The need for nuance in the null hypothesis significance testing debate

Olle Häggström

Abstract

Null hypothesis significance testing (NHST) provides an important statistical toolbox, but there are a number of ways in which it is often abused and misinterpreted, with bad consequences for the reliability and progress of science. Parts of contemporary NHST debate, especially in the psychological sciences, is reviewed, and a suggestion is made that a new distinction between *strongly*, *weakly* and *very weakly* anti-NHST positions is likely to bring added clarity to the debate.

1. Introduction

While the human cognitive machinery is impressive in many ways, it has limitations and biases that hamper our ability to spontaneously do good science. Examples of such biases are (a) our tendency in many situations towards overconfidence in our beliefs, and (b) an overly trigger-happy pattern detection mechanism that tends to see patterns in what is actually just noise; see, e.g., Häggström (2013). This necessitates the use of rigorous statistical methods to circumvent these biases. The most widely used of these, from the 20th century and onwards, is so-called frequentist statistics, of which null hypothesis significance testing (NHST) forms a central part; see Salsburg (2001) for a reasonably balanced history. NHST has, however, come under some fire in recent years from critics who claim that its practice suffers from major problems. And indeed, there are problems, but that does not make all criticisms correct. In this paper, I will try to add some much needed nuance to the discussion, to distinguish various strands of the anti-NHST view, and to evaluate their merits. The main focus will be on the practice of NHST in psychology and related sciences. My eventual verdict will be that, on one hand, we cannot do without NHST in our statistical toolbox – a position that implies that the fiercest critics of NHST who want to abolish the practice altogether go too far – while, on the other hand, there are features of contemporary NHST practice that need to change.

This paper is organized as follows. At the end of this introductory section I will recall basic NHST terminology. Section 2 is a quick tour through some of the main malpractices that haunts present-day NHST practice. In Section 3, I comment on the positions of a few of the most influential contributors to contemporary NHST debate. The reader will come out of that section with the (correct) impression that the debate is multifaceted and somewhat chaotic. In Section 4 I offer an attempt at providing a least a bit of structure to the debate through a distinction between different strands of anti-NHST position; I believe that if everyone who identifies themselves as anti-NHST would make an effort to clarify their own stance with respect to my classification, then the clarity of the debate as a whole would improve nontrivially. The reason for this is explained in Section 5, where I also offer some further conclusions regarding what ought to be done in order to improve statistical practice and science as a whole.

There is no need here for a mathematically fully precise treatment of NHST, but we do need the basic terminology. Statistical hypothesis testing is always relative to some **null hypothesis**, which typically states that some effect or some parameter is zero. Given the data, the **p-value** can loosely be defined as what the probability would be of getting at least as extreme data as those we actually got, if the null hypothesis were true. The looseness of this definition comes from the fuzziness of the term ``extreme´´, which actually needs to be carefully defined before analyzing the data. The standard so-called Fisherian definition gives much flexibility about defining ``extreme´´, whereas the other classical approach – that of Neyman and Pearson – requires the specification of an alternative hypothesis, and defining ``extreme´´ so as to maximize power (see below) with respect to that alternative hypothesis. See, e.g., Lehmann and Romano (2008) for a more careful discussion.

If the p-value ends up below a given threshold – which is called the **significance level** and which (mostly for historical reasons) is most commonly taken to be 0.05 – then **statistical significance** is declared, which is typically considered to provide (at least some) reason for suspecting that the null hypothesis is false. For a given significance level, the **power** of a testing procedure is a function of the effect size (or, more generally, of the precise choice of alternative hypothesis), and it is defined as the probability of obtaining statistical significance under that effect size.

A (Fisherian) 95% **confidence interval** -- or, more generally, **confidence set** – for a parameter ρ is the set of all r such that, with the given data, testing the null hypothesis ρ=r would not yield statistical significance at level 1-0.95=0.05. This guarantees, regardless of the true value of ρ (but provided all other model assumptions are true), that the probability of getting data that yields a confidence interval that includes the true ρ is at least 0.95. The concept generalizes in the obvious way to confidence intervals with confidence levels other than 95%.

2. Not all is well

It is hard to imagine how 20th and early 21st century science could have done without the NHST machinery, and I will argue later in this paper that we still need it. It must, however, be admitted that current NHST practice suffers from a number of endemic problems. The following are amongst the most detrimental.

(a) Any Statistics 101 class teaches that obtaining statistical significance does *not* constitute disproof of the null hypothesis, and likewise that failure to obtain statistical significance does *not* constitute proof of the null hypothesis. Nevertheless misconceptions to the contrary are often implicit, and sometimes even quite explicit, in statements made by highly qualified researchers, including Harvard professors such as Mitchell (2014) who believes that p<0.05 does the same thing to his null hypothesis that an observation of a black swan does to the all-swans-are-white hypothesis; see also Häggström (2014b).

(b) The fallacy of the transposed conditional. Despite wide-spread claims to the contrary, a p-value cannot be interpreted as the probability that the null hypothesis is true. To do so is an example of mistaking a statement about the probability of certain kinds of data given the null hypothesis for a statement about the probability of the null hypothesis given the data, which in turn is an example of the fallacy of the transposed conditional: confusing P(A|B) with P(B|A), which can lead to arbitrarily misleading conclusions.

Likewise, in a situation where a 95% confidence interval has been calculated, it may be tempting to conclude that the true parameter value sits in that interval with probability 0.95, but again this would be to commit the fallacy of the transposed conditional; see Morey et al. (2015) for extensive discussion of this instance of the fallacy.

There is a much-employed but highly unfortunate terminological convention in the use of the word ``likely´´, which I believe contributes to the confusion. Consider the case of statistician Yudi Pawitan's influential book *In All Likelihood*. Discussing an experiment where a coin with unknown heads-probability θ is tossed 10 times, resulting in 8 heads, he writes that we are in a position to conclude that θ ``is very unlikely to be very small´´, and that in contrast ``θ=0.6 or θ=0.7 is likely´´ (Pawitan, 2001, p 21). Since, in ordinary English, ``likely'' is synonymous to ``having high probability´´, this very much *looks like* a case of the fallacy of the transposed conditional. Pawitan, however, is much too sophisticated a statistical thinker to make that mistake; what he does is that he implicitly *defines* the statement ``θ=0.7 is likely´´ to *mean* that the likelihood function L(θ) – defined as the probability of the obtained data provided the parameter value is θ – takes a large value for θ=0.7. For another example, we may note that the same language appears in a recent column by the president of the Association for Psychological Science:

> The information that the binomial likelihood function conveys is extremely intuitive. It says that given that we have observed 7 successes in 10 tries, the probability parameter of the binomial distribution from which we are drawing […] is very unlikely to be 0.1; it is much more likely to be 0.7, but a value of 0.5 is by no means unlikely. (Gallistel, 2015)

It is difficult or impossible for the reader, confronted with a passage like this one, to judge whether the author is under the spell of the fallacy of the transposed conditional, or if he is merely adhering to the same unfortunate terminological convention as Pawitan. In any case, the chosen language invites confusion and almost begs statistically unsophisticated readers to commit the fallacy of the transposed conditional.

(c) If a study shows a statistically significant effect of, say, a medical treatment compared to placebo, on one group such as men, but no statistically significant effect on another group such as women, then it may be tempting to conclude that the data exhibits a statistically significant difference between the treatment's effects on the two groups. However, such a conclusion is unwarranted; a separate test of the difference is needed. A study by Nieuwenhuis et al. (2011), scrutinizing the statistical analyses in neurobiology papers in a range of top journals, found that in 79 of the papers, precisely this mistake was made – while in 78 of them the correct procedure was carried out.

(d) Sizeless science. Consider a clinical trial comparing a new treatment to placebo, and suppose that the true fact of the matter is that the treatment does have a positive effect compared to placebo, but that the effect is too small to be of any practical relevance to patients' health or well-being. If the sample size of the study is sufficiently large, it is nevertheless likely that statistical significance will be obtained. This illustrates the need for considering not only statistical significance, but also what we might call *subject-matter significance*, meaning that the observed effect is large enough to be of subject-matter interest. Ziliak and McCloskey (2008) scrutinized 369 papers involving regression analysis in the prestigious journal *American Economic Review*, and found that 276 of them committed sizeless science – that is, nearly three quarters of the papers.

(e) Failure to account for multiple testing. Consider the case of Bygren et al. (2014) who reported that (quoting the title of their paper) ``change in paternal grandmothers' early food supply influenced cardiovascular mortality of the female grandchildren´´, which sounds like a rather sensational result in epigenetics. The reported p-value is 0 .016, but a closer look at the paper reveals that the authors tested a total of 24 combinations of food supply pattern, maternal vs paternal grandmothers and grandparents, and sex of grandchildren. Under the null hypothesis of no influence or correlation of the kind looked for, this would on average (i.e., in the long run, with repeated experiments) produce 24*0.05=1.2 statistically significant outcomes at level 0.05, and 24*0.016=0.384 outcomes with p-value at most 0.016. Getting one such p-value is not much of a surprise under the null hypothesis, and thus cannot count as much evidence against the null hypothesis. What the authors of the study failed to do was to account for the multiple testing using some of the standard statistical procedure for that purpose (see, e.g., Bretz, Hothorn and Westfall, 2011). The simplest method (so called Bonferroni correction) is just to multiply the smallest p-value by the number of tests, giving a corrected p-value of 0.384 – nothing to write home about, and none of the fancier procedures will change that. There is nothing unique or uncommon about the Bygren et al. (2014) failure to account for multiple testing, but it is bad practice nevertheless, and a case which, due to the naïve press coverage it received, happened to attract my attention (Häggström, 2014a).

A very common special case of failure to account for multiple testing arises when both positive and negative values of a parameter ρ are possible, and the null hypothesis is that ρ=0. Unless there is good reason to expect a nonzero parameter to be in a certain direction (positive or negative) one should normally carry out a so-called two-sided test, which takes the possibilities of both ρ<0 and ρ>0 equally into account. If there is good reason to expect ρ<0 rather than ρ>0, a one-sided test can be decided on beforehand, but the cheap trick to first look at the data and then decide on the direction of the one-sided test to get the best p-value is incorrect; in effect, such a procedure is two-sided but produces a better (half as large) p-value than the correct two-sided procedure. Ruthless (or naïve) researchers can do this and then produce an ad hoc explanation for why the chosen direction was to be expected before the data was collected and analyzed. Doing so can be seen as a special case of the next malpractice.

(f) Publication bias. A variant of the multiple testing malpractice (e) is the phenomenon that statistically significant outcomes are more likely to be published than those that are not. This produces a bias in the literature in favor of statistically significant results. Evidence suggest that the effect is substantial; see, e.g., Decullier et al. (2005) and Fanelli (2010, 2012). This is arguably an even worse phenomenon than (e), where the tests that are unaccounted for in the final statistical analysis are at least reported, making it possible for other researchers to do a more careful analysis. This is not so readily done when the statistically non-significant results do not even appear in the literature, and this causes severe difficulties in, e.g., doing meta-studies to evaluate the totality of evidence on some scientific problem.

(g) Replication crisis. As a consequence of many or perhaps all of the above misconceptions and malpractices (a)-(f), much of science seems to be in a situation which has been described as a replication crisis: a large proportion of published statistically significant findings resist replication. Ioannidis (2005) gave an influential study of this phenomenon, and the problem has also received attention in popular press (Lehrer, 2010). Recently, the severity of the phenomenon in the psychology literature was demonstrated in an ambitious collaborative effort emphasizing effect sizes: in 100 replicated studies, the mean effect size in the replications was just under half of that in the original studies, and only 39% of the replications gave confidence intervals containing the estimated effect size of the original study (Open Science Collaboration, 2015).

3. Some highlights from the debate

During large parts of the 20th century, intense debate took place within the statistics community between on one hand defenders of NHST and other methods in frequentist statistics, and on the other hand advocates of Bayesian statistics. I shall refrain from discussing this history (but see Salsburg, 2001) and instead focus on how researchers in a wide range of applied fields have increasingly begun to highlight their dissatisfaction with NHST practice. Researchers in psychology and related sciences have been especially active in this discourse. Carver (1978) and Cohen (1994) are two relatively early examples, both of them pointing out several of the misconceptions and malpractices discussed in Section 2, plus a few others.

3.1. Cumming. Particularly influential in the NHST debate among psychologists today is Geoff Cumming, who advocates, in papers, op-eds, instructional videos and a textbook, that reporting p-values and statistical significance should be abolished in favor of reporting confidence intervals, which he argues to be much more informative; see, e.g., Cumming (2009, 2012a, 2012b, 2014). For concreteness, let me zoom in on his 2009 YouTube video which is a bit of a pedagogical masterpiece, using a simple computer simulation experiment to demonstrate beautifully the superiority, in a particular setting, of representing the results in terms of confidence intervals as opposed to p-values. The chosen setting is a comparison of a single trait between two groups, with sample size 32 in each group. Within each group, the trait is normally distributed with the same variance but possibly different means, and the task is to say something interesting and informative about the difference. Cumming demonstrates that if the true difference between the means is exactly half of the within-group standard deviation (i.e., the standardized effect size known as Cohen's d equals 0.5), then p-values give very little interesting information about this true difference, while confidence intervals perform better in this respect. I am reasonably convinced that this is so in the chosen setting, but I believe the generalizability to other settings is rather less than the video might lead one to think, in at least three ways.

First, in his example, Cumming chose to pick a combination of sample size and effect size in such a way as to make the power, with respect to significance level 0.05, close to one half (0.52). This maximizes, in a sense, the noise level of the sequence of statistical significances and non-significances in repeated replications of experiments (in more precise mathematical language, it maximizes the entropy of the distribution of the indicator function of statistical significance). With either a much smaller (or zero) effect size, or a much larger one, ``replicability´´ (in Cumming's somewhat inept terminology) of statistical significance becomes much better. It is mathematically obvious that an effect size can be chosen that produces a borderline case as to whether the experiment tends to produce statistical significance or non-significance. With this in mind, the take-home message from Cumming's video is not so much that p-values and statistical significance are useless concepts, but more that power 0.52 at the effect sizes one is looking for (i.e., effect sizes large enough to be of subject-matter interest, yet small enough that one can plausibly hope that they exist) is a way too modest level of power, so a larger sample size ought to be used. Of course I realize that there are costs to increasing sample size, and Cumming describes his chosen combination of sample size and effect size to be typical of experimental studies in psychology, but if this is the case, then it is probably a good idea that researchers in psychology reconsider their priorities and aim for fewer studies but with larger sample size.

Second, Cumming argues that the confidence intervals in his simulation are intuitively self-explanatory in the sense of having the property that the width of the intervals provide a good indication of how much their location is likely to vary from one replication to another. This is true, but it is an artefact of choosing the confidence level equal or close to 1-0.05=0.95. With a higher confidence level, the same heuristic will overestimate the variation in location, while with a lower confidence level it yields an underestimate; this is because the centers of the intervals (and hence the amount of variation in their location) are independent of the confidence level, whereas the interval width increases with increasing confidence level. So for Cumming's argument here to serve as a guide to statistical practice, it will have to involve a continued or increased focus on the particular confidence level 0.95, which is tantamount to significance level 0.05. In Section 5 I will argue that the widespread obsession today with significance level 0.05 is a bad thing.

Third, in the simple situation described by Cumming, calculating confidence intervals is just about as straightforward as obtaining p-values, but in general this is not the case. Computing a confidence interval, or more generally a confidence set, implicitly involves calculating p-values all of the (typically infinitely many) possible null hypotheses corresponding to different effect sizes. Sometimes, as in Cumming's situation, this is doable in a single sweep, but in more complicated situations it can be intractable. And even in cases where it is computationally doable, the confidence set may turn out more complicated and less easy to interpret (or even to represent graphically) than just an interval, especially if we are dealing with not just a single effect parameter, but with multiple parameters. So even if Cumming's recommendation to abandon p-values in favor of confidence intervals can be followed in a range of more-or-less standard situations, it seems infeasible to implement it across all statistical analyses in all sciences.

3.2 The BASP declaration. Perhaps the most dramatic event so far in the NHST debate is the recent declaration by David Trafimov and Michael Marks, in an editorial in their journal *Basic and Applied Social Psychology* (BASP), that NHST methods, including confidence intervals, are ``invalid´´ and therefore banned from the journal (Trafimov and Marks, 2015). As an explanation for the judgement ``invalid´´, they offer this:

> Confidence intervals suffer from an inverse inference problem that is not very different from that suffered by the NHSTP [i.e., NHST procedure]. In the NHSTP, the problem is in traversing the distance from the probability of the finding, given the null hypothesis, to the probability of the null hypothesis, given the finding. Regarding confidence intervals, the problem is that, for example, a 95% confidence interval does not indicate that the parameter of interest has a 95% probability of being within the interval. Rather, it means merely that if an infinite number of samples were taken and confidence intervals computed, 95% of the confidence intervals would capture the population parameter. Analogous to how the NHSTP fails to provide the probability of the null hypothesis, which is needed to provide a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval.

In short, they mean to say that use of NHST is tantamount to committing the fallacy of the transposed conditional (item (b) in Section 2 above). But, as all statistics professors teaching frequentist methods take great pains to explain, in lectures and in textbooks, the logic behind NHST is *not* the (faulty) logic of the fallacy of the transposed conditional, but instead the following, where I quote one of the pioneers of 20th century statistics, in a context where he has obtained, in an applied problem, a fairly low p-value of about 0.00003:

> The probability [...] is amply low enough to exclude at a high level of significance any theory involving [the null hypothesis]. The force with which such a conclusion is supported is logically that of the simple disjunction: Either an exceptionally rare chance has occurred, or [the null hypothesis] is not true. (Fisher 1956, p 39)

In other words, the lower the obtained p-value is, the greater a coincidence we need to accept in order to explain away the data and hold on to the null hypothesis; eventually, as the evidence accumulates, doing so becomes untenable. In my view, this captures well the process of how science marches forward, and the logic is sound. Trafimow's and Marks' verdict, declaring NHST methods to be ``invalid´´, is based on a straw man argument, pretending that its logic is based on committing the fallacy of the transposed conditional. I will try to dig deeper into the Trafimow—Marks position in Section 4.

3.3. Ziliak and McCloskey. As to contributions to the NHST debate from empirical sciences other than psychology and related fields, I think one of the most important is the book *The Cult of Statistical Significance* by economists Stephen Ziliak and Deirdre McCloskey (2008). With particular emphasis on sizeless science (item

(d) in Section 2), they paint an ambitious and alarming image of the quality of statistical practice, mainly in their own discipline economics, while not entirely neglecting other disciplines such as psychology and medicine. The book is very strong on describing the problem, i.e. the prevalence of statistical malpractice, but it is distinctly weaker on suggesting a solution, and as I say in my review of the book in the *Notices of the American Mathematical Society* (Häggström, 2010):

> Sometimes the authors push their position a bit far, such as when they ask themselves: ``If null-hypothesis significance testing is as idiotic as we and its other critics have so long believed, how on earth has it survived?´´ (p. 240). Granted, the single-minded focus on statistical significance that they label sizeless science is bad practice. Still, to throw out the use of significance tests would be a mistake, considering how often it is a crucial tool for concluding with confidence that what we see really is a pattern, as opposed to just noise. For a data set to provide reasonable evidence of an important deviation from the null hypothesis, we typically need *both* statistical *and* subject-matter significance.

It is somewhat odd, but actually quite illustrative of the overly broad strokes that are characteristic not only of Ziliak's and McCloskey's writings but of much of the NHST debate more generally, that in a later contribution (Ziliak and McCloskey, 2010), they point specifically to my review as an example in support of their claim that ``in several dozen journal reviews and in comments we have received, [no one] has tried to *defend* null hypothesis significance testing´´ – a statement that is obviously falsified by the above passage from the review.

4. A taxonomy of anti-NHST positions

I wish that participants of the NHST debate expressing anti-NHST sentiments (including those cited in Section 3) would clearly state whether they wish to rule out as unscientific all arguments involving or building on calculation of p-values, or whether they merely want to ban explicit reference to such calculation in scientific papers. In other words, they ought to declare whether their position is *strongly* or *weakly* anti-NHST, in the terminology that I hereby propose.

The strongly anti-NHST position: *To calculate a p-value (i.e., the probability of getting data at least as extreme as those we actually got, under some well-specified statistical model) is a conceptual error and can never form part of a valid scientific argument.*

The weakly anti-NHST position: *To calculate a p-value can sometimes form part of a valid scientific argument, but when that is the case it should be done in secrecy: scientific papers should never make explicit reference to p-values or to the derived notion of statistical significance.*

Of course, it is conceivable that someone who identifies herself as anti-NHST might find that neither strongly, nor weakly anti-NHST accurately captures her position, but in such cases, it is likely to be illuminating if she went on to explain wherein the mismatch lies.

Cumming (2009, 2012a, 2012b, 2014) advocates the use of confidence intervals, as do (albeit to a lesser extent) Ziliak and McCloskey (2008), and since confidence intervals are derived from p-value calculations, these authors' views do not match the strongly anti-NHST position. So their respective positions are at most weakly anti-NHST, but it remains to be heard whether they find the weakly anti-NHST position to be an accurate description of their respective views.

The case of Trafimow and Marks (2014) is more difficult to interpret. While some of their rhetoric suggests the strongly anti-NHST position, the following passage is more suggestive of the weakly anti-NHST position:

> Will manuscripts with p-values be desk rejected automatically? […] No. If manuscripts pass the preliminary inspection, they will be sent out for review. But prior to publication, authors will have to remove all vestiges of the NHSTP (p-values, t-values, F-values, statements about ``significant´´ differences or lack thereof, and so on).

It may also be noted that as one of their preferred alternatives to NHST, they ``encourage the use of larger sample sizes than is typical in much psychology research, because as the sample size increases, descriptive statistics become increasingly stable and sampling error is less of a problem´´. This is noteworthy, because while their statement about the advantage of larger sample sizes is correct, calculation of hypothetical p-values is necessary in order to derive this monotonicity property and to work out how large a sample size will be needed to attain a desired level of precision.

In my opinion, this last observation demonstrates the untenability of the strongly anti-NHST position. Apparently even Trafimow and Marks recognize that instability of statistical estimates and sampling error are a problem, but to embrace the strongly anti-NHST position would be to deprive ourselves of the ability to calculate *how much* of a problem it is, and how large a sample size we would need to bring the problem down to an acceptable level. (One of the referees suggested that one might be able to judge the sample size needed while remaining strongly anti-NHST by calculating standard deviations rather than p-values. But this overlooks the underlying reason why standard deviations are relevant for this. To go from a small standard deviation to a small probability of a highly misleading statistical estimate one needs to apply Chebyshev's inequality. That step, which is so standard that it is usually left implicit, is precisely the bounding of a p-value.)

On the other hand, the weakly anti-NHST position strikes me as unacceptable for other reasons. To accept a scientific methodology, but to require that it is kept secret, banning any mention of it in scientific publications, flies straight in the face of the transparency that is a cornerstone of good scientific method.

To summarize, it seems that neither the strongly, nor the weakly anti-NHST position is acceptable. Well then, might there be some third strand of anti-NHST that makes better sense – perhaps one that accepts the need to calculate p-values as well as the principle of openness about the scientific arguments that are used, at the same time as it recognizes the need to act in ways to keep in check the various misuses and misinterpretations of NHST discussed in Section 2? Let me propose the following.

The very weakly anti-NHST position: *Giving only p-values and declarations of statistical significance, without further elaboration and explanation, is not an acceptable way to present the results of a scientific study.*

This is in fact a strand of anti-NHST sentiment that I can embrace. I do not claim any originality in taking on this position. On the contrary, I daresay it is held by virtually all professional statisticians. Note that *not* holding the very weakly anti-NHST position is tantamount to accepting the practice of sizeless science (item (d) in Section 2). Since accepting that practice means accepting the scientific literature to be filled with reports of statistically significant findings without mentioning whether the observed effect sizes are large enough to matter, the very weakly anti-NHST position becomes pretty much obligatory for anyone with a reasonable sense of good scientific practice.

5. What needs to be done

That practitioners of statistical methods in a variety of applied disciplines have taken over much of the initiative from statisticians in the debate on how to do statistics is in my opinion a welcome development; a broad interest in these matters is likely to help the scientific community move forward on them. I only wish that fewer of these participants took on poorly grounded anti-NHST positions.

In my humble opinion, statisticians still have a lot to contribute to these discussions, and it is therefore important that they take part. Without their participation, much of the body of knowledge of statistical inference accumulated during the 20th century and onwards risks getting lost. For instance, if the editors of BASP had had contact with professional statisticians, they would probably have been spared the embarrassment of banning NHST methods on the confused grounds that they did. Another case in point – less dramatic but nonetheless interesting – is the paper by cognitive scientist John Kruschke (2010), who advocates the abandonment of NHST methods, arguing that their violation of the so-called **strong likelihood principle** (SLP) is unacceptable. He does not use that terminology, however, being apparently unaware both of SLP being known at least since Savage (1962), and of the long controversy over SLP since then (see, e.g., Cox and Hinkley, 1974, Barndorff-Nielsen, 1985, and Pawitan, 2001). I am sure his paper would have been much more interesting and useful to the scientific community if there had been statisticians around to inform him of this discourse so that he could have taken into account the arguments for and against SLP that have accumulated over the years, rather than trying to reinvent the discussion from scratch.

One of the referees for this paper asked what role I hope the distinction proposed in Section 4 between strongly and weakly anti-NHST positions will play in the debate. I am a bit reluctant to give away my answer to this question, for reasons analogous to why one might not want to ruin a joke by explaining it, but all right, here it is: none. I do not expect or even hope to encounter future debates whose contestants have, for clarity, signed up for the strongly versus and weakly anti-NHST camp, respectively. Rather, my ambition is that anyone with an anti-NHST sentiment reading Section 4, will, faced with the strongly vs weakly anti-NHST distinction, realize that both positions are untenable, and therefore give up being anti-NHST altogether (or switch to the third option of being very weakly anti-NHST, which actually is not an anti-NHST position at all, but rather one about what constitutes good NHST practice). Once that is accomplished, there is no further need for the distinction.

Besides the much-needed presence of statisticians in the debate over what constitutes good statistical practice, their presence is even more badly needed in the direct workings of applied sciences – in the actual research projects. Because, as I've argued in Section 4, NHST cannot be abandoned without seriously hurting science, and therefore the competence level among practitioners of statistical methods needs to improve, so as to avoid statistical pitfalls and malpractices such as those discussed in Section 2. This, I believe, requires *both* a better (qualitatively as well as quantitatively) statistics part of the undergraduate and graduate training in other disciplines, *and* an increased presence of statisticians in research projects in these disciplines.

A particular point where I think an improved understanding among scientists would greatly benefit science itself is the interpretation of a statement like ``p<0.05´´. Suppose everyone understood that statistical significance on level 0.05 on its own is no more than (at most) the observation that data suggest that an effect may well be present, an observation that can be taken to warrant further study of the possible presence of the effect. If scientists in general understood this, then the low success rate of replication studies (item (g) in Section 2) might no longer deserve the term ``replication crisis´´, but would instead represent a reasonably healthy state of science. As things are today, ``p<0.05´´ is way too often taken to mean something like ``beyond reasonable doubt´´, or to serve as an excuse for cocksure statements such as the title of the Bygren (2014) paper that I quoted in item (e) of Section 2. For that kind of interpretation of statistical significance, way lower significance levels are needed – although at this point I am reluctant to give a number, because how low it is reasonable to go depends very much on the particulars of the problem, including how a priori plausible the existence of the searched-for effect seems to be. The p-value alluded to by Fisher (1956) in the passage quoted in Section 3.2 is about 0.00003 – which in some contexts I might consider small enough to state with a high degree of confidence that the null hypothesis is probably false, but certainly not in all. For instance, the Big Data revolution in areas like bioinformatics (see, e.g., Efron 2012) has led to routine experiments involving thousands or even millions of hypotheses to be tested simultaneously; in such situations, a raw p-value of 0.00003, untouched by multiple inference analysis, does not even warrant a raised eyebrow.

A more well-calibrated and realistic interpretation of the evidential value of statements like ``p<0.05´´ would likely (or at least *should*) come hand in hand with placing a higher value on replication studies. Unfortunately, today such studies are generally not held in high regard. The view of Mitchell (2014) is extreme (and utterly confused due to committing fallacy (a) of Section 2), but the dominant view in most empirical sciences today is depressingly similar to the one expressed by the anonymous professor in following anecdote, told by Richard Feynman in a commencement speech at Caltech in 1974:

> When I was at Cornell, I often talked to the people in the psychology department. One of the students told me she wanted to do an experiment. […] It had been found by others that under certain circumstances, X, rats did something, A. She was curious as to whether, if she changed the circumstances to Y, they would still do A. So her proposal was to do the experiment under circumstances Y and see if they still did A.
>
> I explained to her that it was necessary first to repeat in her laboratory the experiment of the other person – to do it under condition X to see if she could also get result A, and then change to Y and see if A changed. Then she would know that the real difference was the thing she thought she had under control.

> She was very delighted with this new idea, and went to her professor. And his reply was, no, you cannot do that, because the experiment has already been done and you would be wasting time. This was in about 1947 or so, and it seems to have been the general policy then to not try to repeat psychological experiments, but only to change the conditions and see what happens. (Feynman, 1985, p 344)

That was in the 1940s, but the same contemptuous view of replication studies is still dominant. It is time we overcome it.

Acknowledgement: I thank Kimmo Eriksson, Jacques de Maré, Mats Rudemo, Rebecca Willén and three anonymous referees for valuable comments on an earlier draft.

References

Barndorff-Nielsen, O.E. (1985) Diversity of evidence and Birnbaum's theorem, *Scandinavian Journal of Statistics* **22**, 513—515.

Bretz, F., Hothorn, T. and Westfall, P. (2011) *Multiple Comparisons Using R*, CRC Press, Boca Raton, FL.

Bygren, L.O., Tinghög, P., Carstensen, J., Edvinsson, S., Kaati, G., Pembrey, M. and Sjöström, M. (2014) Change in paternal grandmothers' early food supply influenced cardiovascular mortality of the female grandchildren, *BMC Genetics* **15**:12.

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

Cohen, J. (1994) The earth is round (p<.05), *American Psychologist* **49**, 997—1003.

Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics*, Chapman and Hall, London.

Cumming, G. (2009) Dance of the p values, *YouTube*, https://www.youtube.com/watch?v=ez4DgdurRPg

Cumming, G. (2012a) *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*, Routledge, New York.

Cumming, G. (2012b) Mind your confidence interval: how statistics skew research results*, The Conversation*, April 18.

Cumming, G. (2014) There's life beyond 0.05, *Observer* **27**(3).

Decullier, E., Lhértier, V. and Chapuis, F. (2005) Fate of biomedical research protocols and publication bias in France: retrospective cohort study, *British Medical Journal* **331**(7507):19.

Efron, B. (2012) *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Cambridge University Press, Cambridge, UK.

Fanelli, D. (2010) "Positive" results increase down the hierarchy of the sciences, *PLOS ONE* **5**(4): e10068.

Fanelli, D. (2012) Negative results are disappearing from most disciplines and countries, *Scientometrics* **90**, 891—904.

Feynman, R. (1985) *Surely You're Joking, Mr. Feynman: Adventures of a Curious Character*, W.W. Norton, New York.

Fisher, R.A. (1956) *Statistical Methods and Scientific Inference*, Oliver & Boyd, Edinburgh.

Gallistel, R. (2015) Bayes for beginners: probability and likelihood, *Observer* **28**(7).

Häggström, O. (2010) Book review: The Cult of Statistical Significance, *Notices of the American Mathematical Society* **57**, 1129—1130.

Häggström, O. (2013) Why the empirical sciences need statistics so desperately, in *European Congress of Mathematics, Krakow, 2-7 July, 2012*, (eds R. Latala et al.), EMS Publishing House, Zürich, pp 347—360.

Häggström, O. (2014a) Om statistisk signifikans, epigenetik och de norrbottniska farmödrarna, *Häggström hävdar* blog, http://haggstrom.blogspot.se/2014/02/om-statistisk-signifikans-epigenetik.html

Häggström, O. (2014b) On the value of replications: Jason Mitchell is wrong, *Häggström hävdar* blog, http://haggstrom.blogspot.se/2014/07/on-value-of-replications-jason-mitchell.html

Ioannidis, J. (2005) Contradicted and initially stronger effects in highly cited clinical research, *Journal of the American Medical Association* **294**, 218—228.

Kruschke, J. (2010) Bayesian data analysis, *Wiley Interdisciplinary Reviews: Cognitive Science* **1**, 658—676.

Lehmann, E. and Romano, J. (2008) *Testing Statistical Hypotheses* (third edition), Springer, New York.

Lehrer, J. (2010) The truth wears off, *The New Yorker*, December 13.

Mitchell, J. (2014) On the emptiness of failed replications, https://web.archive.org/web/20140708164605/http://wjh.harvard.edu/~jmitchel/writing/failed_science.htm

Morey, R., Hoekstra, R., Rouder, J., Lee, M. and Wagenmakers, E.-J. (2015) The fallacy of placing confidence in confidence intervals, *Psychonomic Bulletin & Review*, DOI 10.3758/s13423-015-0947-8.

Nieuwenhuis, S., Forstmann B.U. and Wagenmakers, E.-J. (2011) Erroneous analyses of interactions in neuroscience: a problem of significance, *Nature Neuroscience* **14**, 1105—1107.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science, *Science* **349**, 6251.

Pawitan, Y. (2001) *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press, Oxford.

Salsburg, D. (2001) *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*, W.H. Freeman, New York.

Savage, L.J. (1962) *The Foundations of Statistical Inference*, Methuen, London.

Trafimow, D. and Marks, M. (2015) Editorial, *Basic and Applied Social Psychology* **37**, 1—2.

Ziliak, S. and McCloskey, D. (2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*, The University of Michigan Press, Ann Arbor, MI.

Ziliak, S. and McCloskey, D. (2010) We agree that statistical significance proves essentially nothing: a rejoinder to Thomas Mayer, *Econ Journal Watch* **10**(1), 97—107.