I can see clearly now: reinterpreting statistical significance

Jonathan Dushoff[1], **Morgan P. Kain**[1], **and Benjamin M. Bolker**[1,2]

[1]**Department of Biology, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1 Canada**
[2]**Department of Mathematics and Statistics, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4L8 Canada**

Corresponding author:
Jonathan Dushoff[1]

Email address: dushoff@mcmaster.ca

Running Title

Statistical Clarity

Word Count

2,220

Abstract

1. Null hypothesis significance testing (NHST) remains popular despite decades of concern about misuse and misinterpretation. There are many recent suggestions for mitigating problems arising from NHST, including calls for abandoning NHST in favor of Bayesian or information-theoretic approaches. We believe that NHST will continue to be widely used, and can be most usefully interpreted as a guide to whether a certain effect can be seen *clearly* in a particular context (e.g. whether we can clearly see that a correlation or between-group difference is positive or negative).

2. We believe that much misinterpretation of NHST is due to language: significance testing has little to do with other meanings of the word "significance". We therefore suggest that researchers describe the conclusions of null-hypothesis tests in terms of statistical "clarity" rather than "significance". We illustrate our point by rewriting common misinterpretations of the meaning of statistical tests found in the literature using the language of "clarity".

3. The meaning of statistical tests become easier to interpret and explain when viewed through the lens of "statistical clarity".

4. Our suggestion is mild, but practical: this simple semantic change could ~~substantially~~ enhance clarity in statistical communication.

Key Words

Statistical philosophy; Statistical clarity; Hypothesis testing; p-value

Introduction

Statisticians and scientists have bemoaned the shortcomings of null hypothesis significance testing (NHST) for nearly a century (Cohen, 1994). Books and articles proposing the de-emphasis or abandonment of the p-value have been cited thousands of times (Cohen, 1994, Goodman, 1999, Wilkinson, 1999, Ziliak and McCloskey, 2008, Wasserstein and Lazar, 2016). These works plead for a focus on effect sizes and confidence intervals, and point out that null effects that truly have zero magnitude are unrealistic or impossible in most fields outside of the hard physical sciences (Meehl, 1990, Tukey, 1991, Cohen, 1994). Yet, p-values without confidence intervals (or even effect sizes) and references to null effects still pervade the scientific literature at all levels up to and including articles in high-impact journals.

In a meta-analysis of 356 studies Bernardi et al. (2017) found that 72% of studies contained an ambiguous use of the term "significant", 49% interpreted non-significant effects as zero effects, and 44% failed to report a comprehensible effect size. The misuse and misinterpretation of NHST is so frequent that there have been recent calls for drastically reducing (Szucs and Ioannidis, 2017) or abandoning (McShane et al., 2017) its use. Other prescriptions have included the complete abandonment of frequentist statistics (The, 2011), or the use of a stricter significance threshold (e.g. $p < 0.005$: Benjamin et al. (2018)); however, the former seems impractical, while the latter is unlikely to reduce the misuse and misinterpretation of p-values, or the publication bias imposed by any p-value threshold (Ridley et al., 2007).

Here, we argue that NHST remains useful, and that pervasive misuse can be reduced through a linguistic change: using the language of statistical "clarity" instead of statistical "significance".

The null hypothesis is false

In most biological studies, the null hypothesis is known or believed to not be strictly true. Even in cases where the null hypothesis is sensible (e.g., particle physics, Staley (2017)), NHST does not provide evidence that a difference is exactly zero. This being the case, it is worth asking how NHST has survived "if it is as idiotic as . . . long believed" Ziliak and McCloskey (2008, cited in Krämer (2011)).

The value of NHST ~~can be seen in something like a permutation-based t-test (Good, 2000, Chapter 1): it provides a simple, robust framework to ask~~ is that asking whether we can ~~tell *which* mean is bigger. More generally, testing~~ reject the null hypothesis is a proxy for asking whether we ~~clearly see a signal of how~~ see clearly *how* our data differs from it. ~~In many cases, this comes down simply to~~ For example, in a t-test, we are nominally asking whether we can ~~be confident of the~~ see a *difference* between two means, but the scientific question is whether we are confident which of the two means is *larger*; similarly, tests for whether two values are correlated are a proxy for whether we are confident about the *sign* of ~~a difference or a~~ the correlation coefficient (Robinson and Wainer, 2001). In other cases (e.g., a one-way ANOVA), it may not be simple to describe the difference we see, but NHST is still a reasonable, widely accepted way to evaluate whether an effect has been seen clearly.

The "idiocy", if any, comes in the interpretive step. A statistical fact ("we have *seen* a difference between the groups", which should immediately prompt the question "what have you learned *about* that difference?") is interpreted as a scientific fact ("there *is* a 'significant' difference between the groups), which is often seen as an end in itself: "we showed that the groups differ".

The p-value is a property of the study

We often see sentences like, "X et al. showed that there is no significant effect of Y on Z" with the implication that this effect can now be assumed to be absent (or unimportant). In fact, the sentence is erroneous even before we get to the implication: significance tests provide information about *a data set* – that is, about a study, not about the study system (Hoenig and Heisey, 2001). Indeed, a very small effect can lead to $p < 0.05$, when data is abundant (or noise is small); or a very large one can lead to $p > 0.05$ when the sample is small or noisy.

The statement “X et al. showed that Y has a statistically significant effect on Z” is similarly misleading. Frequentist statistics effectively assume that the effect is present (or at least, admit that it can’t be disproven). The question is whether it is seen in a particular data set. The statement “X et al. were able to see the effect of Y on Z” is not only more accurate, but it appropriately implies that something is missing: *What* effect did they see?

Statistical clarity

The language of “statistical clarity” could help researchers escape various logical traps while interpreting the results of NHST, allowing for the continued use of NHST as a simple, robust method of evaluating whether a data signal is clear (see Abelson (1997) for arguments for NHST). The use of “significance” to describe the results of hypothesis tests is deeply, and sometimes subtly, misleading, because it is at odds with other meanings of the word: the p-value is not an accurate gauge of whether a result is large in magnitude, biologically important, or relevant. “Clarity,” on the other hand, is an apt term for what NHST actually evaluates. Jones and Tukey (2000) and Robinson and Wainer (2001) suggest that researchers should report $p > 0.05$ using language such as “the direction of the differences among the treatments was undetermined”. This is a step in the right direction. Replacing “significance” with “clarity” takes this idea further, and has the ~~promise to substantially~~ potential to improve statistical communication.

For example, the sentence “X et al. showed that the effect of Y on Z is statistically unclear”, is noticeably awkward. It seems less like a statement about the study system, and suggests the more straightforward “did not find a statistically clear effect.” Similarly, “We did not find a clear difference in response between the control and sham groups” is both more colloquial and harder to transform into a misleading statement than "We did not find a significant difference . . . ". Bernardi et al. (2017) complained that "... sociological and social significance are sacrificed on the altar of statistical significance". Describing statistical tests in terms of clarity would allow "significant" to reclaim its common English definition and reduce conflation between statistical results and substantive significance.

Descriptions of statistical results using the language of clarity should begin with reference to the effect. For example, "The difference between the control and treatment group was not statistically clear." Table 1 shows published examples of statements that misinterpret p-values in three different ways and demonstrates how to rephrase them in the language of clarity. We have attempted to do this thoughtfully, and therefore the language on the right differs from the language on the left in many ways. The point is not that changing "significance" to "clarity" is a substantial improvement by itself, but that it can prompt rethinking and reinterpretation. We also hope that, by drawing attention to effects, the language of clarity will encourage more reporting of effect sizes and confidence intervals.

Caveats

The suggested change in language should help focus attention on improved statistical practice, but does not guarantee it. In particular, there is a danger that "statistical clarity" will eventually come to be seen as end to itself, the way that "significance" is now. The idea of "statistical clarity" will work best if it remains linked to the principles of attaching clarity to the study, not the system, and of focusing on effects.

The use of statistical clarity could be confounded with other meanings of the word clarity. We have found that scientists' understanding of our proposed use of statistical clarity is enhanced by explicitly connecting clarity statments to statements about P values.

ADD a bit about winners' curses, plausibility and the like here (and take some of it out of the Conclusions. Should we say specifically that we are agnostic about (supportive of?) the idea that NHST and/or frequentist stats are overused?

Conclusions

We believe that NHST is useful as a simple, robust way to ask whether an effect can be seen clearly in a particular data set (Robinson and Wainer, 2001), and that careful, clarity-based language can reduce misinterpretation and miscommunication.

We agree with Cohen (1994) and others (Goodman, 1999, Ziliak and McCloskey, 2008, Wasserstein and Lazar, 2016), that scientific communication and understanding will be improved by a shift away from p-values to effect sizes and confidence intervals. We argue that the use of "statistical clarity" reinforces the need for confidence intervals and effect sizes by making it clearer that bald statements about p-values are insufficient. The statement "The difference between our control and treatment groups was not statistically clear ($p = 0.30$)" is noticeably incomplete; an effect size and confidence interval are required to complete the story.

Improving language will not by itself solve all of the known problems with current statistical practice. We echo previous statements in favor of "neglected factors" (prior and related evidence, plausibility of mechanism, study design and data quality, real world benefits, novelty and other factors) (McShane et al., 2017) and reporting of *a priori* analysis of statistical power to avoid emphasis on implausibly large effects given low statistical power (the "winner's curse" Gelman and Carlin, 2014, Szucs and Ioannidis, 2017, Bernardi et al., 2017). Additionally, we support the idea of writing a statistical journal that chronicles all steps in the analytical process (Kass et al., 2016), and clearly delineating the boundary between inferences based on *a priori* hypotheses and discoveries from *post hoc* data exploration.

Whether or not our recommendations are broadly adopted by authors, reviewers, and editors, they can be useful for individual researchers who want to help themselves think clearly about NHST results. We have found that rephrasing NHST statements that we encounter (in the literature, or in seminar presentations) in terms of clarity has already helped us with both interpretation and communication.

Acknowledgments

We thank members of the Dushoff and Bolker labs for helpful comments on the first draft of the manuscript.

REFERENCES

Abelson, R. P. 1997. On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science 8*(1), 12–15.

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, et al. 2018. Redefine statistical significance. *Nature Human Behaviour 2*(1), 6.

Bernardi, F., L. Chakhaia, and L. Leopold 2017. 'Sing me a song with social significance': The (mis) use of statistical significance testing in European sociological research. *European Sociological Review 33*(1), 1–15.

Cohen, J. 1994. The earth is round (p $<$.05). *American Psychologist 49*(12), 997.

Gelman, A. and J. Carlin 2014. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science 9*(6), 641–651.

Gelman, A. and H. Stern 2006. The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician 60*(4), 328–331.

Good, P. 2000. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer.

Goodman, S. N. 1999. Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine 130*(12), 995–1004.

Hoenig, J. M. and D. M. Heisey 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician 55*(1), 19–24.

Jones, L. V. and J. W. Tukey 2000. A sensible formulation of the significance test. *Psychological Methods 5*(4), 411.

Kass, R. E., B. S. Caffo, M. Davidian, X.-L. Meng, B. Yu, and N. Reid 2016. Ten simple rules for effective statistical practice. *PLoS Computational Biology 12*(6), e1004961.

Krämer, W. 2011. The cult of statistical significance–what economists should and should not do to make their data talk. *Schmollers Jahrbuch 131*(3), 455–468.

McShane, B. B., D. Gal, A. Gelman, C. Robert, and J. L. Tackett 2017. Abandon statistical significance. *The American Statistician 70*.

Meehl, P. E. 1990. Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports 66*(1), 195–244.

Ridley, J., N. Kolm, R. Freckelton, and M. Gage 2007. An unexpected influence of widely used significance thresholds on the distribution of reported p-values. *Journal of Evolutionary Biology 20*(3), 1082–1089.

Robinson, D. H. and H. Wainer 2001. On the past and future of null hypothesis significance testing. *ETS Research Report Series 2001*(2).

Staley, K. W. 2017. Pragmatic warrant for frequentist statistical practice: the case of high energy physics. *Synthese 194*(2), 355–376.

Szucs, D. and J. Ioannidis 2017. When null hypothesis significance testing is unsuitable for research: a reassessment. *Frontiers in Human Neuroscience 11*, 390.

The, B. 2011. Significance testing: are we ready yet to abandon its use? *Current Medical Research and Opinion 27*(11), 2087–2090.

Tukey, J. W. 1991. The philosophy of multiple comparisons. *Statistical Science*, 100–116.

Wasserstein, R. L. and N. A. Lazar 2016. The ASA's statement on p-values: context, process, and purpose.

Wilkinson, L. 1999. Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist 54*(8), 594.

Ziliak, S. and D. N. McCloskey 2008. *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. University of Michigan Press.

Language from published articles	**Rewritten using "clarity"**
***Accepting the null hypothesis** ($p > 0.05 \not\Rightarrow$ **no effect**)*	
Toxins accumulate after acute exposure but have no effect on behaviour	Toxins accumulate after acute exposure but their effects on behaviour are statistically unclear
There was no effect of elevated carbon dioxide on reproductive behaviors	The effect of elevated carbon dioxide on reproductive behaviors was statistically unclear
The finding that species richness showed no significant relationship with the area of available habitat is surprising because richness is usually strongly influenced by landscape context	Although species richness is usually strongly influenced by landscape context, we were unable to find a statistically clear relationship in this study
Inferring weak effects from large p-values (Wasserstein and Lazar, 2016)	
. . . differences between treatment and control groups were nonsignificant, with P values of at least 0.3, and most in the range $0.7 \leq P \leq 0.9$.	. . . differences between treatment and control groups were not statistically clear (all $P > 0.05$) [*since smallness is no longer implied the authors might now think of adding confidence intervals.*]
The difference between "clear" and "not clear" is not clear (Gelman and Stern, 2006)	
This correlation was significant in males ($\rho = 0.35$, P <0.05) but not females ($\rho = 0.35$, NS). . . . [*The authors later write as though they have demonstrated a difference between males and females*]	Although males and females show the same correlation coefficient ($\rho = 0.35$), the sign of the coefficient is statistically clear only in males . . . [*Again, this phrasing may suggest to the authors that confidence intervals are called for.*]
. . . risk of low BMD [bone mineral density] remained greater in HCV-coinfected women versus women with HIV alone (adjusted OR 2.99, 95% CI 1.33–6.74), but no association was found between HCV coinfection and low BMD in men (adjusted OR 1.26, 95% CI 0.75–2.10). . . . The precise mechanisms for the association between viral hepatitis and low BMD in HIV-infected women but not men remain unclear.	. . . risk of low BMD [bone mineral density] remained greater in HCV-coinfected women versus women with HIV alone (adjusted OR 2.99, 95% CI 1.33–6.74), but the association between HCV coinfection and low BMD in men was not statistically clear (adjusted OR 1.26, 95% CI 0.75–2.10). . . . Pursuing biological differences between women and men in the effect of HIV on BMD would be premature given these results.

Table 1. Examples of misleading language in peer-reviewed papers (citations available by request), and revisions using our proposed language of clarity.