Perpetuating the 'preposterous' use of significance at the expense of better science

Stephen Gorard

My brief paper on the abuse of statistics ('abuse' being a synonym for 'misuse' in the Oxford English Dictionary, and explained as meaning 'to use mistakenly or for a bad purpose') argued three main points.

1. Once random sampling variation is eliminated as a possible explanation for any apparent finding, analysts need to focus on alternative explanations based on design, measurement, bias, and attrition rather more than they do now.
2. The widespread use of probability calculations such as significance tests for eliminating random sampling variation as a possible explanation is based on a misunderstanding.
3. The kinds of probability calculations involved in significance tests and confidence intervals cannot be used with non-random cases, such as convenience samples and population data, anyway.

It is heartening to read support for these ideas in the responses from such a high-profile influence in the area (Gene Glass, from whom I have borrowed the term 'preposterous') and an emerging influence on methods in social science (Patrick White, the only respondent to engage fully with the second point above). Their response papers take these points forward and provide elegant examples with both clarity and some subtle humour. Yet, in a sense I cannot understand why *everyone* does not comprehend and support these three relatively simple points. They were not made by me in a spirit of 'purism' (Putwain), or of 'iconoclasm' (Howe). These three points are simply the truth of the matter. They have long been recognised as the truth by some before us, and papers and books have even been written to try and explain the political, financial and career reasons why others continue to ignore their obvious truth.

A few important misunderstandings

It is intriguing that the four respondents who wish to retain the use of significance etc. all focus on the third point above, because this is surely the hardest to defend against. The computation of significance, standard errors and confidence intervals (and the associated algorithms, now hidden by software) are clearly predicated on true random sampling. The fact that the software still operates and provides an 'answer' when the data does not come from randomisation and contains no probabilistic uncertainty is merely an illustration of the well-known garbage-in garbage-out principle. I guess that for these respondents to accept the truth (and in clearer terms than Putwain's 'technically correct', and the double-negative 'would not disagree', or Styles' 'not strictly justified', for example) would mean the end of statistics as it is practiced, the end of purported expertise in these methods, and the casting of doubt over prior work.

Howe states resolutely 'I do not agree that sampling theory techniques should be deemed inappropriate just because the sample was based on opportunity, convenience or snowballing'. To some extent this is merely a description of current practice and so many researchers must agree with Howe. But to see it so blatantly in print like that is truly shocking. It is the exact opposite of everything that appears in reputable methods texts, and it is the kind of error that leads Glass to say 'The fiction that probability statements are meaningful in the absence of random acts underlying them is preposterous'.

Common responses when I write or lecture about the abuse of statistics are 'everyone does it', 'it has happened for a long time' and eventually 'we already know all this but what should we do instead?'. I hope readers can see that none of these is a valid counter-argument, and that they remain invalid when deployed by four respondents here who largely re-state their own existing practice. Styles claims that in rejecting the use of significance and CIs I am rejecting any attempt to consider uncertainty in research findings. This is not true, and the original paper urges researchers to consider a wider and more important range of factors that lead to uncertainty but which are ignored by the significance approach (such as design bias or respondent attrition). I feel I am the more concerned because I do not just want to pretend I am assessing uncertainty via an invalid technique.

Howe claims that I argued that we should not use convenience samples and quotes APA guidance suggesting that convenience samples are perfectly proper. They are, and I never suggested otherwise. In fact, I clearly stated that we often have no practical alternative. What APA does not say is that we should use significance tests with convenience samples. As ever, it is presumably easier to mis-portray what I said and argue with that. Van Daal and Ader do something similar. I showed that denying the consequent is only valid in logic when the premises are certain, and that the modus tollens argument fails once any premise is uncertain or probabilistic. They portray this as me saying that probabilistic argument in general is invalid, including weather forecasting as an illustration. But weather forecasting does not employ this 'denying the consequent' argument structure at all. These three examples show the lengths that commentators have to go to in order to try and defend the indefensible.

There are some less common variants in the responses that try to maintain the edifice of significance testing. If we have a non-random sample we could randomly sample from within that and then use significance with the sub-sample (Van Daal and Ader). This seems truly desperate. Purportedly, there are techniques to 'fix' a non-random sample and make it back into a random one (Van Daal and Ader). No there are not, because if we knew the key values for the missing cases then we would have a complete sample. If not, we can only use the values we do have to make up for what is missing, so enhancing the bias caused by the missing cases in the first place (Gorard 2013). The same practical problem eliminates Putwain's suggestion that if the achieved sample looks similar to what the random sample would have been if available, then using significance is justified. To imagine a random sample based on a convenience sample, and then try to compute real probabilities accurately based on that imagination is surely incorrect.

Even stranger is the notion that the super-, hyper- or virtual population invented to help differentiate between theoretically finite and infinite populations can then be used to justify treating actual population data as a random sample (Styles). Just envisage what Styles means when he writes 'we imagine the trial being run many times on students in the same schools at the same time in a virtual population from which we did sample randomly'. And note that this entirely ignores the logical problems raised by point two at the outset. I have written about this absurdity many times before (e.g. Gorard 2008), and White handles this briefly but well in his response.

The search for an alternative or what to do 'instead' of significance, especially with non-random samples/allocation, is an odd one. Since the existing approaches do not work we must abandon them. A Bayesian approach would certainly be more logical but is no panacea and no substitute for judgement. As my original paper outlined, we do not need alternatives as such since we should be considering all competing substantive, design and methods explanation for any apparent finding anyway (even if we want to eliminate chance first). But significance testing has

somehow come to replace such real analysis, perhaps because the latter is not push-button. Nowhere is this more apparent than in the Carr and Marzouq (2012) paper cited in my original paper. I would have thought that my views on that paper were plain, and that once the goobledegook is removed nothing of scientific value is left. But I do not intend to pursue this further. That paper was merely chosen as a recent example to represent hundreds similar in this journal and the many thousands in journals worldwide.

The abuse of significance is a big problem

I did not, in my brief article, explain just how widespread the abuse of sampling theory with non-random samples is, nor how often the results of statistical analyses are poorly reported. I assumed that respondents like van Daal and Ader would know (and realise the damage that ensues). They do tell readers that 'convenience samples dominate in the social sciences' – which surely means that significance tests should hardly ever be encountered, as their computation depends entirely on prior randomisation. However, these tests are still widely encountered, as van Daal and Ader should and probably do know, and as Putwain helpfully illustrates via consideration of articles in the Journal of Educational Psychology. None of the 24 articles involving numbers was based on random samples, and all of them used inferential statistics (incorrectly). This does not surprise me and I have done similar analyses of education journals and found the same thing (e.g. Gorard 2008). The problem is so widespread that it is almost universal, and perhaps in some strange way that makes it hard for van Daal and Ader to notice. Similarly, although the meaning of significance tests is often correctly described in methods texts, these tests are then generally misused in the examples and in research practice. The error is now usually only implicit since analysts simply take the probability of the data observed given the truth of the null hypothesis as being the same as or closely related to the probability of the null hypothesis being true given the data observed. That is, they 'reject' the null hypothesis where the probability of the data observed under that hypothesis is low, and they do so without explanation or justified argument.

This key error of confusing pHyp|Data with pData|Hyp is again almost universal. That is why I proposed asking all researchers to explain the steps in the 'logic' they are using explicitly, since once they have written the argument down in clear (if they can) it should be obvious to them and their readers that the argument is invalid (see the worked example in Chapter 5 of Gorard 2013). Significance tests just do not work as intended by their users, even when applied to random sampling/allocation. The situation is even worse for confidence intervals because I have yet to encounter an analyst who can correctly explain what CIs mean. CIs are as bad as p-values since they "suffer from similar flaws to p-values, exaggerating both the size of implausible effects and their significance" (Matthews 1998, p.5), yet they are even harder to describe. And why should the fact that 95% of hypothetical repeated sample figures lie within 1.96 population standard deviations of the population average then imply that the population average will lie within 1.96 one-sample standard deviations of a specific one-sample average 95% of the time? It will not and it does not, as anyone can see if they think about it clearly. Please let us not worry about whether it is a 'standard deviation' or a 'standard error', a 'population' or a 'sampling distribution'. Let the reader insert their terms of choice, and the argument of CIs remains just as clearly nonsense.

These problems are not just commonplace, they are also dangerous. At best, they make research reports harder to read, perhaps confusing readers, and certainly wasting people's time with producing, publishing and consuming fake results. In a worse case, they waste public funding, causing needless opportunity costs, and they waste people's energy pursuing what turn out to be

all too easily predictable vanishing breakthroughs. At worst, these errors damage lives and kill people (see examples from diet, cancer research, and heart treatments in Matthews 1998).

> "70 years ago Ronald Fisher gave scientists a mathematical machine for turning baloney into breakthroughs, and flukes into funding" (Matthews 1998, p.1).

This is especially shocking because research in the UK is largely funded by the public (taxpayers and charity-givers). Where social science has impact in practice, the effect is largely on the public in areas of policy like education, crime, housing, transport and health. Yet ethical committees and guidelines still largely ignore the interests of the wider public in their focus on possible harm to the researchers and the researched (Gorard 2002). I urge Putwain to pursue the implications of this second principle of ethics, now creeping into ethical guidelines such as those of the SRA. Public money is being wasted, and public lives are being made worse (or at least not improved as much as they could be), by this invalid practice of significance testing. To call it merely a 'cult' is to downplay its importance. It should cease. Now.

References

Gorard, S. (2002) Ethics and equity: pursuing the perspective of non-participants, *Social Research Update*, 39, 1-4

Gorard, S. (2008) *Quantitative research in education: Volumes 1 to 3*, London: Sage

Gorard, S. (2010) All evidence is equal: the flaw in statistical reasoning, *Oxford Review of Education*, 36, 1, 63-77

Gorard, S. (2013) *Research Design: Robust approaches for the social sciences*, London: SAGE

Matthews, R. (1998) *Bayesian Critique of Statistics in Health: The great health hoax*, http://www2.isye.gatech.edu/~brani/isyebayes/bank/pvalue.pdf