STILL AGAINST INFERENTIAL STATISTICS: REJOINDER TO NICHOLSON AND RIDGWAY

STEPHEN GORARD
School of Education, Durham University
s.a.c.gorard@durham.ac.uk

PATRICK WHITE
Department of Sociology, University of Leicester
pkw4@leicester.ac.uk

ABSTRACT

In their response to our paper, Nicholson and Ridgway agree with the majority of what we wrote. They echo our concerns about the misuse of inferential statistics and NHST in particular. Very little of their response explicitly challenges the points we made but where it does their defence of the use of inferential techniques does not stand up to scrutiny. Their statements are either contradictory, agreement 'dressed up' as disagreement, appeals to authority, semantic slights of hand, or irrelevant to our original claims. It is not clear why such a response was needed.

Keywords: *Teaching statistics; Abuse of statistics; Inferential statistics; Significance testing*

EVERYTHING IS SETTLED IN PRACTICE

In our paper we claimed that significance tests were widely abused. They are routinely used in the wrong context – such as with population data, non-randomised cases, or samples where data or cases are missing. In almost all real-life research, analysts will be working with these kinds of data rather than with the simple random sample (with no non-response or dropout) that significance testing, and all other inferential statistical techniques, assume as a mathematically necessary basis. This means that significance tests should not generally be used in practice.

We also pointed out that even in a situation where these assumptions are met, the probability provided by NHST is the probability of the data given the null hypothesis $p(D|H_0)$ rather than the probability of the null hypothesis given the data we have $p(H_0|D)$. So, as we explained in our original paper, even when used 'correctly', significance testing does not generate a probability that provides the analyst with any useful information.

As Nicholson and Ridgway (NR) agree that NHST (and other inferential techniques) should not be used when the correct assumptions are not met and that significance testing can only produce $p(D|H_0)$, there should be little more to discuss. Researchers should not use, editors should not publish, and reviewers should not accept significance tests in research reports when the assumptions they require are not met. And even if these assumptions are met, the probabilities produced can rarely, if ever, be interpreted in a useful way. Because of this, significance testing and other inferential statistics should not have pride of place, if indeed they should have any place, in numeric analysis teaching. So what are NR's concerns that drove them to write a response?

Statistics Education Research Journal, 16(1), 74-79, http://iase-web.org/Publications.php?p=SERJ

NICHOLSON AND RIDGWAY'S RESPONSE

NR's response runs to 3,500 words, with much of this either agreeing with our original points or addressing issues that we did not raise (e.g., the section on "NHST and experimental design: sample size"). Shorn of these elements, there are only about 1,000 words in the response that are relevant to our paper. And even large parts of this are simply restatements of the authors' conviction that significance tests work, but with no demonstration, argument or evidence. For example (p. 66): "One of the strengths of using the NHST framework (assuming a well-designed study) is that the power function of the test, and effect size, taken alongside the *p*-value, give a good assessment of the likely robustness of the evidence collected […]." There are also areas of hidden agreement, made to look as though they were disagreements. For example:

> Contrast this with the account given by WG (p. 55): "As researchers, what we want to know is the probability of the null hypothesis being true (or false) given the data obtained, or $p(H_0|D)$. We want to know the probability that the difference or relationship that we observed in the sample (or experimental data) is due to the vagaries of random sampling (or allocation) rather than being a true reflection of data in our population." Researchers might want to know the probability of the null hypothesis being true [$p(H_0|D)$], but this is unknowable. The ASA statement on the interpretation of *p*-values makes it clear that NHST does not give $p(H_0|D)$.

This whole section merely repeats what we wrote. Significance tests do *not* provide good evidence on the probability of the null (or any other) hypothesis being true. That is why we have said that users are wrong to use the *p*-values generated by significance tests as though they assisted with judgements about the truth of the null hypothesis.

The response also includes a lengthy discussion of permutation analysis (based on a short aside in our paper). We do not intend to discuss this further in detail. PA is based on population data, and so does not suffer from the same problems as sampling theory derivatives. It does not, therefore, raise the same issues as significance tests, just as we said in our paper. Of course the allocation of cases must still be completely random, so even PA is of almost negligible use in practice (as above). There are only a few brief passages where NR respond directly to the substantive points made in our paper. Below, we show why each of these responses is incorrect.

WHAT ARE *P*-VALUES THE PROBABILITIES OF?

Given that NR agree with us about not using significance tests when the cases are incomplete or not randomised (i.e., every real-life study we have ever seen) the most fundamental and important point is this – they say (p. 66): "NHST makes no claims about $p(H_0|D)$, and so does not conflate conditional probabilities." Yet again, they agree with us that the *p*-value generated by a significance test is $p(D|H_0)$, and not $p(H_0|D)$. But then what use is $p(D|H_0)$? This is the key point they fail to address directly, because there is no sensible answer. In real-life, users of significance tests compute $p(D|H_0)$ to assess the likelihood of the null hypothesis, $p(H_0|D)$, being true. If this likelihood is low, they 'reject' the null hypothesis and assume that their finding requires an explanation other than chance. This is clearly what happens in practice and how significance testing is regularly taught. But using $p(D|H_0)$ in this way to assess the likelihood of the null hypothesis being true is invalid. The two probabilities are not the same, and one can be large while the other is small (see below). To use $p(D|H_0)$ in this way *is* a logical error.

NR try to disguise their defence of this logical error by referring to rejection of the null hypothesis as like asking (p. 65) "are any differences observed consistent with the

behaviour expected or observed with a plausible random mechanism? […] If not, look for another explanation, i.e., the alternative hypothesis."

The 'alternative hypothesis' is the alternative to the null hypothesis, and so looking for another explanation is exactly what most users do when they report 'rejecting' the null hypothesis. This is just NR playing with words. In order to assess whether to look at the alternative they must use $p(D|H_0)$, looking at the alternative only when $p(D|H_0)$ is small. And yet this is what they agree is fallacious when others do it. The only other way we can envisage them working is that they run a significance test, compute $p(D|H_0)$ and then ignore it! But that is no defence of significance testing either. We cannot look at what NR actually do in practice because neither of their websites lists an empirical study using significance tests. Like others who insist that *p*-values are useful if interpreted 'correctly', they cannot point to an example where this has actually been done. It is hard to understand what NR's objection really is, as their explanation is contradictory. For example, they say (p. 65):

> We disagree with the second sentence – this probability does not require an assumption that H_0 is true – it is the probability of obtaining these data, or data less likely than the data actually observed, *if* H_0 is true. There is an important distinction here in the logic […].

This says that the *p*-value generated by a significance test is based on assuming that the null hypothesis is true ("if H_0 is true"), but at the same time it says that "this probability does not require an assumption that H_0 is true". This is an internally contradictory sentence. The *p*-value depends on the null hypothesis being true. The null hypothesis does not *have* to be true, of course (and probably is not in many real-life scenarios). But if it is not true, then the *p*-value based on it being true no longer has any value.

As NR note, we illustrate this problem in our teaching by presenting students with a bag filled with 100 marbles of two colours. If we pull 10 marbles from the bag, the colour of these marbles provides us with no information about proportion of the colours of the 90 marbles left in the bag. This fact is obvious to all students watching the demonstration. If you had to estimate the ratio of the colours of marbles left in the bag, you might predict that this is the same as the ratio in the 10 cases you know about. But no *exact* probabilities can be calculated based only on this information and, of course, no significance testing would be involved in making this type of prediction.

As an alternative illustration, we tell the students what the ratio of marbles was in the bag before any were taken out, and all of them can quickly see that it would be possible to calculate the probability of obtaining the proportion of colours we did in our sample of 10. This is what a significance test does, but it relies on us *knowing* or at least *assuming* what is in the bag. If we know what is in the bag, then the significance test is pointless. If we do not know this, we have to make an assumption, and if that assumption is not true then the significance test approach *cannot* tell us what is in the bag. This is clearly an intractable problem and students are generally quick to understand this.

If we simply make up an assumption as our null hypothesis, such as that there are 50 marbles of each colour, then the *p*-values we calculate for any sample of marbles will be the same whether there are actually 50 of each, or whether they are distributed 70:30 or 30:70 or any other valid figure. Its accuracy depends on the accuracy of our assumption, which is the very thing we are trying to test. This is why *p*-values are useless in practice. The samples involved and so all calculations based on them (whether for *p*-values or confidence intervals) contain no information about the marbles remaining in the bag. Unfortunately there is no way around this problem, however much we would like there to

be. And it is this problem that undermines not just significance testing but all related inferential statistical techniques.

What NR are doing when they treat $p(D|H_0)$ as being a clue to the size of $p(H_0|D)$ (and so justifying looking for an 'alternative hypothesis' in their words) has been termed the base rate fallacy. Unfortunately, they ignore the unconditional likelihood of the null hypothesis being true in the first place – p(H) as opposed to $p(H_0)$. The following example demonstrates the problem with this practice.

The US Institute of Education Sciences has funded a large number of randomised control trials and found that just over 10% of the interventions involved were effective. Accepting this figure for illustrative purposes, this suggests that if 1,000 different trials were conducted around 100 should be found to be effective, and about 900 ineffective. If the sample sizes for the trials were large enough that 80% were correctly identified as effective based on using 5% significance as the criterion for 'effective' then 80 of the 100 interventions would be identified as useful (80% is the 'power' here, in traditional significance testing parlance). But 5% of the other 900 trials would also be identified (incorrectly) as being effective. This means that at least 36% (45/(45+80)) of the results treated as significant would be incorrect. This is far larger than 5%, and is even considerably larger than the 20% allowed by the supposed power calculation. The same result appears more formally if Bayes' theorem is used (see Colquhoun, 2014, for a more detailed explanation of this problem).

So, the proportion of incorrect positive results obtained when using significance tests can be far higher than usually envisaged, because that proportion depends on the independent probabilities of H_0 and D, which no one performing a significance test actually knows. If p(H) is lower than 10% then the level of incorrect results from significance testing increases dramatically. For example, an intervention – using the same 80% power and 5% alpha levels – that was actually only 1% likely to be effective (or true) then over 86% of the positive results from significance tests would be incorrect. And since an analyst using a significance test in the traditional manner has no idea of, and takes no account of, p(H) this means that their 'significant' findings are very likely to be invalid unless they are dealing with something that is already very likely to be true. It is no wonder that there is a replicability crisis (although the use of significance tests is not the only reason for this).

THE WEIGHT OF 'AUTHORITY'

In the original paper we noted that many commentators agreed that ISTs do not work as intended. NR responded by saying that a recent ASA statement does not explicitly acknowledge this and that the organisation seeks to 'advance' and 'improve the practice of' statistical inference. Our original point was merely intended to acknowledge that while we recognise that our stance on NHST and inferential statistics is far from universal, we are by no means the first to make these points. But appeals to authority are, in any case, a weak basis for deciding on the truth of a claim. We, like others before us, have restated the logical impossibility of transforming the information provided by ISTs – used according to conventional practice – into something useful for researchers. As no one seems to be able to provide a solution to this problem, or even an example of 'correctly' interpreted outputs of IST in practice, it would not matter if we were the first, or only, commentators to raise this point.

It is perhaps unsurprising that an organisation of professional statisticians has not yet abandoned significance testing. Many ASA members will have invested a great deal in the development and refinement of inferential statistical techniques, and the current ASA

position is likely to be a compromise between those members who are highly critical of significance testing and those who still do not realise the extent of the problems that underlie its use. We are confident, however, that the tide is turning in the right direction – albeit more slowly than we would like.

There are many reasons to be optimistic. The Institute of Education Sciences advises against use of significance tests, as does the American Psychological Association. Prestigious journals in epidemiology, public health, ecology, education and psychology have already banned their use. Key figures such as Berk, Cohen, Freedman, Glass, Jeffreys, Meehl, Rozeboom and Tukey have written critically about significance testing for decades and there has been a long history of writers calling for significance testing to be banned altogether – including Bakan (1966), Meehl (1967), Morrison and Henkel (1970), Walster and Cleary (1970), Carver (1978), Guttman (1985), Berger and Sellke (1987), Loftus (1991), Falk and Greenbaum (1995), Nester (1996), Schmidt (1996), Hunter (1997), Daniel (1998), Nix and Barnette (1998), Tryon (1998), Nelder (1999), Nickerson (2000), Fidler, Thomason, Cumming, Finch, and Lehman. (2004), Lipsey et al. (2012), Cumming (2014), and many others.

The weight of evidence is certainly against significance testing. No one can explain away the logical problem that dogs the interpretation of *p*-values. And no one has yet provided us with an example of their 'correct' use in empirical research, yet examples of abuse are commonplace. The fact that organisations such as the ASA, APS and BPS are becoming more sceptical about their use is presumably evidence of a change of culture among the organisations representing major stakeholders. There has certainly been less progress in the recognition that other outputs of inferential statistics – such as SEs and CIs – share the same problems as *p*-values, both in terms of the assumptions underlying their use and the logical problems with their interpretation. As with their responses relating to *p*-values, however, NR provide nothing to suggest that our concerns are unfounded. We believe that once the problems with NHST become widely accepted, the flaws inherent in these other outputs will soon be recognised as condemned by the same illogic and abuse.

CONCLUSION

Inferential techniques continue to be at the 'core' of most statistics courses and their inappropriate use in research reports is standard practice. The dominance of these techniques in the curriculum is so complete that many students (and some researchers) cannot separate inferential procedures from the descriptive and modelling ones underlying them. This distinction is certainly not made clear in most textbooks, where the assumptions underlying the use of inferential tests are also downplayed or even ignored.

Although academic communities have become more critical of the use of inferential statistics – especially over the last two decades – this appears to have had only a very limited impact on what is taught in classrooms, included in textbooks or reported in published empirical outputs. Given the reluctance of many established researchers to abandon the use of these flawed techniques, it is all the more important to ensure that a new generation of researchers becomes aware of the problems – both theoretical and practical – underlying their use. Continuing to teach something to new researchers that is so widely abused, almost entirely misunderstood, and dangerous when used in real-life applications, is worse than a waste of curriculum time. It is the perpetuation of bad practice that hinders the conduct of thoughtful and robust analysis. There are many more useful and important elements of statistical analysis to cover instead. It is time to change the focus and direction of statistics teaching.

REFERENCES

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*(6), 423–437.

Berger, J. O. & Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of *p* values and evidence (with comments and rejoinders by the authors). *Journal of the American Statistical Association, 82*(397), 112–39.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*(3), 378–399.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of *p*-values. *Royal Society Open Science, 1*, 1–16. [Online: doi.org/10.1098/rsos.140216]

Cumming, G. (2014). The new statistics: why and how. *Psychological Science, 25*(1), 7–29.

Daniel, L. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools, 5*(2), 23–32.

Falk, R. & Greenbaum, C. W. (1995). Significance tests die hard: the amazing persistence of a probabilistic misconception. *Theory and Psychology, 5*(1), 75–98.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine. *Psychological Science, 15*(2), 119–126.

Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis, 1*, 3–10.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8*(1), 3–7.

Lipsey, M., Puzio, K., Yun, C., Hebert, M., Steinka-Fry, K., Cole, M., Roberts, M., Anthony, K., & Busick, M. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. Washington DC: Institute of Education Sciences.

Loftus, G. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology, 36*, 102–105.

Meehl, P. (1967). Theory-testing in psychology and physics: a methodological paradox. *Philosophy of Science, 34*(2), 103–115.

Morrison, D. E. & Henkel, R. E. (1969). Significance tests reconsidered. *The American Sociologist, 4*(2), 131–140.

Nelder, J. A. (1999). Statistics for the millennium: from statistics to statistical science. *Journal of the Royal Statistical Society: Series D* (*The Statistician), 48*(2), 257–269.

Nester, M. R. (1996). An applied statistician's creed. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 45*(4), 401–410.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 5*(2), 241–301.

Nix, T. W. & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools, 5*(2), 3–14.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychological Methods*, *1*(2), 115–129.

Tryon, W. W. (1998). The inscrutable null hypothesis. *American Psychologist*, *53*(7), 796.

Walster, G. W. & Cleary T. A. (1970). A proposal for a new editorial policy in the social sciences. *The American Statistician, 24*(2), 16–19.

STEPHEN GORARD
School of Education, Durham University
Leazes Road, DH1 1TA, UK