All evidence is equal: the flaw in statistical reasoning
Gorard, Stephen

Link to publication on Research at Birmingham portal

All evidence is equal: the flaw in statistical reasoning

Stephen Gorard

The School of Education

The University of Birmingham

B15 2TT

s.gorard@bham.ac.uk

Keywords

Significance debate, statistical testing, conditional probability, methods paradigms

Abstract

In the context of existing 'quantitative'/'qualitative' schisms, this paper briefly reminds readers of the current practice of testing for statistical significance in social science research. This practice is based on a widespread confusion between two conditional probabilities. A worked example and other elements of logical argument demonstrate the flaw in statistical testing as currently conducted, even when strict protocols are met. Assessment of significance cannot be standardised and requires knowledge of an underlying figure that the analyst does not generally have and can not usually know. Therefore, even if all assumptions are met, the practice of statistical testing in isolation is futile. The question many people then ask in consequence is - what should we do instead? This is, perhaps, the wrong question. Rather, the question could be – why should we expect to treat randomly sampled figures differently from any other kinds of numbers, or any other forms of evidence? What we could do 'instead' is use figures in the same way as we would most other data, with care and judgement. If all such evidence is equal, the implications for research synthesis and the way we generate new knowledge are considerable.

All evidence is equal: the flaw in statistical reasoning

Abstract

In the context of existing 'quantitative'/'qualitative' schisms, this paper briefly reminds readers of the current practice of testing for statistical significance in social science research. This practice is based on a widespread confusion between two conditional probabilities. A worked example and other elements of logical argument demonstrate the flaw in statistical testing as currently conducted, even when strict protocols are met. Assessment of significance cannot be standardised and requires knowledge of an underlying figure that the analyst does not generally have and can not usually know. Therefore, even if all assumptions are met, the practice of statistical testing in isolation is futile. The question many people then ask in consequence is - what should we do instead? This is, perhaps, the wrong question. Rather, the question could be – why should we expect to treat randomly sampled figures differently from any other kinds of numbers, or any other forms of evidence? What we could do 'instead' is use figures in the same way as we would most other data, with care and judgement. If all such evidence is equal, the implications for research synthesis and the way we generate new knowledge are considerable.

Social science research and the schism

In the last decade or so, there has been considerable international discussion about the quality and relevance of social science research, especially in applied fields such as education, health promotion, and crime prevention. In my experience, progress towards enhanced research capacity sometimes founders on the entrenched schism between work that is purportedly 'qualitative' and that which is termed 'quantitative'. The barriers to improvement generated by this schism include demands for different kinds of criteria for judging the quality of different forms of evidence. The barriers to knowledge accumulation include demands for separate approaches to evidence synthesis for each type of data. Indeed, in some current approaches to systematic review, work that is less than a certain scale, or not of a particular kind, is simply

excluded from consideration. The outcome is that work involving large-scale numbers tends to be privileged in review and, at least rhetorically, in evidence-informed policy-making and practice. Ironically, this sense of privilege may then reinforce the schism for those whose work is excluded.

It has not been satisfactorily explained why evidence in the form of numbers should be treated differently, *per se*, to other forms of evidence; nor why textual, archival, audio, visual, and other sense data are deemed so similar as to be in one silo, distinct to numbers. There seems to be no philosophical or theoretical basis for the schism (Gorard 2004a, 2004b). As so-called 'mixed methods' studies show, there is no fundamental barrier to the successful combination of all of these data forms within or across studies. In fact, it is even possible to mount an argument that mixing methods is a natural way to conduct high quality relevant research of the kind that commentators have been asking for (Gorard and Cook 2007). This paper addresses one potential obstacle which appears to suggest that numbers should be routinely treated in a very different way to other forms of evidence – this obstacle is the domination of number-based work by a sampling theory approach derived from agricultural trials (Gorard 2006a).

A reminder of the basis for statistical testing

The basis of statistical testing derived from sampling theory is the calculation of a conditional probability. This probability becomes the p-value used for significance testing, and also for standard errors, confidence intervals and often in deciding which variables to retain in complex statistical modelling.[1] The calculation of the probability is grounded in various assumptions, such as a random or randomised sample and complete measurement of all cases in the selected sample (de Vaus 2002). There are two different widespread uses of significance tests (based loosely on the Fisher and Neyman-Pearson traditions). Apart from the terminology their logic, as far as this

---

[1] This paper focuses on significance testing. However, readers should recall that the same issues arise in calculating standard errors and confidence intervals – which are calculated on the basis of a null hypothesis and both of which are subject to similar misinterpretations as p-values. For example, the 95% confidence interval is often misunderstood as a bound within which we can be 95% confident that

discussion is concerned, is the same. The test is intended to help decide whether two sets of measurements (perhaps two sub-samples) could have come from the same overall group (population). Here is a fairly standard summary of this process, as it appears in methods resources.

The usual process of hypothesis testing consists of four steps.

1. Formulate the null hypothesis (commonly, that the observations are the result of pure chance) and the alternative hypothesis (commonly, that the observations show a real effect combined with a component of chance variation).

2. Identify a test statistic that can be used to assess the truth of the null hypothesis.

3. Compute the p-value, which is the probability that a test statistic at least as significant as the one observed would be obtained assuming that the null hypothesis were true. The smaller the p-value, the stronger the evidence against the null hypothesis.

4. Compare the p-value to an acceptable significance value (sometimes called an alpha value). If p<alpha, that the observed effect is statistically significant, the null hypothesis is ruled out, and the alternative hypothesis is valid. (Wolfram Mathworld - http://mathworld.wolfram.com/HypothesisTesting.html, 29/5/08)

I have several concerns about the way the process is described here, and everyone may have their own favoured way of expressing it, but that is not my main point. The steps are, in general terms, those agreed in resources and used in practice. An extended treatise of the nature and use of statistical testing says:

Application of NHST [null hypothesis significance testing] to the difference between two means yields a value of p, the theoretical probability that if two samples of the size of those used had been drawn at random from the same population, the statistical test would have yielded a statistic (e.g., t) as large or larger than the one obtained. (Nickerson 2000, p.242)

---

a specific population parameter falls, rather than the bound within 95% of sample values would fall *if* the null hypothesis is true.

Nickerson (2000, p.242) continues by saying of the null hypothesis that 'Often… the term is intended to represent the hypothesis of "no difference" between two sets of data'. If the p-value calculated as above is less than a certain level, traditionally 5%, then the null hypothesis is rejected.[2] Standard reputable texts on methods for statistical analysis are in agreement over the conditional nature of null hypothesis testing, even though they may all express this condition in different ways. Here is a selection of partial descriptions of the same process from a variety of sources

A traditional A-level text book says:

> If the experiment gives results which have a high probability on the null hypothesis the hypothesis will not be rejected. (Brookes and Dick 1951, p.148)

A more advanced text from the same period says (where H_0 is the null hypothesis):

> In advance of the data we specify the set of all possible samples that could occur when H_0 is true. From these, we specify a subset of possible samples which are so extreme that the probability is very small, if H_0 is true, that the sample we actually observe will be among them. (Siegel 1956, p.8)

A more recent, more general methods text says:

> The final step in the testing process is to see whether the proportion from the random sample is sufficiently far from the proportion assumed by the null hypothesis to warrant the rejection of the null hypothesis. (Fielding and Gilbert 2000, p.248)

And another says:

---

[2] For simplicity, this paper discusses only the 'nil' null hypothesis of no difference, common in the Fisher tradition. Other kinds of null hypothesis are possible, and all would be affected by the logical flaws discussed here.

This [i.e. the standard 95% alpha level] means that if we drew 100 samples, we are recognizing that as many as 5 of them might exhibit a relationship when there is not one in the population'. (Bryman 2001, p.233)

Perhaps the most widespread reference says:

The null hypothesis is presumed true until statistical evidence, in the form of a hypothesis test, indicates otherwise — that is, when the researcher has a certain degree of confidence, usually 95% to 99%, that the data does not support the null hypothesis. (Wikipedia – 29/05/08)

The (nil) null hypothesis is also variously described:

The null hypothesis states that the experimental group and the control group are not different with respect to [a characteristic of interest] and that any difference found between their means is due to sampling fluctuation. (Carver 1978, p.381)

…the hypothesis that certain observed data are a merely random occurrence. (Borowski and Borwein 1991, p.411)

What all of these and other texts like them make clear in their different ways is that calculation of the probability of the data/sample encountered in any new research is conditional upon the null hypothesis (rather than the other way around). And that the probability so generated is routinely used to accept or reject the null hypothesis, where the null hypothesis is usually one of no difference. Why does this matter?

The logic is a modified form of the argument of *modus tollendo tollens,* or denying the consequent. If the null hypothesis is assumed true we can calculate the probability of observing data as (or more) extreme than the data we did observe. If this probability is very small it suggests that the null hypothesis is not likely to be true. Of course, the formal *modus tollens* is not based on probability but certainty (in a Platonic tradition). The argument goes:

If A then B

Not B

Therefore, not A

This argument is indisputable, and its soundness can be proved by logic trees or truth tables. However, when converted to a likelihood argument (perhaps in a more Aristotelian tradition), the argument says:

If the null hypothesis (H) is true then we can calculate the probability of observing data (D) this extreme

The probability of D (given H) is very small

Therefore, the probability of H is very small

This second argument is more complex than the simple *modus tollens*. Unlike the formal logic version, it allows for a false positive result (or Type I error) when H is rejected incorrectly because the very small probability of observing D has occurred despite H (as it will on a small percentage of occasions).[3] Many specific events, when closely described, can be deemed low probability, and the occurrence of a low probability event is not, in itself, any evidence of it not being due to chance (as this term is usually interpreted, but see Gorard 2002). Sitting down to a game, the gamer would not declare a die biased just because it rolled two threes in succession. Nor would they do so if they were dealt three cards in the same suit from a standard pack. Yet both of these gaming events are less likely than the 5% threshold used in traditional statistical analysis.

This likelihood version of *modus tollens* also allows for a false negative result (or Type II error) where the probability of D is found to be not very small, and so H is not rejected even where it is not true. In general, users and teachers of statistics tend to understand these limitations of the method, either formally or intuitively. They may reason that as long as the conditional probabilities of D and H are clearly linked, so

---

[3] By definition a Type I error (incorrect rejection of the null hypothesis) is only possible if the null hypothesis is true. The truth of the null hypothesis itself is not derivable from p since p only exists assuming the null hypothesis to be true.

that a low value for p(D|H) means a low value for p(H|D) and *vice versa*, the approach is useful and valuable.[4] As Nickerson (2000, p.251) says:

> Even to many specialists, I suspect, it seems natural when one obtains a small value of p from a statistical significance test to conclude that the probability that the null hypothesis is true must also be very small.

There have been many criticisms of statistical testing over decades (Gorard 2006a), and many examples of misuse of the method such as widespread acceptance of p-values based on non-probability samples, or dredging datasets via multiple use of a technique whose probability calculations are predicated on one-off use (Wright 2003). Several alternatives to significance tests have been suggested, such as greater use of effect sizes, and descriptive, graphical or everyday approaches. Perhaps the most obvious alternative is to treat numeric judgements as we should treat judgements about any of the other kinds of data available to us in social science. That is we need to argue clearly and logically why a specific difference, pattern or trend is valuable and worthy of further consideration – in such a way that readers can see why and so make their own judgement to agree or disagree. There is no general technique available that we can use which means readers must agree with our judgements. Nor perhaps should there be. But these alternatives are still seen by the wider field merely as additions to the otherwise useful and valuable process of significance testing. What this paper hopes to demonstrate is that the easy assumption that a low value for p(E|H) means a low value for p(H|D) is false. *Modus tollens* does not work with probabilities – or, expressed differently, it requires a further probability in order to make it work at all. To illustrate why, I present a simple example of calculating two conditional probabilities.

A simple example of conditional probability

---

[4] The format p(D|H) is used to denote the probability of D given H, or D calculated on the assumption that H is true. Reversing D and H leads to two statements that are clearly not equivalent in real life (or indeed in science, logic, maths or statistics). The probability of carrying an umbrella if it is raining is not, necessarily, the same as the probability of it raining if one carries an umbrella.

Imagine that around 1% of the population will develop a specific disease or illness, and that a predictive diagnostic test has been developed to help identify this unfortunate 1% so that remedial or palliative action can be prepared. Research has shown that the test is 90% accurate, in the limited sense that of the people who go on to develop the disease, 90% would be predicted to do so by the prior test. It is a useful test, even though 10% of the positive cases are likely to be missed (false negatives). However, as with all tests there is a downside. Of those people who do not go on to develop the disease, 10% would still be predicted to do so by the test (false positives). This could cause considerable but unnecessary distress, even though 90% of those who will *not* develop the disease are also correctly identified. Given this situation, and assuming that the research on the accuracy of the test is believable, consider this problem. If a large number of people from the general population are tested and someone you know obtains a positive result from the test (i.e. they are predicted to develop the disease) what is the probability that they will actually develop the disease?

To make this easy to visualise, consider the average results for 1000 people being tested. Of these, 10 (1%) will likely develop the disease, of whom 9 (90%) will test positive (i.e. be correctly predicted to have the disease). Of the remaining 990 (1000-10) who will not develop the disease, 99 (10%) will test positive for the disease (falsely). Thus, in any large group of randomly selected people, someone testing positive is only 9/108 (8%) likely to develop the disease – a high risk but much less likely than the 90% some might fear, and that even relevant professionals such as physicians and counsellors have suggested (Gigerenzer 2002). The frequencies are summarised in Table 1.

Table 1 – Frequencies of test results and disease for 1,000 people (1% prevalence)

	Test positive	Test negative	Total
Develop disease	9	1	10
Not develop disease	99	891	990
Total	108	892	1000

What this example shows is that there is a clear difference between the probability of someone developing the disease having tested positive (8%) and the probability of someone generating a positive test result given that they will develop the disease (90%). Not only are these two conditional probabilities completely different, but also there is no way of calculating or deducing the value of one from the value of the other (in isolation). The conversion of one value to the other requires knowledge of the underlying unconditional probability of developing the disease (1%) and a computation like the one in Table 1.[5] If we keep everything in the example the same, but change the prior probability of getting the disease from 1% to 50% then the probability of someone generating a positive test result given that they will develop the disease is still 90%. However, the probability of someone developing the disease having tested positive is now 90% as well, as explained in Table 2.

Table 2 – Frequencies of test results and disease for 1,000 people (50% prevalence)

	Test positive	Test negative	Total
Develop disease	450	50	500
Not develop disease	50	450	500
Total	500	500	1000

The probability of someone developing the disease having tested positive – $p(D|T)$ - can be very small even when the probability of testing positive given that they will develop the disease – $p(T|D)$ – is very large. These instances of inverse relationship happen when the unconditional prior probability of D is small. Or the two conditional probabilities can be identical (when the prior probability of D is 50%). Or they can be positively related, with $p(D|T)$ even greater than $p(T|D)$ when the unconditional prior probability of D, or $p(D)$, is large.

---

[5] Bayes' Theorem
The more general relationship between two conditional probabilities of the sort discussed in this paper is (where A and B are events, p signifies probability, | signifies given, and ' denotes the inverse of an event):
$p(A|B) = p(A).p(B|A)/(p(A).p(B|A)+p(A').p(B|A'))$
In the example in the paper, if we substitute developing the disease for A and testing positive for B, the relevant probabilities are A=1%, B|A=90%, A'=99%, B|A'=10%. Therefore:
$p(A|B) = 90 / (90 + 990)$
Or 8%, which is the same answer as obtained by considering Table 1. Since we know the overall probability of developing the disease, $p(A)$, we can convert the probability of getting a positive test result if one was going to develop the disease into the much more useful diagnostic probability of developing the disease if one gets a positive test result.

The flaw in statistical reasoning

What is the relevance of this to statistical testing? I hope that readers are already seeing the link between this example and the modified *modus tollens* argument in statistical reasoning. As shown above, the p-value underlying all statistical tests is calculated as the conditional probability of observing data as extreme (or more extreme) as was actually observed in the sample assuming that the sample and population (or two sub-samples) are identical in terms of the measured characteristic. Put in a similar way to the disease example above, a p-value for a statistical test is the probability of getting the data we did given that the null hypothesis of no difference between sample and population (or two sub-samples) is true. For the p-value to be calculated the null hypothesis must be assumed as true. By definition, therefore, the p-value by itself cannot be used to judge or even help judge the likelihood of the truth of the null hypothesis. Yet what analysts seem to want from statistical testing is precisely that ability to judge the probability of the null hypothesis. And this is what they usually report the p-value as being, and how they use the p-value in practice.

> [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).

These countless analysts over generations are making the same mistake as someone believing that a positive result in the diagnostic test above means a 90% chance of developing the disease. What these analysts want is the probability of the null hypothesis given the data they found, but what they actually calculate is the probability of finding that data given the null hypothesis. As the disease example above shows, these two figures are different and one cannot be calculated or even guessed from the other. The widespread practice of significance testing rests on a crucial confusion between these two conditional probabilities.

The assumptions of significance testing only really work when the prior probability of the null (and alternate) hypothesis is 50%. In this unlikely circumstance the probability of the data observed given the null hypothesis - $p(D|H)$ - is the same as

p(H|D) and so the null hypothesis could be plausibly retained or rejected with likelihood p(D|H).[6] In other circumstances this is not possible, which means that statisticians not only have to claim a series of rather unrealistic assumptions such as a perfect random sample with full response, no dropout and no measurement error, they also have to claim that all of their null hypotheses have a 50% likelihood before any data is collected or analysed. In other words, statistical analysis must assume no prior knowledge of any kind!

Nickerson (2000, p.247) provides numerous examples of specialists in statistics routinely ignoring this distinction between pD|H and pH|D, and shows that false beliefs about these conditional probabilities are 'abundant' in the literature. These examples include social scientists and statisticians 'of some eminence', such as Fisher himself.[7] Any reader who has worked in this area will be able to add more examples (including, of course, from my own work when I started out as an academic in 1996/97). This 'belief that p is the probability that the null hypothesis is true' (Nickerson 2000, p.247) is an even more fundamental error than the less common but still prevalent mistaken belief that 1-p is the probability of any specific alternate hypothesis being true.

The situation for statistical analysis is equivalent to the diagnostic test earlier. We can convert the probability of getting the data we did given the null hypothesis into the more useful probability of the null hypothesis given the data we obtained, as we did above using frequencies, or by using Bayes' Theorem. But the problem for analysts is that to do this requires us to know p(H) – the unconditional probability of the null hypothesis being true in the first place. So, apparently tautologically, in order to find the empirical probability of the null hypothesis we must know how likely it is beforehand. Analysts might reasonably use their *a priori* subjective judgement of the null hypothesis being true to substitute for p(H). In which case the new p-value is an estimate of how far that subjective view of the null hypothesis might be affected by the new empirical evidence (see Gorard et al. 2004a). But none of the training texts quoted above, and almost none of the material published in mainstream social science

---

[6] Although of course it is still not clear that one study leading, however convolutedly, to p(H|D)<0.05 should be able to over-ride a prior p(H) of 0.50 in this manner.
[7] Despite being a great pioneer of agricultural trials, Fisher has been known to err (see Gorard 2005).

research journals, follow such a procedure. They simply appear to misinterpret the conditional probability of the new data as though it were the probability of the null hypothesis conditional upon the new data. Falk and Greenbaum (1995) claim that the lack of a clear relationship between pD|H and pH|D discredits the whole logic of significance testing, which answers a question we would never knowingly ask, but leads to us to conclude we have some kind of answer to the questions we might actually want answered – such as how probable is the hypothesis, how reliable are the results, and what is the size of the effect found?

Reprise of argument

In order to clarify this flaw in statistical testing, before moving to some of the implications, I repeat the argument in what I hope is a clear and simple form.

1. In general, the probability of event A given event B is not the same as the probability of event B given event A.

2. The values $p(A|B)$ and $p(B|A)$ can be very different – with one near 100% and another near 0 for example – or they can be similar.

3. It is not possible, generally, to deduce or calculate $p(B|A)$ from $p(A|B)$, without also knowing an *a priori* prevalence figure – the unconditional value $p(B)$.

4. Significance testing in statistics is based on calculating the probability of the finding the evidence/data (D) observed assuming the null hypothesis (H).

5. $p(D|H)$ is not the same as $p(H|D)$.

6. The values $p(D|H)$ and $p(H|D)$ can be very different – with one near 100% and another near 0 for example – or they can be similar.

7. It is not possible, generally, to deduce or calculate $p(H|D)$ from $p(D|H)$, without also knowing the *a priori* unconditional value $p(H)$.

8. If we had a good and reliable figure for p(H) we would presumably not bother to calculate p(H|D) anyway in many situations where we currently use significance tests.

9. We do not generally know the *a priori* figure p(H) when conducting a significance test, and it is almost never used in the literature. There would rarely be a good argument for using the implied underlying value of 50%, since something is almost always known about any topic prior to the research.

10. In these circumstances, therefore, we cannot use p(D|H) to assess the likelihood p(H|D).

11. In summary, with a good prior figure p(H), significance testing is largely pointless. Without p(H), significance testing cannot be defended as logical.

What are the implications?

Conclusions

I have argued here that p(D|H) is not the same as p(H|D), that significance testing involves the probability of findings given the null hypothesis, and so that to treat a statistical test as leading to an estimate of the probability of the null hypothesis being true must be a logical error. I believe that this conclusion is inescapable. The p-value from a statistical test is the probability of such an extreme finding if the null hypothesis is true. If the null hypothesis is not true then the p-value clearly has no meaning. It also follows that purportedly standard thresholds for statistical significance, such as 5%, are not really standard at all. The import of any p-value for the likelihood of any null hypothesis depends, as I have shown, on an *a priori* probability – p(H). Where p(H) is small, as it was in the disease example, a value of 5% or less for the data-generated p-value is less remarkable, and so less suspicious, than if p(H) is nearer 100% at the outset, such as when p(H) is at the 50% level that statisticians are forced by their own arithmetic to imply. As with so many things in social science research, we need an essential base comparator before we can judge the

security of any results. We cannot and should not use statistical tests, in isolation, to make judgements about the likelihood of the null hypothesis. What we would also need to make such a judgement is the *a priori* probability of the null hypothesis being true – which we generally do not know and hardly ever use (but see Gorard et al. 2004).

Of course, significance testing has many other defects as well, many of them well-known. It is very rare to have a full random sample with complete measurement, and yet researchers routinely ignore these assumptions and publish statistical tests anyway, for example. But this further issue of mistaken conditional probability is so serious that, in my opinion, the practice of teaching, using, and publishing statistical tests, in isolation from p(H), should cease altogether and immediately. There really is no sensible alternative to that step, as has long been argued by a number of commentators. Since at least the writing of Jeffreys (1937) it has been well established that a small value of p(D|H) (such as less than 0.05) can be associated with a probability of H that is actually near 1. This problem is the skeleton in the cupboard for traditional statistics.[8] A null hypothesis significance test is:

> …based upon a fundamental misunderstanding of the nature of rational inference, and is seldom if ever appropriate to the aims of scientific research. (Rozeboom 1960, p.417)

> …essential mindlessness in the conduct of research. (Bakan 1966, p.436)

> …not only useless, it is also harmful because it is interpreted to mean something it is not. (Carver 1978, p.392).

> [and 27 years later] - Surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students. (Rozeboom 1997, p.335)

---

[8] The fact that a low posterior probability of H can come from a high pD|H and vice versa is sometimes referred to as Lindley's paradox (even though it is not a paradox) - http://en.wikipedia.org/wiki/Lindley_paradox

Readers may be thinking – 'If a small value of p does not provide a basis for this conclusion, what is the purpose of doing a statistical significance test?' (Nickerson 2000, p.251). This is a very common response to strong criticism of NHST, and one that I have encountered many times when arguing that a significance test may not be suitable in a specific situation (such as when dealing with high non-response or with census data). The first response is often that surely the process must be alright because so many other people are doing it.[9] Even when convinced that this is not a sound argument people ask what should be done instead of NHST with poor samples or populations, as though implying that we should continue with an invalid process unless I (or others) can devise a perfect replacement. This even happened in the reviewing of this paper, where an otherwise positive reviewer was nonetheless concerned with what happened instead, and wanted the paper to focus more on the incomplete alternatives to NHST. My view is that if what appears in this paper were taken seriously and acted upon then the impact for social science methods training and use (particularly in education and psychology) would be so great that this *alone* would provide justification for the paper and the argument it portrays.[10] We are in a peculiar position that those few who see the logical flaws in NHST as currently conducted are, quite properly, moving on to consider what happens instead. But a large number of users, including a less positive reviewer of this paper, still believe that NHSTs are logically sound and that any problems can be written off as merely poor use by a sub-set of users.

The conclusion here should not be confused with an argument against numeric analysis, rather the reverse. We all use numbers successfully and relatively unproblematically everyday but do not need a significance 'test' to decide which of two products in a supermarket is the cheaper or which of two students got the higher mark in a test, for example (Gorard 2006b). Rejecting significance tests makes the use

---

[9] Actually the first (irrelevant) response is usually that I do not have a good understanding of whatever issue is being discussed. A reviewer for this paper said 'It is fairly clear that the author does not have a very clear grasp of basic statistical theory'. Such a comment is irrelevant in that if it were true it would be obvious in what I wrote and so a reviewer could mount an easy counter-argument comprehensible to all readers without adding the insult as well. This is a further example of the lack of intellectual engagement I have previously noted in this field, where reviewers are even prepared to misquote what I have said in order to make the point easier to argue with (see Gorard and Fitz 2006).

[10] For example, standard methods texts in social science statistics and research methods do not warn their readers that a low p-value is very, very weak evidence (if it is evidence at all) against the truth of the null hypothesis.

of numbers less problematic and so encourages their use in research. We should continue where possible to randomise our cases to treatment groups in trial designs, and to select samples at random in passive designs where population figures are not possible.[11] [12] We do this because it helps minimise systematic bias, not because it means we can then use significance tests. We could adopt the Bayesian approach and decide on a subjective synthetic value for the null hypothesis at the outset and use each new finding merely to modify it (although this is neither a complete nor a straightforward solution).[13] Or we could adopt much clearer reporting and judgement of numeric results. Either way, we end up treating numeric data in pretty much the

---

[11] A reviewer for this paper claimed that, 'contrary to what the author seems to believe', use of population data is 'rare'. It will probably surprise this reviewer to learn, therefore, that in the previous year of Oxford Review of Education at time of writing (issues 33, 5 to 34, 4), there were 11 papers presenting numeric analyses of which six used population data, two used convenience samples for which no probabilistic calculations are possible, two used existing large-scale survey data such as PISA, and one was a new study based on a kind of probability sample. Thus, the majority of pieces in this snapshot were based on population figures and hardly any would or should require tests of significance. Perhaps also of relevance to this paper is the finding that only one paper based on population figures presented these as census data (Gorard 2007). The papers by Allen and Vignoles (2007), Evangelou et al. (2007), Flouri and Ereky-Stevens (2008) and Tooley et al. (2007) all present p-values or tests of significance but invalidly based on population data with no probabilistic element. A further paper by Snell et al. (2008) used a convenience sample with no probabilistic element but still invalidly presented p-values.

[12] The same reviewer for this paper also claimed that 'many data analysts would often use' 'interval estimates' 'instead of or in addition to significance tests'. It would presumably also surprise this reviewer to learn that none of these 11 papers used interval estimates in the way the reviewer had suggested. On both of these peculiar claims I believe that the reviewer is wrong – at least for education – and I base this belief on the multiple methodological reviews I have conducted of education literature (including, for example, Gorard et al. 2004b).

[13] Imagine conducting a systematic review of the available evidence on the viability of a particular approach to teaching the understanding of fractions in secondary school mathematics. We would set out to assemble as much of the literature as possible of relevance to the question. We could divide all of these results into those susceptible to conversion for a meta-analysis of effect sizes and those which were more impressionistic. The former we could assess in terms of relatively standard quality checks, such as respondent refusal or drop-out, measurement error and so on. The latter we would present to a mixed panel of relevant experts for detailed reading. We would ask them to use their prior knowledge and experience, coupled with what they have learnt from reading the evidence to rate, in an overtly subjective way, the likelihood that this approach to teaching is effective. We can then either continue with each subjective probability separately or find their overall mean. This figure(s) becomes the prior probability for a calculation using successive numeric studies in sequence as posterior probabilities. Bayes' theorem allows us to adjust the subjective judgements of the experts, using the additional information generated by future studies. For example, the expert opinion based on classroom experience and observations may be that this approach to teaching fractions is very effective. A meta-analysis of randomised controlled trials may show no discernible in using this technique compared to a control. In current procedures, the meta-analysis results would be published separately and used to argue that this approach to teaching is ineffective. The prior expertise would be largely ignored, while practitioners may ignore the important results of the synthesis because it does not accord with their own experience in the classroom. In this version everybody loses. What a synthetic analysis does, more properly, is to weigh the two versions of the evidence in terms of the common `currency' of subjective probabilities. It asks: is the posterior evidence so strong that it should substantially change the minds of the experts? This is a much tougher proposition for the evidence than the more usual question: is the result `significant'. For discussion of some of the practical problems that such an approach faces, see Gorard with Taylor (2004).

same way as we should also treat text, visual or auditory data, by judging patterns, trends or exceptions and then reporting them in such a way that readers have enough information to attempt the same judgement. I repeat – this is not an argument for not using numbers in research. In fact, it is part of a plea for numbers to be used more but more simply, as they often are in everyday life, by a much wider range of researchers than at present.

The error of confusing two different conditional probabilities could be part of the reason why social science statistical results so often lead to non-effective treatments and 'vanishing breakthroughs' (Harlow et al. 1997). Because the real results depend on a value, p(H), not available for the standard calculation. It could be why some newcomers to research have such a hard time understanding statistics as it is practiced (because it does not actually make sense).[14] And so it could be partly why we end up with a methods schism based on a confused and wary majority eschewing numeric analysis, while a minority routinely misuses complex statistical procedures but largely escapes rigorous critique by peer-reviewing each others work. Ceasing the illogical, unjustified and largely useless practice of statistical significance testing may thus yield several quick benefits.

Research methods training would be less threatening, as statistical testing and analysis is frequently reported as the most off-putting aspect of research methods for most students. The apparent need to learn about different kinds of logic for use with different kinds of data would be reduced. This might lead to a more general and widespread use of simple numbers, as students realise that setting the context with a table of frequencies, or explaining a problem to be researched in terms of a numeric difference between social groups, does not entail entering some kind of 'quantitative' paradigm. Instead of starting each new study or analysis from a position of feigned ignorance, as we are forced to do when conducting a significance test as currently envisaged, we can estimate a subjective prior probability for any hypothesis (or idea) and use future evidence to modify rather than replace that probability. This quasi-

---

[14] Of course, part of the reason for the generally acknowledged difficulty faced by students and new researchers when dealing with statistics is that the logic of NHST 'does not seem to accord with what would be the mode of reasoning in ordinary rational discourse' (Berkson 1942, p.267). Perhaps because, as this paper argues, NHST is not logical or rational.

Bayesian approach to knowledge accumulation, as well as using past evidence efficiently, also has the advantage that it is independent of the scale and type of evidence. A small study with a certain effect size would modify our existing knowledge to a lesser extent than a large study with the same effect size. But we do not need to ignore small scale work, and so student and practitioner research can be included easily in future syntheses of evidence. In fields like education and social work, it also means that professional experience and judgement can rightly be included in helping to form the prior (or underlying) probability. More generally, this prior probability can be based on existing evidence of all types. Like several other purported barriers to mixing methods, significance testing is a rhetorical illusion, hindering scientific progress for no clear gain.

References

Allen, R. and Vignoles, A. (2007) What should an index of segregation measure?, *Oxford Review of Education*, 33, 5, 643-668

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

Berkson, J. (1942) Tests of significance considered as evidence*, Journal of the American Statistical Association*, 37, 325-335

Borowski, E. and Borwein, J. (1991) *The Harper Collins dictionary of mathematics*, New York: Harper Collins

Brookes, B. and Dick, W. (1951) *Introduction to statistical method*, London: Heinemann

Bryman, A. (2001) *Social research methods*, Oxford: Oxford University Press

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

de Vaus, D. (2002) *Analyzing social science data: 50 key problems in data analysis*, (London: Sage)

Evangelou, M., Brooks, G. and Smith, S. (2007) The Birth to School Study: evidence on the effectiveness of PEEP, *Oxford Review of Education*, 33, 5, 581-609

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

Flouri, E. and Ereky-Stevens, K. (2008) Urban neighbourhood quality and school leaving age, *Oxford Review of Education*, 34, 2, 203-216

Gigerenzer, G. (2002) *Reckoning with risk*, London: Penguin

Gorard, S. (2002) The role of causal models in education as a social science, *Evaluation and Research in Education*, 16, 1, 51-65

Gorard, S. (2004a) Scepticism or clericalism? Theory as a barrier to combining methods*, Journal of Educational Enquiry*, 5, 1, 1-21

Gorard, S. (2004b) Three abuses of 'theory': an engagement with Nash, *Journal of Educational Enquiry*, 5, 2, 19-29

Gorard, S. (2005) Revisiting a 90-year-old debate: the advantages of the mean deviation, *The British Journal of Educational Studies*, 53, 4, 417-430

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

Gorard, S. (2006b) *Using everyday numbers effectively in research:* Not *a book about statistics*, London: Continuum

Gorard, S. (2007) What does an index of segregation measure?, *Oxford Review of Education*, 33, 5, 669-677

Gorard, S. and Cook, T. (2007) Where does good evidence come from?, *International Journal of Research and Method in Education*, 30, 3, 307-323

Gorard, S., with Taylor, C. (2004) *Combining methods in educational and social research,* London: Open University Press

Gorard, S., Roberts, K. and Taylor, C. (2004a) What kind of creature is a design experiment?, *British Educational Research Journal*, 30, 4, 575-590

Gorard, S., Rushforth, K. and Taylor, C. (2004b) Is there a shortage of quantitative work in education research?, *Oxford Review of Education*, 30, 3, 371-395

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

Jeffreys, H. (1937) *Theory of probability*, Oxford: Oxford University Press

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

Rozeboom, W. (1960) The fallacy of the null hypothesis significance test, *Psychological Bulletin*, 57, 416-428

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

Snell, M., Thorpe, A., Hoskins, S. and Chevalier, A. (2008) Teachers' perceptions and A-level performance, *Oxford Review of Education*, 34, 4, 403-423

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

Tooley, J., Dixon, P. and Gomathi, S. (2007), *Oxford Review of Education*, 33, 5, 539-560

Wright, D. (2003) Making friends with your data: improving how statistics are conducted and reported, *British Journal of Educational Psychology*, 73, 123-136