society, although this article is written in his personal capacity