The widespread abuse of statistics by researchers: what is the problem and what is the ethical way forward?

Stephen Gorard
Durham University
s.a.c.gorard@durham.ac.uk

Abstract

The paper presents and illustrates two areas of widespread abuse of statistics in social science research. The first is the use of techniques based on random sampling but with cases that are not random and often not even samples. The second is that even where the use of such techniques meets the assumptions for use, researchers are almost universally reporting the results incorrectly. Significance tests and confidence intervals cannot answer the kinds of analytical questions most researchers want to answer. Once their reporting is corrected, the use of these techniques will almost certainly cease completely. There is nothing to replace them with but there is no pressing need to replace them anyway. As this paper illustrates, removing the erroneous elements in the analysis is usually sufficient improvement (to enable readers to judge claims more fairly). Without them it is hoped that analysts will focus rather more on the meaning and limitations of their numeric results.

Which kind of statistics is being abused?

The term 'statistics' is an ambiguous one. It emerged from the collation and use of figures concerning the nation state from the seventeenth century onwards in the UK, and subsequently in the USA and elsewhere (Porter 1986). Such figures involved relatively simple analyses, and 'political arithmetic' was largely used to lay bare inefficiencies, inequalities and injustice (Gorard 2012). However, more recently and for many commentators the term has come to mean a set of techniques derived from sampling theory, and/or the products of those techniques. It is the abuse of such techniques that is the subject of this new paper. These techniques include the use of standard errors, confidence intervals and significance tests (both explicitly and disguised within more complex statistical modelling). They are supposedly used to help analysts to decide whether something that is found to be true of the sample achieved in a piece of research is also likely to be true of the known population from which that sample was drawn. All of these statistical techniques, including confidence intervals, are based on a modified form of an argument *modus tollendo tollens*. In formal logic, the argument of denying the consequent is as follows:

> If A is true then B is true
> B is not true
> Therefore, A is not true also

This is a perfectly valid argument, and the conclusion must be true, as long as the premising statements are all definitive. If B is not true then it is certain that A is not true. However, as soon as tentativeness or probability enters the argument fails:

> If A is true then B is probably true also

B may not be true
Therefore, A may not be true also

This is not really a valid argument, and the truth of the conclusions is contingent on many factors beyond pure logic. Characteristic A may or may not be true. If it is true, characteristic B could be true as well, or not. The observation that B may (or may not) be true says almost precisely nothing about the truth of A. The first premise may be likened to the null hypothesis in statistical analysis, and the second to the evidence from the achieved research sample. The probabilistic argument is now contingent upon the frequency with which A and B are true together in reality, and on the accuracy of the research finding about the likelihood of B being true. Knowing both of these facts, it would be possible to draw a probabilistic conclusion about the likelihood of A being true (the desired research conclusion). But in reality, neither of these facts would be known. In fact, the main supposed objective of the analysis would be to help decide on the accuracy of the research finding that B may not be true. The analysis assumes from the outset something about that which it is supposed to be assessing. The misunderstanding caused by this assumption is widespread.

Using sampling theory techniques in inappropriate contexts

However, the most obvious abuse of sampling theory techniques is their use in situations for which they were not designed and for which they ought not to be used. Data for a population cannot have a standard error, by definition. The standard error is defined as the standard deviation of a random sampling distribution, of samples drawn repeatedly from a population. It is used (but incorrectly, see below) to try and estimate the proximity of the sample mean to the population mean. When working with population data the population mean is known, therefore such an estimation is neither needed nor valid. Of course, the population data may be incomplete due to missing cases or missing values, but this is a cause of bias not a consequence of random sampling variation. Bias ought to be addressed in any analysis (although it rarely is addressed by those who use 'statistics' instead) but it cannot be addressed through significance tests and the like. None of the techniques of sampling theory statistics can or should be used with population data. When commentators like Goldstein (2008, p.396) advocate the use of confidence intervals with population based data, they are betraying ignorance of the meaning of confidence intervals (see below), misleading policy-makers and other researchers, and harming those who will be affected by supposedly evidence-informed decisions.

Exactly the same applies to samples other than the random samples on which sampling theory techniques such as significance tests are based (Fielding and Gilbert 2000). Opportunity, convenience, snowball samples and the like also do not have a standard error, by definition. Findings derived from such samples have no probabilistic uncertainty; they will just have bias. In the same way that findings from population data can be tentatively generalised to other cases not in the population, so findings from non-random samples can be generalised to other cases. But in both situations the generalisation can only be based on judgement, and how well the sampled cases match the non-sampled ones in terms of what it already known. In reality, the judgement is not a generalisation *from* the sample (or population) but a decision about what is already known about non-sampled cases and how well they match the sampled ones. None of this concerns random sampling variation. When researchers like Carr and Marzouq (2012), to take just one of many available examples, cite significance tests and p-values derived from two complete classes of children in one primary school, they are making

a key analytical error. Their probabilities cannot mean anything in the context where only a convenience sample of a year group from one school is involved. Even if their results had been based on a random sample, the statistical population which such results could be generalised to does not exist outside the sample. Such abuse of statistical techniques simply has to cease. As with Goldstein's use of confidence intervals for population data, such abuse of non-random samples leads to errors, wasted opportunities, vanishing breakthroughs, and unwarranted conclusions.

The final situation for this first kind of abuse is when random samples are planned but not achieved. Strictly speaking an incomplete random sample is not a random sample at all. Rolling 1,000 unbiased dice to estimate the probability of gaining each outcome would be a (pseudo-)random process. Rolling the dice and then re-rolling any that showed a 6 would not lead to a good estimate of the probability of gaining each outcome. This is obvious. In the same way, selecting 1,000 cases by chance from a known population is very different from selecting 1,000 cases and then replacing 100 of these because they refused to participate. This means that in almost all real-life research situations, sampling theory statistical techniques are not relevant, do not mean anything and must not be used. In a sense, the paper could end at this point, because it would be rare for an analyst to be dealing with a complete random sample.

Misunderstanding and misrepresenting the outputs of significance tests

However, there is a second kind of widespread abuse of statistics that is even worse but somewhat harder to explain. This is because there is such a common misunderstanding of this form of analysis. Put simply, statistical analysis even when conducted appropriately and with all underlying assumptions met does not do what most analysts want and what many methods instructors portray that it does. The nature of the conditional probabilities involved is commonly and mistakenly reversed, whether through incompetence or intention to deceive.

This confusion between the probabilities for a sample and a population is clear in the logic of significance testing and the quotation of p-values. As with the modus tollens argument above, a significance test assumes from the outset that what is being 'tested' is true for the population, and so calculates the probability of obtaining a specific value from the random sample achieved (Siegel 1956). Analysts then generally mistake this (pData|Hyp) as being the probability of what is being 'tested' also being true for the population, given the value obtained from the random sample achieved (pHyp|Data). These two probabilities are clearly very different, and neither can be safely inferred from the other. One may be small and the other large, or vice versa, or any combination in between (Gorard 2010). The p-value calculation depends on the initial assumption of a null hypothesis about what is true for the population. As soon as it is allowed that the null hypothesis may not be true, the calculation goes wrong. The actual computation for a significance test involves no real information about the population, and this means that the same sample from two very different populations would yield the same p-values. A sample mean of 50 would, quite absurdly, produce the same p-value if the population mean were 40, 50, 60 or 70 etc. This is because the population value is not known (else there would be no point on conducting the significance test), and the entire calculation is based only on the achieved sample value.

To illustrate the common misunderstanding of this, consider a simplified situation. There is a bag, containing 100 well-shuffled balls of identical size, and the balls are known to be of only

two colours. A sample of 10 balls is selected at random from the bag. This sample contains 7 red balls and 3 blue balls. The analytical question to be addressed is: how likely is it that this observed difference in the balance of the colours between the two samples is also true of the original 100 balls in each bag? The situation is clearly analogous to many analyses reported in social science research. The bag of balls is the population, from which a sample is selected randomly. A moment's thought shows that it is not possible to say anything very much about the other 90 balls in the bag. The remaining 90 might all be red or all blue, or any share of red and blue in between. Yet the purpose of such a significance test analysis is to find out via sampling something about the balance of colours in the bag. Without knowing what is in the bag there is no way of assessing how improbable it is that the sample has ended up with 7 red balls. Once this impossibility is realised, the pointlessness of significance testing becomes clear.

What a significance test does instead is to make an artificial assumption about what is in the bag. Here the null hypothesis might be that the bag contains 50 balls of each colour at the outset. Knowing this it becomes relatively easy to calculate the chances of picking 7 reds and 3 blue in a random sample of 10. If this probability is small (traditionally less than 1 in 20, or 0.05) it is customary to claim that this is evidence that the bag must have contained an unbalanced set of balls at the outset. This claim is obviously nonsense. The assumption of the null hypothesis tells us nothing about what is actually in the bag. For example, imagine that the bag started with 80 red balls and 20 blues. The sample is drawn as above, and contains 7 reds. The significance test approach assumes that there are 50 reds in the bag and calculates a probability of getting 7 in a sample of 10. This probability will be clearly incorrect because the balls are less balanced in fact than the null assumption requires. Now imagine that the sample is still the same but that the bag had 80 blue balls and only 20 red originally. The significance test approach again assumes that there are 50 reds in each bag and calculates the same probability of getting 7 red from one and 5 from the other. This probability will also be clearly incorrect because the balls are less balanced than the null assumption requires. More absurdly, this second probability *must* be the same as the first one since they are both calculated in the same way on the same assumption. So the significance test would give exactly the same probability of having drawn 7 reds in a random sample from a bag of 80% reds as from a bag of 20% reds. This absurdity happens because the test takes no account of the actual proportion of each colour in the population. It cannot, since finding out that balance is supposed to be the purpose of the analysis.

Of course the probability of getting 7 reds from a bag containing 80 reds is different, a priori, to the probability of getting 7 reds from a bag containing 20 reds. But the significance test is conducted post hoc. There is no way of telling what the remaining population is from the sample alone. To imagine otherwise, would be equivalent to deciding that rolling a 3 followed by a 4 with a die showed that the die was biased (since the probability of that result is only 1/36, which is much less than 5% of course).

For anyone who has spotted this misunderstanding, there is little doubt that their use of significance testing would cease (Falk and Greenbaum 1995). No one wants to know the probabilistic answer the tests actually provide (about the probability of the observed data given the assumption), and the test cannot provide the answer analysts really want (the probability of the assumption being true given the data observed). This conclusion is not new (Harlow et al. 1997). It has been known for a long time, perhaps since their earliest adoption, that significance tests do not work as hoped for, and may well be harmful because their results are so widely misinterpreted (Carver 1978). Yet unwary methods resources and

purported experts continue to peddle the fiction that p-values are, or are closely related to, the probability of the sample result being 'true', real or relevant. Relatively recent examples among many include the following in a textbook on social science methods:

> [Statistical significance is] 'the likelihood that a real difference or relationship between two sets of data has been found' (Somekh and Lewin 2005, p.224).

And perhaps even more worrying is the 'explanation' (in relation to statistical modelling) given during the training of heavily selected UK national experts in rigorous evaluation:

> Significance of b_4 indicates whether there is evidence of an interaction effect (Connolly 2013, slide 5).

Both of these explanations are the wrong way around. The 'significance' value is really the likelihood of finding a fake 'difference' or 'effect' if none actually exists. This is a very different value to the likelihood of there actually being a difference or effect. It is like saying the probability of being a professional footballer if a person is over six feet tall is the same as the probability of a person being over six feet tall if they are a footballer. The first of these values will be much, much smaller than the second. To confuse the two as the supposed experts above do is to make a very serious mistake. It *is* possible to convert one figure to the other using Bayes' theorem, as long as the unconditional probabilities are already known (such as what proportion of people are footballers and what proportion are over six feet tall). But there would be no point in conducting a significance test in this situation since both conditional probabilities would be calculated precisely.

Misunderstanding and misrepresenting confidence intervals

Faced with increasing criticism of significance testing and its abuse, in 1999 the American Psychological Association (APA) set up a Task Force on Statistical Inference. This considered a ban on the reporting of such tests in all APA journals. Unfortunately, their final recommendation fell short of such a radical but useful step, and APA instead focused on moving beyond significance to a consideration of the 'precision' of any research findings. Its influential publication manual now states that:

> [Null hypothesis significance testing] is but a starting point and that additional reporting elements such as effect sizes, confidence intervals, and extensive description are needed (APA 2010, p. 33).

This is a shame because confidence intervals use the same underlying logic as significance tests, share the same fatal flaws, and are at least as widely misunderstood. For example, talking about confidence intervals, Goldstein (2008, p.399) says of their use in value-added calculations:

> A confidence interval provides a range of values that, with a given probability – typically 0.95 – is estimated to contain the true value of the school score.

Connolly (2007, p. 149) says that a 95% confidence interval shows that:

> There is a 95 percent chance that the true population mean is within just under two standard errors of the same mean.

Both of these statements are wrong. With population data, or where the true population value (such as its mean) was already known, there would be no need for confidence intervals (CIs). A CI is calculated only from the sample value, and no reference at all is made to the true population value (how could it be?). Instead of the above, a CI for a sample value means precisely this:

> If we assume that the value from a complete random sample is identical to the true population value, then the CIs of many repeated complete random samples of the same size would contain the population value for 95% (or selected interval) of these samples.

This is why any reported CI for a specific sample is centred around the sample value. Of course, based on this correct definition (of how a CI is actually calculated) the technique is completely useless. It cannot be used to assess how close the sample value is to the unknown population value, because it is based on the assumption that the two are identical from the outset. As soon as it is allowed that the two might differ at all, then the calculation of the CI fails. If the sample mean is not at the precise centre of the normally distributed population (or sampling distribution) then it is not true that 95% of the population will lie within 1.96 standard deviations from the sample mean. The absurdity of this kind of artificial calculation is perhaps even clearer when considering what happens in an example. Imagine that a sample mean was 50, and that this was drawn from a population with mean 60. The CI would have a particular range centred around 50. Now imagine that all else remains the same but that the population mean was actually 70. The CI would remain the same because the CI is unrelated to the actual population mean. This suggests that a CI based on an estimate of 50 for a real value of 60 would imply the same level of accuracy as for a real value of 70. In practice, and even when used as intended, CIs are pointless. Worse than this, because even purported authorities are explaining their interpretation incorrectly, they are being used to draw invalid inferences. Again, money and research effort are being wasted and those intended to benefit from research may be being harmed. Simply stating the number of cases underlying any sample value is sufficient and valid.

What should happen instead?

There is a tendency to want to cling to traditional statistics, not understanding them or even knowing that they do not makes sense, due to not being sure what to do instead. In general, the answer is that nothing should be done instead. Removal of the error is improvement enough. In the paper used as an example above (among countless others), Carr and Marzouq (2012) present a Table 1 (p.7) as below and textual discussion of these findings (p.6):

> As seen in Table 1 children endorsed all four of the achievement goals to similar degrees. However, the range of responses for both the mastery-approach and mastery avoidance scales were narrower than the performance scales and were focused at the top end of the scale. Correlations between goals (Table 1) are consistent with the 2 x 2 framework where goals sharing a dimension (mastery/performance *or* approach/avoidance) are positively correlated while those not sharing a dimension are unrelated (Elliot & McGregor, 2001; Elliot & Murayama, 2008). Although this pattern

of correlation is evident in this sample the association between mastery approach and performance approach goals is smaller than expected, just approaching significance.

Table 1: Descriptives and intercorrelations for achievement goal responses.

Variable	*M*	*SD*	*Min*	*Max*	*Range*	*1*	*2*	*3*
1. Performance Approach	3.39	1.07	1.00	5.00	4.00	–		
2. Performance Avoidance	3.58	.88	1.67	5.00	3.33	.69** –		
3. Mastery Approach	4.51	.53	3.00	5.00	2.00	.20+	.10 –	
4. Mastery Avoidance	3.92	.89	3.00	5.00	3.00	.13	.41**	.42**

+p<0.06; *p<0.05; **p<0.001 (1-tailed).

Clearly, much of this reporting is incorrect. With a convenience sample of 58 children from one school, Carr and Marzouq (2012) should not be discussing statistical 'significance' or quoting p-values. Therefore, parts of the report such as the gobbledegook at the foot of the table can be simply removed. In addition, the use of decimal places should be curtailed. It is unlikely that the reported means are really accurate to five one thousandths of a unit in a study measuring things as vague as 'performance approach' with only 58 cases. The result could look like this

> As seen in Table 1 children endorsed all four of the achievement goals to similar degrees. However, the range of responses for both the mastery-approach and mastery avoidance scales was narrower than the performance scales and focused at the top end of the scale. Correlations between goals (Table 1) are consistent with the 2 x 2 framework where goals sharing a dimension (mastery/performance *or* approach/avoidance) are positively correlated while those not sharing a dimension are unrelated (Elliot & McGregor, 2001; Elliot & Murayama, 2008). Although this pattern of correlation is evident in this sample the association between mastery approach and performance approach goals is smaller than expected.

Table 1: Descriptives and intercorrelations for achievement goal responses.

Variable	*M*	*SD*	*Min*	*Max*	*Range*	*1*	*2*	*3*
1. Performance Approach	3.4	1.1	1.0	5.0	4.0			
2. Performance Avoidance	3.6	0.9	1.7	5.00	3.3	+0.7		
3. Mastery Approach	4.5	0.5	3.0	5.00	2.0	+0.2	+0.1	
4. Mastery Avoidance	3.9	0.9	3.0	5.00	3.0	+0.1	+0.4	+0.4

Nothing much has changed with the invalid p-values removed. If the findings of the paper were important (or not) before, they remain so now that they are reported without abusing

statistics. It is entirely possible that making the results simpler, and not misleading readers or even the researchers themselves with false probabilities, would encourage a greater emphasis on the analytical issues that really matter and on the substantive (or not) nature of the results. Key issues in this example appear to be whether the measures are measuring anything at all, whether they can measure it accurately, how they could be calibrated, what the bias might be in the sample, the nature of any non-response, and how any of these initial errors might propagate through ensuing calculations. The answers to these questions and others like them will help readers and researchers decide whether the results warrant the claim in the paper - that the researchers have *tested* 'the 2 x 2 achievement goal model' (p.6). Moving away from the convenient but invalid push-button approach to analysis might yield benefits beyond mere cessation of the abuse. It might introduce more transparency and judgement in reporting (Gorard 2006).

Conclusion

When an analyst is trying to decide on the substantive importance of an apparent research finding, they are faced with a number of alternative explanations. If they have used a random sample then one of these explanations is that the result is a fluke introduced by sampling variation. This is the explanation that significance testing, confidence intervals and associated statistical techniques are meant to address (but which they do not). However, this is only one explanation. Other methods-based explanations include design errors, bias in the sample, errors in measuring or recording data, researcher effects and so on. These other explanations ought to be considered and discussed whether the sample is a random one or not, or even if a population is involved. But the current abuse of significance testing seems to have replaced all other considerations. What should happen instead of the false logic of statistics is a greater focus on the meaning and authority of the evidence that analysts uncover, using transparent judgement to decide whether a difference is worth pursuing or whether a coefficient is worth retaining in a model. There are a number of simple techniques than can assist in making and portraying these judgements, including graphical displays, and a range of effect sizes from odds ratios to R^2.

Of course, none of the above is any kind of argument against measurement or the crucial role of numbers in social science research. This paper is rather an argument that researchers should take numbers more seriously, and think rather more carefully than at present about their meaning. Similarly, this is not an argument against the random selection of cases in a sample, or the random allocation of what is in effect population data to treatment groups in a trial design. Randomisation is the best protection against imbalance or bias in the sample or the groupings, both in terms of known characteristics that could be matched and in terms of the unknown characteristics that any attempted matching procedure is forced to neglect (Gorard 2013). But random sampling is used to minimise bias, not so that significance tests can be run. With a high quality random sample the best estimate of any equivalent value for the linked population will be the sample value. No amount of dredging with the sample data alone (as happens with standard errors, significance tests and CIs) can improve this estimate.

Real people, their lives, well-being, health and education are affected by research evidence-informed decisions in policy and practice. At present, these decisions are (unknown to policy-makers and practitioners) overly influenced by a superstitious ritual that few seem to understand but many seem happy to follow and pass on to new researchers. This ritual was described by Rozeboom (1997, p.335) as:

> Surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students.

As illustrated above, the techniques of sampling theory statistics do not work as intended, and can give misleading results leading to vanishing breakthroughs and even harmful interventions. It is time for this wasteful and dangerous nonsense to cease.

References

American Psychological Association (2010) *Publication manual of the APA* (6th ed.), Washington, DC

Carr, A. and Marzouq, S. (2012) The 2 x 2 achievement goal framework in primary school: Do young children pursue mastery-avoidance goals?, *The Psychology of Education Review*, 36, 2, 3-8

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

Connolly, P. (2013) *Analysis of Randomised Controlled Trials (RCTs)*, presentation to Conference of EEF Evaluators: Building Evidence in Education, London, http://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=4&ved=0CDwQFjAD&url=http%3A%2F%2Feducationendowmentfoundation.org.uk%2Fuploads%2Fpdf%2FSession_4_-_analysis_and_reporting.pptx&ei=ethoUtWZDsSl0QWW14GYCg&usg=AFQjCNFDqiLZLV5rnHSwUmMmlcPP8sDkaA

Connolly, P. (2007) *Quantitative data analysis in education*, New York: Sage

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

Fielding, J. and Gilbert, N. (2000) *Understanding social statistics*, London: Sage

Goldstein, H. (2008) Evidence and education policy – some reflections and allegations, *Cambridge Journal of Education*, 38, 3, 393-400

Gorard, S. (2006) Towards a judgement-based statistical analysis, *British Journal of Sociology of Education*, 27, 1, 67-80

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

Gorard, S. (2012) The increasing availability of official datasets: methods, opportunities, and limitations for studies of education, *British Journal of Educational Studies*, 60, 1, 77-92

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

Harlow, L., Mulaik, S. and Steiger, J. (1997) *What if there were no significance tests?*, Marwah, NJ: Lawrence Erlbaum

Porter, T. (1986) *The rise of statistical thinking*, Princeton: Princeton University Press

Rozeboom, W. (1997) Good science is abductive not hypothetico-deductive, in Harlow, L., Mulaik, S. and Steiger. J (Eds.) *What if there were no significance tests?*, New Jersey: Erlbaum

Siegel, S. (1956) *Nonparametric statistics for the behavioural sciences*, Tokyo: McGraw Hill

Somekh, B. and Lewin, C. (2005) *Research Methods in the Social Sciences*, London: Sage