In support of null hypothesis significance testing

Michael Mogie

Centre for Mathematical Biology, Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK (m.mogie@bath.ac.uk).

Recd 21.08.03; Accptd 10.09.03; Online 22.10.03

Many criticisms have been levelled at null hypothesis significance testing (NHST). It is argued here that although there is reason to doubt that data subjected only to NHST have been subjected to sufficient analysis, the search for clear answers to well-formulated questions derived from substantive hypotheses is well served by NHST. To reliably draw inferences from data, however, NHST may need to be complemented by additional methods of analysis, such as the use of confidence intervals and of estimates of the degree of association between independent and dependent variables. It is argued that these should be seen as complements of, rather than as substitutes for, NHST since they do not directly test the strength of evidence against a null hypothesis.

Keywords: null hypothesis significance testing; p -value; confidence intervals

1. INTRODUCTION

Given its widespread utilization in data analysis, it is clear that most biologists assume that null hypothesis significance testing (NHST) provides a reliable means of interpreting the world. NHST has many critics, however. Criticisms are particularly evident in the psychology literature, where NHST has been variously described as 'a disaster' (Hunter 1997), as involving 'two-valued (traditional, but silly) logic' (Harris 1997) and as 'surely the most bone-headedly misguided procedure ever institutionalised in the rote training of science students' (Rozeboom 1997). For an overview of the debate, see Harlow *et al.* (1997) and the open peer commentary in volume 21 of *Behavioural and Brain Sciences* (1998, pp. 194–239). Concern about NHST is not as apparent in the biological literature (but see Johnson 1999; Anderson *et al.* 2000; Poole 2001), with many journals giving no or scant instructions to authors on the presentation of statistics. Exceptions include *Behavioural Ecology*, which reminds authors to draw appropriate conclusions from non-significant results, and the *Proceedings of The Royal Society of London Series B: Biological Sciences*, which requires authors to state that power calculations were undertaken to ensure that sample sizes were adequate to test the hypothesis being investigated.

NHST compares sample data with a hypothetical population whose distribution is specified by the null hypothesis (NH). It involves the calculation of a conditional probability value (henceforth, p -value) that is the long-run

relative frequency with which the data observed, or data more extreme than those observed, would be obtained if the NH is true—i.e. data are collected randomly, and the assumed model is correct (e.g. that data are distributed normally). Perceived problems with NHST include the widespread use of implausible and/or uninteresting NHs, and fixed-level testing (e.g. reject the NH if $p \leq 0.05$, otherwise do not reject). It is argued here that these problems have been exaggerated, and that there is no clear case for biologists to adopt the pessimistic approach to NHST that is prevalent in much psychology literature. In particular, many of the problems associated with NHST can be ameliorated with additional methods of data analysis, for example through the complementary use of confidence intervals (CIs) and of estimates of the degree of association between independent and dependent variables.

2. THE LOGIC OF IMPLAUSIBLE AND UNINTERESTING NULL HYPOTHESES

The currency of NHST is the p -value, and the pivotal point is the NH. Often, the p -value is calculated for a NH that is implausible. Implausibility is particularly likely when NHs are point hypotheses (which propose exact values for parameters), since these identify circumstances that are so specific that they are unlikely to be true (e.g. a single precise value for a mean, or identical means in two populations; see Swaddle & Cuthill 1997). This has led critics of NHST to question whether p -values are meaningful when NHs are unlikely to be true (Johnson 1999). This criticism is mistaken to the extent that asserting the truth or otherwise of the NH is not a role of NHST. It does have a function of asserting the probability of the observed data, or more extreme data, given the sampling distribution specified by the NH. But this p -value is a mathematical property of this distribution rather than a reflection of the 'truth content' of the NH (Batanero 2000).

Critics may dismiss this as a correct but weak defence of implausible NHs. A stronger defence can be constructed by recognizing that an implausible NH can contribute to the completeness of a statistical analysis by acting as a logical complement of a plausible alternative hypothesis. But this defence exposes a weakness as well as a strength of NHST. A useful way to appreciate this is to place the null and alternative statistical hypotheses (respectively, H_0 and H_1) in the context of Chow's (1996) description of the nature of a scientific investigation, which can be summarized as involving a substantive hypothesis, a research hypothesis, an experimental hypothesis, an alternative statistical hypothesis and a NH. A substantive hypothesis is a speculative account of a phenomenon that usually cannot be investigated directly, for example that the morphology of insect-pollinated flowers has been influenced by the need to attract insect pollinators. An observable implication of the substantive hypothesis gives a research hypothesis (altering the morphology of insect pollinated flowers will reduce their attraction to insect pollinators). Identifying a well-defined dependent variable of the research hypothesis gives the experimental hypothesis (the number of pollinator visits will be greater for unaltered than for altered flowers). An implication of this forms the alternative statistical hypothesis (the mean number of visits for altered flowers (μ_1) will be less than that for unaltered flowers (μ_2); that is, $H_1: \mu_1 < \mu_2$). This hypothesis is used to identify the NH.

There are two options for H_0 that are obvious. The first is to formulate it as a point hypothesis: the mean number of visits will be *the same* for unaltered and altered flowers ($H_0: \mu_1 - \mu_2 = 0$). The benefit of this option is that only a point NH can specify a distribution against which samples can be compared. A potential cost is that it gives a one-tailed test that leaves the hypothesis $H_2: \mu_1 > \mu_2$ untested. That is, the investigator ends up with null and alternative statistical hypotheses that are logical contraries of each other (statistical hypotheses are logical contraries if, and only if, (a) if one is true the other is false, and (b) but both might be false). It would be unwise to adopt this option unless there was a strong justification for assuming that (b) was extremely unlikely, that is, unless there was a strong justification for rejecting *a priori* $\mu_1 > \mu_2$. Otherwise, the investigator should ensure that H_0 and H_1 are logical complements of each other (statistical hypotheses are logical complements if, and only if, (a) if one is true the other is necessarily false, and (b) if one is false the other is necessarily true). But when $H_1: \mu_1 < \mu_2$, this requires that the NH is that the mean number of visits will be *the same or less* for unaltered than for altered flowers ($H_0: \mu_1 - \mu_2 \geq 0$). It is here that a weakness of NHST is exposed since there is a considerable cost to this move. This is that H_0 , being no longer a point NH, cannot specify the distribution of the test statistic against which the sample is to be compared, hence it cannot provide a p -value (Frick 1998). This problem can be resolved by retreating to the point NH, $H_0: \mu_1 - \mu_2 = 0$, and forcing logical complementarity by replacing $H_1: \mu_1 < \mu_2$ with $H_1: \mu_1 - \mu_2 \neq 0$. This gives a two-tailed test of a NH that can specify a distribution and therefore provide a p -value. But the cost of making this move is that the logical flow of argument from substantive hypothesis to alternative statistical hypothesis has been lost. We have ended up with a statistical test that appears to lack utility since, with respect to theory corroboration/falsification, we are interested in the abandoned $H_1: \mu_1 < \mu_2$ and $H_0: \mu_1 - \mu_2 \geq 0$, not in their uninteresting replacements $H_1: \mu_1 - \mu_2 \neq 0$ and $H_0: \mu_1 - \mu_2 = 0$.

This is not a reason to abandon NHST, although it is a reason to doubt that data subjected only to NHST have been subjected to sufficient analysis. If the result of the statistical test is that the NH is not rejected then our confidence in the substantive hypothesis should be diminished (but see § 3). If the NH is rejected, further steps can be taken to infer whether it is rejected in favour of $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$. By far the most straightforward step is to examine the CI, which will contain values either only to the left or only to the right of zero. The former will justify the inference $\mu_1 < \mu_2$, the latter the inference $\mu_1 > \mu_2$. Hence, judicious use of the CI has, in effect, restored the logical flow of argument from the substantive hypothesis to the alternative statistical hypothesis.

3. BEHAVIOURAL RULES VERSUS STATISTICAL INFERENCE

The discussion in § 2 needs to be qualified. The strength of evidence against a sample being representative of the hypothetical population specified by the NH is expressed in terms of the p -value. But sample size influences p -values, with the latter decreasing as the former increases. One consequence of this is that a difference

between a sample and a NH that is (or is not) shown to be statistically significant may not (or may) be of biological importance. Failure by some investigators to make this distinction provides critics with a further argument against NHST. But criticism here should be directed at investigators who deploy NHST uncritically rather than at NHST itself.

The potential for an uncritical deployment of NHST arises because the most common approach to NHST is to identify a significance level, α , above which a NH is not rejected and at or below which it is rejected. This is fixed-level testing. Commonly, $\alpha = 0.05$, a value that results in a rate for type I errors (the rejection of a true NH) of *ca.* 5%. Critics point out that fixed-level testing leaves no room for statistical inference since its practitioners are essentially directed to behave in a certain way (Goodman 1999). But the force of this criticism dissipates once it is recognized that this is with respect only to the NH. In a well-designed study, support for the alternative statistical hypothesis will be inferred from rejection of the NH, at which point the knowledge and experience of investigators may result in them inferring support for the substantive hypothesis. Knowledge and experience is crucial here since fixed-level testing does not explain why compatibility of the data with the alternative statistical hypothesis or the substantive hypothesis is, or is not, indicated. For example, a failure to reject the NH may be due to low power, and as Mulaik *et al.* (1997) note, systematic error and experimental artefact should always be considered a possibility when rejecting the NH.

4. ALTERNATIVES TO NHST?

There are ways other than NHST of assessing evidence. For example, the calculation of a likelihood ratio also achieves statistical hypothesis comparison but without binding its practitioners to a behavioural rule (Gregoire 2001). Its utility is that it indicates the extent to which the evidence supports a statistical hypothesis over alternatives. But it provides no information about the extent to which the evidence supports a statistical hypothesis *per se*. Hence, there is no clear pathway of inference from a likelihood ratio to a substantive hypothesis except for the case when the statistical hypotheses being compared are logical complements of one another. With this important exception, likelihood ratios are best viewed as useful complements of, rather than substitutes for, NHST.

CIs are also useful complements of NHST, as are estimates of the degree of association between the independent and dependent variables (e.g. eta-squared (η^2), omega-squared (ω^2), r^2). The latter provide measures of the size of effects (and hence of their biological importance). CIs enable inference both about the direction of effects shown to be significant following NHST, and about the extent to which plausible values of a variable fall into a range that is of biological importance (Sims & Reid 1999), providing a relatively straightforward way of distinguishing between biological importance and statistical significance. But it is arguable that approaching data analysis through NHST entails a different kind of basic logic from approaching it through CIs, with NHST testing the strength of evidence against a NH and with CIs directly describing a pattern of population parameters (Loftus 1995). Owing to this, it is arguable that the

reporting of CIs does not provide a substitute for NHST (but see Poole 2001), although it does provide an important additional tool with respect to drawing inferences from data. The use of both is to be recommended.

Acknowledgements

The author thanks two anonymous reviewers for comments on an earlier draft of this manuscript.

- Anderson, D. R., Burnham, K. P. & Thompson, W. L. 2000 Null hypothesis testing: problems, prevalence, and an alternative. *J. Wildl. Mgmt* **64**, 912–923.
- Batanero, C. 2000 Controversies around the role of statistical tests in experimental research. *Math. Think. Learn.* **2**, 75–98.
- Chow, L. S. 1996 *Statistical significance: rationale, validity and utility*. London: Sage.
- Frick, R. W. 1998 Chow's defense of null-hypothesis testing: too traditional? *Behav. Brain Sci.* **21**, 199.
- Goodman, S. N. 1999 Towards evidence-based medical statistics. 1: the P value fallacy. *Ann. Internal Med.* **130**, 995–1004.
- Gregoire, T. G. 2001 Biometry in the 21st century: whither statistical inference? Keynote address. In *Conf. Forest Biometry Information Sci.*, June 2001. See <http://cms1.gre.ac.uk/conferences/iufro/proceedings/gregoire.pdf>.
- Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (eds) 1997 *What if there were no significance tests?* Mahwah, NJ: Erlbaum.
- Harris, R. J. 1997 Significance tests have their place. *Psychol. Sci.* **8**, 8–11.
- Hunter, J. E. 1997 Needed: a ban on the significance test. *Psychol. Sci.* **8**, 3–7.
- Johnson, D. H. 1999 The insignificance of statistical significance testing. *J. Wildl. Mgmt* **63**, 763–772.
- Loftus, G. R. 1995 Data analysis as insight: reply to Morrison and Weaver. *Behav. Res. Meth. Instrum. Comput.* **27**, 57–59.
- Mulaik, S. A., Raju, N. S. & Harshman, R. A. 1997 There is a time and a place for significance testing. In *What if there were no significance tests?* (ed. L. L. Harlow, S. A. Mulaik & J. H. Steiger), pp. 65–115. Mahwah, NJ: Erlbaum.
- Poole, C. 2001 Low P-values or narrow confidence intervals: which are more durable? *Epidemiology* **12**, 291–294.
- Rozeboom, W. W. 1997 Good science is abductive, not hypothetico-deductive. In *What if there were no significance tests?* (ed. L. L. Harlow, S. A. Mulaik & J. H. Steiger), pp. 335–391. Mahwah, NJ: Erlbaum.
- Sims, J. & Reid, N. 1999 Statistical inference by confidence intervals: issues of interpretation and utilization. *Phys. Ther.* **79**, 186–195.
- Swaddle, J. P. & Cuthill, I. C. 1997 The biological relevance of testing for perfect asymmetry. *Anim. Behav.* **54**, 475–476.