Austral ECOLOGY A Journal of ecology in the Southern Hemisphere

Austral Ecology (2009) **34**, 447–468

Misprescription and misuse of one-tailed tests

CELIA M. LOMBARDI[1]* AND STUART H. HURLBERT[2]

[1]*Consejo Nacional de Investigaciones Científicas y Técnicas, Museo Argentino de Ciencias Naturales, Av. Angel Gallardo 470, C1405DJR Buenos Aires, Argentina (Email: celia7@sigmaxi.net); and*

[2]*Department of Biology, San Diego State University, San Diego, California, USA*

Abstract One-tailed statistical tests are often used in ecology, animal behaviour and in most other fields in the biological and social sciences. Here we review the frequency of their use in the 1989 and 2005 volumes of two journals (*Animal Behaviour* and *Oecologia*), their advantages and disadvantages, the extensive erroneous advice on them in both older and modern statistics texts and their utility in certain narrow areas of applied research. Of those articles with data sets susceptible to one-tailed tests, at least 24% in *Animal Behaviour* and at least 13% in *Oecologia* used one-tailed tests at least once. They were used 35% more frequently with nonparametric methods than with parametric ones and about twice as often in 1989 as in 2005. Debate in the psychological literature of the 1950s established the logical criterion that one-tailed tests should be restricted to situations where there is interest only in results in one direction. 'Interest' should be defined; however, in terms of collective or societal interest and not by the individual investigator. By this 'collective interest' criterion, all uses of one-tailed tests in the journals surveyed seem invalid. In his book *Nonparametric Statistics,* S. Siegel unrelentingly suggested the use of one-tailed tests whenever the investigator predicts the direction of a result. That work has been a major proximate source of confusion on this issue, but so are most recent statistics textbooks. The utility of one-tailed tests in research aimed at obtaining regulatory approval of new drugs and new pesticides is briefly described, to exemplify the narrow range of research situations where such tests can be appropriate. These situations are characterized by null hypotheses stating that the difference or effect size does not exceed, or is at least as great as, some 'amount of practical interest'. One-tailed tests rarely should be used for basic or applied research in ecology, animal behaviour or any other science.

Key words: directional tests, hypothesis testing, non-parametric tests, one-sided tests, S. Siegel, significance tests.

INTRODUCTION

For roughly 50 years debate has gone on in the scientific and statistical literature over the use of one-tailed statistical tests. The debate has concerned primarily the criteria for determining when their use is appropriate. The frequency of their use has varied over time and varies markedly from one scientific discipline to another. Periodically, authors have noted their general inappropriateness. Burke (1953) reported 'a disturbing increase in the use of one-tailed tests' in the psychological literature following publication of misguided advice from Marks (1951) and Jones (1952). Fleiss (1987) noted in the medical literature 'a growing tendency towards using one-tailed significance tests in clinical trials', a use he regarded as 'inappropriate'. Peace (1989, 1991) determined that 34% of experimental medical studies published during 1975–1988 used one-tailed tests and concluded 'that a sizable segment of the research community believes that there are settings where one-sided *p* values are appropriate'.

*Corresponding author.

Accepted for publication July 2008.

In the fields of ecology and evolution, recent reviews have claimed (Gaines & Rice 1990; Rice & Gaines 1994) or implied (Underwood 1990, 1997) that one-tailed tests are not used as often as they should be. Freedman *et al.* (1991, 1998) say it doesn't make much difference which type of test is used as long as one makes clear which is used. Many have heeded the first part of that prescription, and many fewer the second. For example, Feinstein (1974) reported that 'tailedness' was not indicated for the 'vast majority' of 757 statistical procedures found in 404 articles published in 1973 in five general medical periodicals. McKinney *et al.* (1989) observed that of 56 medical articles using Fisher's exact test, 59% did not indicate whether they were using a one- or two-tailed version; and of 20 psychology articles using the *t*-test, 85% neglected to provide this information (Pillemer 1991). Our own interest in this issue was stimulated initially by observation of frequent misuse of one-tailed tests in the animal behaviour and ecological literature. Unaware of the depth of this tar pit, we undertook to prepare a short note on it.

In the course of the debate over criteria, many technical and conceptual issues relating to the

doi:10.1111/j.1442-9993.2009.01946.x

interpretation and conduct of one-tailed tests also have been raised. Errors concerning these have been published in abundance, confused the debate over criteria, and almost certainly led to publication of many incorrectly calculated *P* values for one-tailed tests. The objective of this report is to clarify and resolve some key issues relating to the use of one-tailed tests. In doing so, we come to the conclusion that their use in all basic and most applied scientific research is inappropriate. Their use is appropriate, however, in certain types of applied research where the null hypothesis is other than 'no effect' and where the *P* values yielded by statistical tests might determine actions to be taken, as implicit in the Neyman-Pearson decision theoretic framework for statistical hypothesis testing.

We begin this analysis by establishing our terminology, notation and concepts, by defining the nature of the central problem and by commenting on a subsidiary technical issue, the relationship of *P* values to 'directional' conclusions. We then state what we, and some others, believe to be the only acceptable criterion for use of one-tailed tests, that of the collective interest of science and society. Next we present and analyse the results of surveys of use of one-tailed tests in two journals, *Animal Behaviour* (AB) and *Oecologia* (OE), in 1989 and 2005. Then we critique the treatment of the issue in both older and recent statistics books. In the final sections, we discuss some special types of applied research where one-tailed tests are quite appropriate.

NATURE OF THE PROBLEM

Basic concepts

We have measured the same dependent variable for two samples *a* and *b*, and wish to make an inference about a difference that might exist between the respective real or abstract sampling universes, *A* and *B*. Frequently we wish to make an inference about the difference between the unknown means, $\mu_A - \mu_B$, on the basis of the difference between the observed sample means, m_A and m_B. If we explicitly define $\delta = \mu_A - \mu_B$ and $d = m_A - m_B$, then our interest might consist in asking whether the true difference departs from some particular magnitude ($\delta = c$?) or whether the difference exceeds ($\delta > c$?) or is less than ($\delta < c$?) some particular magnitude. The difference, *c*, might be zero, might be an absolute increment or decrement, or might be a percentage of one of the means, for example, μ_A. To answer any of these three questions a test or assessment of significance, such as provided by a *t*-test, is helpful. This calls for establishment of a null hypothesis (H_0) and specification of the maximum acceptable probability (α) of rejecting H_0 when the H_0 is true. Then one estimates the probability (*P*) of obtaining a value of *d* representing a departure from H_0 that is as extreme as or more extreme than the observed *d*. The null hypothesis, H_0, is the one that we must in some sense find evidence against in order to show the tenability of the alternative hypothesis, H_1. The H_1 most often includes all possible relations of μ_A and μ_B not anticipated by H_0.

In the classical decision theoretic framework, when *P* is sufficiently small, that is $\leq \alpha$ the investigator rejects H_0 in favour of H_1. If this is carried out and if H_0 is in fact true, however, the investigator is said to commit a type I error. If the investigator fails to reject H_0 when H_1 is true, a type II error is said to occur. Of course no actual error is made if the investigator withholds judgment in this latter case and does not affirmatively accept the high *P* value as evidence in favour of H_0 over H_1.

Test 'tailedness'

Our problem begins with the fact that in analysing this simple two-sample data set we usually would like to discriminate among the three possible situations: $\delta = c$, $\delta > c$ and $\delta < c$. In virtually all basic research we desire this discriminatory power. Only in certain narrow types of applied research, to be discussed at the end of this article, are we satisfied merely with distinguishing one of these possibilities from the other two taken collectively.

Let us consider the common situation where $c = 0$ and any of three different sets of hypotheses can be analysed with a *t*-test:

$$\text{Set 1: } H_0{:}\delta = 0,\ H_1{:}\delta \neq 0$$
$$\text{Set 2: } H_0{:}\delta \geq 0,\ H_1{:}\delta < 0$$
$$\text{Set 3: } H_0{:}\delta \leq 0,\ H_1{:}\delta > 0$$

Conventionally, we apply a *t*-test to Set 1 if we are interested in detecting any difference, positive or negative, between μ_A and μ_B. Such a test is variously referred to as a two-tailed, two-sided, or nondirectional test. If our concern is to determine only whether $\mu_A < \mu_B$ or only whether $\mu_A > \mu_B$, then we apply a *t*-test either to Set 2 or Set 3. Such a test is termed a one-tailed, one-sided, or directional test. The advantage of a one-tailed test is that, for fixed α, it has greater power than the two-tailed test for detecting a difference in the direction tested.

There has long existed some confusion in this terminology. The reason is that most discussion of the one-tailed *versus* two-tailed issue has been carried out in the context of the *t*-test. As there is a one-to-one correspondence, under H_0, between the symmetric *t*-distribution, centered on $t = 0$, and the symmetric sampling distribution of *d*, centered on $\delta = 0$, writers have been able to be imprecise without penalty. A 'tail'

doi:10.1111/j.1442-9993.2009.01946.x

of negative t values corresponded to a 'tail' of negative d values, and likewise for the positive half of the distributions. Critical regions for rejection of null hypotheses can be defined or seen for either the t- or the d-distribution.

Confusion can develop when the test statistic is one such as χ^2, which might be used to compare proportions in a 2×2 contingency table. For a χ^2-test the critical or rejection region usually occupies only a right hand portion of the distribution of the test statistic regardless of which of the three sets of hypotheses (above) is being tested. As P can range up to 1.0, this portion can be much larger than is implied by a 'tail'. All tests will be '1-tailed' in that the rejection region will not include the lower or left hand tail of the distribution of the test statistic (χ^2); but the test for Set 1, though not the tests for Sets 2 and 3, will be two-tailed in relation to the sampling distribution of d. As δ is the parameter of prime interest to the investigator, we will follow widespread usage and define one- and two-tailed tests on the basis of whether or not the null hypothesis is of the form $H_0{:}\delta \leq c$ or $H_0{:}\delta \geq c$ (one-tailed tests) or is of the form $H_0{:}\delta = c$ (two-tailed test).

One-tailed procedures are available, of course, for a great variety of other testing situations in addition that of the two-sample problem. Commonest of these would be (i) the comparison of a single sample to a hypothetical standard; (ii) the testing of correlation or regression coefficients; and (iii) 'ordered alternatives' tests where the number of groups or treatments being compared numbers more than two. The basic conceptual issues are the same.

Redefining α and/or H_0

Interpretation of the typical one-tailed test should be straightforward. For example, if $H_0{:}\delta \leq 0$, observed $d = -10$ and $P > 0.50$, then one concludes only that δ is very unlikely to be positive. If P is much larger than 0.50, however, it will be evident that δ probably is negative, that is strongly in the direction in which we supposedly have no interest. For example, if $P = 0.98$ that tells us that had we used a two-tailed test of $H_0{:}\delta = 0$, we would have obtained $P = 0.04$.

Confusion sometimes exists as to how such unexpected or unpredicted results are to be interpreted. This reflects certain variations in how H_0 and α are defined in one-tailed testing situations. The customary approach is as outlined above. A strong departure in the unexpected direction is regarded as possible but of no interest. What often is not made clear with this approach is that the definition of α is not the same as in the standard two-tailed test. It has changed from 'the probability of rejecting $H_0{:}\delta = 0$ when it is true' to 'the maximum probability of rejecting $H_0{:}\delta \leq 0$ when it is true'. In the one-tailed test, the probability of a type I error will equal α if $\delta = 0$ but be less than α if $\delta < 0$. Alpha becomes only an upper bound (e.g. Burke 1953; Meehl 1967).

Occasionally authors imply that when a one-tailed test has been selected *a priori* and a result strongly in the unpredicted direction has been obtained, the result must be considered a chance outcome and not evidence against the idea that $\delta = 0$ (e.g. Glass & Hopkins 1984; Koch & Gillings 1988). Such an interpretation is rarely appropriate. It would be valid only in those rare situations where we can justify testing $H_0{:}\delta = 0$ with a one-tailed test on the grounds that it is absolutely certain that a result in the unpredicted direction is impossible. In such a situation, the definition of α reverts to the standard one; it is no longer merely an upward bound.

The decision theoretic framework

In the classical decision theoretic framework, one specifies *a priori* that a decision will be taken to either 'reject' or 'accept' H_0 and one specifies α, the maximum probability of a type I error that one is willing to accept (Neyman & Pearson 1933). If it turns out that $P \leq \alpha$, then we reject H_0 and state that the difference between d and c (the δ specified by H_0) is 'significant'. If $P > \alpha$, we say this difference is 'not significant'. This framework gives a clear formalization of the concepts of type I error, type II error and power. It also is a useful general framework for quality control procedures and for those rare analyses intended to lead directly to black-and-white, reject-not reject decisions.

For other types of investigations, including virtually all basic and most applied scientific studies, the decision theoretic framework is deficient as a general one for the implementation and interpretation of statistical analyses (Fisher 1960, 1971). The many who have used it and continue to use it as such have been led, perhaps often against their own intuition, to irrational analyses, awkward language and illogical conclusions. A companion article analyses its weaknesses and their role in fomenting use of one-tailed tests (Hurlbert & Lombardi 2009). Following its publication, the decision theoretic framework developed a large following among mathematical statisticians, especially in the USA. Salsburg (1992) claimed that Neyman never championed that framework after the mid 1930s and seemingly 'agreed in principle with most of Fisher's criticisms' of it. In fact Neyman presented the framework *in extenso* in his textbook (Neyman 1950: 250ff) and briefly but favourably in a later historical essay (Neyman 1976). But Fisher's original, if unclear, arguments against the decision theoretic framework have received extensive support from statistics scholars for many decades (Hurlbert & Lombardi 2009).

doi:10.1111/j.1442-9993.2009.01946.x

Options following a surprise

To discuss the consequences of the 'tailedness' of a test, we use the decision theoretic framework, despite its inadequacies for other purposes (Hurlbert & Lombardi 2009), as the language of that framework is familiar to statisticians and scientists.

All scientists certainly wish to keep the risk of type I errors low regardless of what value they might individually assign to 'low' and of the formality of the procedures they wish to use in the task. The costs of such errors can include the slowing of advance in knowledge, decrease in quality and increase in price of the scientific literature, economic and other damage caused by introduction of worthless or unsafe products or technologies, embarrassment, and so on.

Our principal protection against type I errors comes from generally selecting for study, only those independent variables we regard as likely to have an effect on or to help explain variation in our dependent variables. Findings of no effect or no relationship are rarely an objective, nor can high P values be taken as evidence that there is no effect. In these situations, then, our principal concern is to get an acceptably precise estimate of δ, not to avoid a type I error. Nevertheless, for reasons of parsimony and consistency as well as to anticipate those situations when H_0 is true, we do have to show an acceptably low P value before concluding H_0 is false. And in defining 'acceptably low' we come right back to the idea of fixing α.

In the case of our simple two-sample data set, let us say we wish to set, for the overall decision procedure, the maximal probability of a type I error at $\alpha = 0.05$. There are various ways in which this α might unintentionally be increased. For example, we could use t-tests to test separately all three null hypotheses given earlier. For $H_0{:}\delta = 0$ a two-tailed test would be used and for $H_0{:}\delta \leq 0$ and $H_0{:}\delta \geq 0$, one-tailed tests would be used. Logically, of course, if either one-tailed test is carried out, the two-tailed test is redundant, its outcome containing no information not provided in the one-tailed test. But there are situations that might tempt the naive investigator into carrying out two of the three possible tests in a way that leads to an unacknowledged increase in the probability of a type I error.

In the most common of these situations the investigator has more interest in one tail of the distribution than another, for example, more interest in positive values of d than in negative ones. Perhaps theory or prior evidence suggests d should be positive, perhaps only a positive value will validate the investigator's most cherished hypothesis, or perhaps only a positive value will indicate that a new product or process is superior to an older one.

In any such situation the investigator who understands the concept of type I error faces a mild dilemma. The investigator's primary interest might seem best served by carrying out the one-tailed test corresponding to $H_0{:}\delta \leq 0$. The decision to do that would appropriately be made before the collection of data. If d turns out to be positive, the one-tailed t-test of $H_0{:}\delta \leq 0$ will yield a P value equal to exactly half of the P value that would have been obtained by the two-tailed test of $H_0{:}\delta = 0$ (e.g. at least in the case of a t-test). This simply reflects the fact that, for fixed α, power to detect a positive δ is greater for the corresponding one-tailed test than it is for the two-tailed test.

The dilemma arises from the possibility that d might turn out to be negative, that is, of a sign opposite that anticipated or desired. If the investigator had decided *a priori*, and on grounds of primary interest and statistical power, that the one-tailed test for $H_0{:}\delta \leq 0$ was to be used, a later determination that d is negative leaves four options (Goldfried 1959):

> *Option 1*: Report the value of P (which will be >0.5), argue or imply that the unexpected result contains no information that is useful, interesting, or important to the investigator, her vaguely defined audience, or science generally, and that therefore further statistical testing is not needed. If space permits, provide all the information needed by readers who may want to carry out a 2-tailed test.

> *Option 2*: Ignore the first value (>0.50) of P obtained and calculate and report the P value obtained in carrying out the 2-tailed test on $H_0{:}\delta = 0$. If the same value (α_i) is used in each of these two tests, then with respect to $H_0{:}\delta = 0$, the overall α is 1.5 α_i, that is 0.075 [= (0.5)(0.1) + (0.5)(0.05)], if $\alpha_i = 0.05$. In other words, P values will be 'underestimated'.

We can modify this two-stage decision procedure if we wish to set overall α at any desired level. This can be carried out by adjusting the calculated P values upwards by multiplying them by 1.5. Or we can achieve exactly the same thing by carrying out both the one-tailed and the two-tailed test with $\alpha_i = 2\alpha/3$.

The only cost of this decision procedure is slightly reduced power for detecting differences in the expected direction. The overall procedure represents a two-tailed test with unequal apportioning of α to the two tails of the t-distribution, a procedure that, in different forms, occasionally has been recommended though it lacks any logical rationale (S.H. Hurlbert & C.M. Lombardi 2008, unpub. data).

> *Option 3*: Ignore the first value (>0.50) of P obtained and calculate and report the P value obtained in carrying out the 1-tailed test for the unexpected direction, that is, for $H_0{:}\delta \geq 0$. If the same value (α_i) is used in each of these two tests, then with respect to $H_0{:}\delta = 0$, we now have an overall $\alpha = 2\alpha_i$, that is, 0.10 if $\alpha_i = 0.05$. This decision procedure is exactly

doi:10.1111/j.1442-9993.2009.01946.x

equivalent to carrying out a classical 2-tailed test, where α is apportioned equally between the two tails of the t-distribution.

The cost of this option is that overall α is doubled relative to its value in either test considered by itself. This is true even if d turns out to be in the expected direction and the prescribed second one-tailed test is not actually carried out. This doubling of α is prevented if, as a matter of procedure, the P value obtained for the test in the observed direction is itself doubled.

Option 4: Acknowledge that there may be a real difference in the unexpected direction and refrain from any testing for significance. If there is sufficient interest in testing for a difference in the unexpected direction, repeat the study, and analyze the new data with either a 2-tailed test or the other 1-tailed test. The cost of this option is the time and resources required to repeat the study. The main benefit is a truly celestial level of 'statistical purity'.

We began this section with an investigator who had decided *a priori* to use a one-tailed test to test $H_0: \delta \leq 0$ and who then had to consider what to do when d turned out negative. Can at this point any general recommendations be made? We suggest Option 4 will never be an appropriate one as, except where the cost of repeating the study is trivial, it represents a great waste of resources and information. Options 2 and 3 represent different types of two-tailed tests. If either is selected the investigator is simply admitting that she should have explicitly specified use of a two-tailed test when she began the study. She also is unlikely to be candid about how her α has been affected. The appropriateness of Option 1 will depend entirely on the force of the arguments in its favour that can be developed for general classes of situations or for any specific situation; rarely, in our opinion, will this force be great. Before considering those arguments let us review a few technical matters concerning two-tailed tests.

Referent statement for P?

An important aspect of two-tailed tests was called to our attention by Kaiser (1960) that had been much neglected in earlier literature and that has continued to be neglected by both statisticians and scientists with only rare exceptions (e.g. Leventhal & Huynh 1996a,b; Harris 1997). One reason for the later neglect might be that the early part of Kaiser's paper contains some *Sturm und Drang*, exaggerated statements that are not supported by his careful analysis in the later part of the paper. Moreover, he did not present his major conclusion clearly.

Kaiser (1960) started by claiming that 'A correct interpretation of the traditional two-sided test would appear to make this classic procedure of essentially negligible interest'. And a bit later: 'It seems difficult to imagine a problem for which this traditional test could give results of interest'. This language will be frightening to anyone looking back on a career littered with two-tailed tests!

His central point, however, boiled down to this: when a classical two-tailed test is carried out, the probability statement or P value obtained refers only to the probability, assuming that H_0 is true, of obtaining a $|d|$ as great as or greater than the observed $|d|$ and not to any conclusion about the sign of d, the direction of the departure from the null hypothesis. This is correct. A statement such as 'There was a significant difference between the means for groups A and B ($P = 0.02$)' is allowable, but one such as 'There was a significantly higher mean for group A than for group B ($P = 0.02$)' is not. 'Allowable', of course, only under the illogical decision-theoretic framework (Eysenck 1960; Altman 1991, p. 168; Hurlbert & Lombardi 2009) that requires specification of α and dictates use of the 'significant/not significant' terminology.

Kaiser (1960), however, stated this conclusion in too general terms: 'We cannot logically make a directional statistical decision or statement[,] when the null hypothesis is rejected[,] on the basis of the direction of the difference in the observed sample means'. Scientists reasonably accept broader meaning for such language and will tend to respond: If P from a two-tailed test is sufficiently small and $m_A > m_B$, then the conclusion that $\mu_A > \mu_B$ *is* a logical statistical decision.

Bakan (1966) criticized Kaiser's paper and central point, saying it represented '*reductio ad absurdum*'. We do not agree. Bakan rhetorically asked, 'If Sample Mean A is greater than Sample Mean B, and there is reason to reject the null hypothesis, what other direction can it [δ] *reasonably* [our emphasis] be?' Kaiser's point, however, was that *possibly* $\mu_A < \mu_B$ even though $m_A > m_B$ and P is low.

To conclude that $\mu_A > \mu_B$ when, in fact, $\mu_A < \mu_B$, is to make what Kaiser termed the 'particularly repugnant', 'repulsive' and 'nasty error of the third kind', which he symbolized with γ. An increased probability of this gamma error is what we risk when we interpret the P value or probability statement from a conventional two-tailed test as applying not only to the existence of a difference but also to the direction of the difference.

Occasionally, it is claimed that the mere use of a two-tailed test implies the investigator has no interest in reaching 'directional' conclusions (e.g. Ferguson 1971; Braver 1975). It seems likely, however, that every investigator who has ever carried out a two-tailed test has been keenly interested in the direction of any departures from the null hypothesis. The question raised by Kaiser (1960) is whether we should calculate P in such a way that it refers to our directional conclusion. Kaiser recommends that we should. That

doi:10.1111/j.1442-9993.2009.01946.x

approach has in its favour the fact that it assigns a larger part of the full inferential process to objective procedure and quantification.

In contrast, we can acknowledge the massive precedent embodied in the existing literature and continue to use the classical approach. This requires simply acknowledging that P values from the two-tailed tests do not formally apply to the directional interpretations. It in no way precludes our reaching 'logical' conclusions concerning the directions of departures from null hypotheses.

We suggest the latter approach, unwilling to fight history. But as can be shown, the difference between the two approaches is both small and simple.

In practical terms, Kaiser's (1960) recommendation consists of what we termed Option 3 in the preceding section: the carrying out of the two one-tailed tests, or at least the setting up of this as the formal decision procedure. If for each test we set the critical $\alpha = \alpha_i$, then the general procedure is exactly equivalent to a classical two-tailed test with overall $\alpha = 2\alpha_i$, even though operationally only one of the one-tailed tests would be carried out, that where H_1 corresponds to the observed direction of the result. The P value yielded by such a test can simply be doubled ($2P$) if we wish our probability statement to refer to 'direction' (Gibbons & Pratt 1975; Harris 1997). This doubling will not be a valid procedure, however, if the probability distribution under H_0 for different possible outcomes is not symmetric, as will often be the case, for example, with categorical data in asymmetrical contingency tables.

Thus we gain the additional protection against the 'nasty' gamma error simply by doubling our risk of a type I error. This doubling also confers additional protection against type II errors, the probability of which is symbolized as β.

Demonstration that the difference between the two approaches is small requires consideration of the magnitudes of these three types of error (α, β, γ). As indicated by Kaiser (1960) and clarified by Peizer (1967), Shaffer (1972) and Leventhal and Huynh (1996a,b), γ has a maximum value of $\alpha/2$, which is attained asymptotically as δ approaches zero, and decreases rapidly as δ becomes larger (γ is undefined, of course, when $\delta = 0$). Thus γ will always be small, assuming that a low α is specified; and if we follow Kaiser's recommendation (= our Option 3), γ will be exactly half what it will be with the classical two-tailed test. Equivalently, γ is equal to the difference in power $(1 - \beta)$ between the classical two-tailed test and Option 3. As δ increases, this difference in power approaches zero. As δ decreases, the power of the classical two-tailed test approaches α. If δ is small should we care, for example, whether $1 - \beta$ equals 0.05 instead of 0.025? Or whether γ equals 0.025 instead of 0.0125?

Kaiser (1960) thus brought an interesting matter to our attention, one that requires us to be careful in describing the results of two-tailed tests. Abandonment of the classical two-tailed test seems not required, however. If we insist on having P values that formally apply to directional conclusions, in most situations we can simply double the P values yielded by the two-tailed test. Otherwise we can continue to reach directional conclusions informally on the basis of the undoubled P values, agreeing with Schulman (1992) that the 'extra risk [of a gamma error] is so small it can safely be ignored'.

Lest the reader think this is a long dead issue, we cite five modern restatements and elaborations of Kaiser's thesis. Hand and McCarter (1985) have stated that 'ordinary two-tailed test[s] of significance are completely useless insofar as practical actions are concerned' and provide 'no basis' for deciding on the sign of δ even when the sign of d is known and P is low. They recommend Option 3 which they term the 'directional two-tailed test'. Casella and Berger (1987) stated that 'The testing of a point null hypothesis [e.g. $H_0: \delta = 0$] is one of the most misused statistical procedures . . . Few experimenters of whom we are aware, want to conclude that "there is a difference". Rather, they are looking to conclude that "the new treatment is better". Thus there is a direction of interest in many experiments and saddling an experimenter with a two-sided test would not be appropriate'. In reference to medical research, Peace (1989) claimed 'that if the [clinical] trial is designed to be confirmatory, then the alternative [hypothesis] *cannot* be two-sided and still be logical'.

Leventhal and Huynh (1996a) recommended and thoroughly reviewed the 'directional two-tailed test'. They reiterated the claim that conventional two-tailed tests 'do not provide for directional decisions'. They emphasized the importance of considering γ while simultaneously admitting that it will be trivially small except where α is large and δ is small. They gave an excellent summary of the various options for calculating power, confidence intervals and minimum sample sizes and how these relate to the 'tailedness' of a test. But no new arguments were advanced for the 'directional two-tailed test'. Moreover, their framework is the rigid decision theoretic one where α must be specified, high P values lead to 'acceptance' of H_0, and the primary objective is to make an up-or-down decision, not to evaluate the strength of evidence against H_0 or to estimate the precision of d.

In sum, those authors favouring 'directional two-tailed tests' seem opposed to use of classical one-tailed tests. They are two-tailers. They differ from those who suggest the standard two-tailed tests only in the secondary matter of whether a reported P value should be that for the tail corresponding to the observed direction of d or, as is conventional, that for both tails. With either approach to reporting the P value, logical, directional decisions are possible.

doi:10.1111/j.1442-9993.2009.01946.x

Finally, in an extended review of the subject, Harris (1997) initially seems to be taking a hard line, stating '. . . as Kaiser (1960) (first?) pointed out, a researcher who adheres to the two valued logic of the traditional two-tailed tests can never come to any conclusion about the direction of the effect tested . . .' It soon becomes apparent, however, that Harris's complaint is not with two-tailed tests but rather with the pervasive misinterpretation of them.

What Harris terms 'three-valued logic' or 'three-alternative null hypothesis significance testing' is the logic that has always been inherent, albeit often ignored, in two-tailed tests, standard or 'directional' (Tukey 1991; Abelson 1995; Harris 1997; Tryon 2001). The resultant P value and observed sample means allow one of three conclusions: (i) the difference between the true or population means is probably positive; (ii) the difference is probably negative; or (iii) we don't have adequate information to make a confident statement about the sign of the difference – assuming one exists, as is almost always the case.

On the principled assumption that H_0 is false, our confidence that the sign of the real difference between two groups has the same sign as the observed difference between sample means can be measured as $(1 - P/2)$ (Jeffreys 1939; Oakes 1986). This could also be expressed as an odds ratio, that is, $(1 - P/2)/(P/2)$. So if a statistical test for a difference between two groups yields $P = 0.20$, and $(m_A - m_B) > 0$, then we can conclude that the odds are 9–1 that $\mu_A > \mu_B$. This should give us a sharper sense of the weight that P values greater than the conventional 0.05 can have.

THE COLLECTIVE INTEREST CRITERION

For fixed overall α, a one-tailed test gives us slightly greater power of detecting a difference in the direction tested and zero power for detecting a difference in the opposite direction. When the latter occurs, the preceding review shows that there is only one valid and honest way of avoiding the waste of information and resources that can result from this lack of power: Use one-tailed tests only in situations where a strong difference in the untested direction would truly be of no interest. In such cases, nothing is lost by use of a one-tailed test and a little additional power is gained. Such cases rarely occur in either basic or applied research.

This criterion was implicit in the writings of Neyman (1937) on one-tailed tests and one-sided estimation, for example, where he referred to 'frequent practical cases [where] we are *interested only* [our emphasis] in one limit'. Neither he nor many others who cursorily referred to this idea of 'interested only' over the next decades attempted explication of the criterion or a translation of it into operational terms.

Finally, Kimmel (1957) and then Welkowitz *et al.* (1971) and Pillemer (1991), in essays on the general inappropriateness of one-tailed tests, defined the matter very clearly. Kimmel stated:

> This limitation [zero power of 1-tailed test for an effect in the direction opposite that tested] cannot be shrugged off by the comment, 'We have no interest in a difference in the opposite direction'. Scientists are interested in empirical fact regardless of its relationship to their preconceptions . . . Use a one-tailed test when results in the unpredicted direction will, under no conditions, be used to determine a course of behavior different in any way from that determined by no difference at all.

This suggests the behaviour of the investigator would be the operational criterion for assessing whether she is truly 'interested only' in the one direction tested for. This behaviour would include actions taken relative to the dissemination of results and to the manner in which they are described and interpreted.

That we cannot know with certainty the behaviour, past, present, or future, of individual investigators is no obstacle to a general prescription. Our extensive observation of investigators on several continents shows them all to be *gente reviva*, folks quick to appreciate opportunity. '[A]lthough the experimenter might be anticipating, or hoping for, an outcome in one tail of the distribution, he (or she) will surely not disregard an extreme result in the opposite tail of the distribution . . .' (Upton 1992). In every d associated with a low P value, regardless of sign, there is a good story. And we have never known a colleague who shirked at its telling. Such behaviour is predictable. Editors, reviewers and statistical consultants should operate on the assumption that it is universal. They will have to be prepared, of course, to argue with those (e.g. Royall 1997) who will tell them that it is 'presumptuous and condescending' to assume that researchers will take note of unexpected results and change their behaviour in response to them.

More specifically, we recommend an interpretation of Kimmel's criterion that makes the use of one-tailed tests independent of the idiosyncratically defined and *post hoc* reporting of the interests or expectations of individual investigators. 'Interest' should be defined only institutionally, that is on the basis of the collective interest of science and society. This need not be carried out formally by any particular body. It will be enough that a consensus is reached that in every area of research, there will always be a collective interest in knowing of results that are the opposite of those predicted by or of interest to the original individual investigators. Exceptions will be so few and so distinctly *sui generis* that we can ignore them in our general prescription that one-tailed tests be avoided. This 'collective interest' interpretation of Kimmel's

doi:10.1111/j.1442-9993.2009.01946.x

(1957) 'interest only' criterion is consistent with his own arguments opposing the use of one-tailed tests.

The best explications of this 'collective interest' criterion we have found are by Welkowitz *et al.* (1971; Appendix S1) and Pillemer (1991).

Welkowitz *et al.* (1971) provided the first cogent textbook treatment of one-tailed tests that we know of. They stated:

> It is our belief, however, that one-tailed tests should *not* be used in place of two-tailed tests of significance. One reason for this position is that the user of one-tailed tests is placed in an embarrassing position if extreme results are obtained in the direction opposite to the one expected. . . . [I]n almost all situations in the behavioural sciences, extreme results in either direction are of interest. Even if results in one direction are inconsistent with a given theory, they may suggest new lines of thought. Also, theories change over time, but published results remain immutable on the printed page. Those perusing the professional journals (perhaps years after the article was written) should not be bound by what theory the original researcher happened to believe . . . [E]xtreme results in the 'opposite' direction invariably have important implications for the state of scientific knowledge. . . . If a one-tailed test is used and results are obtained in the 'opposite' direction which would have been significant, the experiment should be repeated before conclusions are drawn.

We do not believe that it would ever be desirable to repeat a well-conducted experiment for the *sole* reason that the result did not coincide with the investigator's expectation. But otherwise the advice of Welkowitz *et al.* (1971) on this subject is clearer and wiser than that offered by any other textbook we know.

That same advice was retained, with only minor changes in wording, up through the fourth edition of their text (Welkowitz *et al.* 1991). In the fifth edition (Welkowitz *et al.* 1999), however, they completely rewrote their section on 'One-tailed tests of significance' and withdrew from the field of battle. They omitted all the advice quoted above, recast their discussion in terms of confidence intervals and said nothing about the appropriateness of one- *versus* two-tailed tests. It is difficult to imagine what could have precipitated such a complete retreat from their earlier cogently argued position. Doubtless it was one or more of the attacks of the last 30 years on significance testing by the many critics who have failed to distinguish between the useful properties of significance tests on the one hand and the misuse of such tests by researchers and statisticians on the other.

In December 1999, following some earlier correspondence, we sent the above text to Joan Welkowitz asking for her appraisal. Though she did not respond, in the next (sixth) edition the authors (Welkowitz *et al.* 2006) reinserted a warning against use of one-tailed tests. Less clear and forceful than the original, it now reads, 'Although behavioral researchers often have a theory that predicts the direction of their results, to be conservative . . . they usually report their results in terms of two-tailed tests. . . . Because the research community does not want to discourage the reporting of paradoxical findings, . . . the two-tailed test is the norm. However, one-tailed tests may be justified if the results would make no sense in the opposite direction . . .'.

In his excellent review of the inappropriateness of one-tailed tests in educational research, Pillemer (1991) states:

> The criteria for choosing a hypothesis testing strategy should be refocused away from the concerns of individual scientists and toward the needs [read: *interests*] of the broader scientific community, present and future . . . . Educational research should continue to be guided by hypotheses, but a researcher's personal beliefs and predictions should not dictate what outcomes will become a part of the public record. . . . Researchers should routinely present . . . exact two-tailed probabilities for all major statistical comparisons within a study.

Thus we conclude: one-tailed tests are generally inappropriate. We save for a section at the end of this article, discussion of certain types of applied research where the one-tailed test can be a useful tool.

USAGE IN TWO JOURNALS

Method of survey

We now report the results of our survey on the frequency of the use of one-tailed tests in the 1989 and 2005 volumes of the journals AB and OE. We selected the former journal because of our observation that misuse of one-tailed tests was widespread in the animal behaviour literature in general. We selected the latter for the purpose of making a comparison of behaviour with ecology, where our impression, based partly on Hurlbert and White (1993), was that such misuse was much less common. An earlier version of this manuscript was based on only the 1989 volumes. Here we have analysed an additional 585 articles in the AB and OE volumes for 2005, allowing us to document some marked changes in statistical practice over this 16-year interval.

For every one of the 1169 articles in these four volumes, we examined the 'tailedness' of procedure for all comparisons of two samples or treatments, all comparisons of a sample with a standard or hypothetical

doi:10.1111/j.1442-9993.2009.01946.x

value, and all one- and two-sample tests of correlation and regression coefficients.

Many papers used more than one type of statistical test and often used a given type for two or more data sets. We examined every individual test reported and calculated frequency of usage of one-tailed tests on both per article and per procedure bases. Consider three articles, one reporting six two-tailed *t*-tests, a second reporting three two-tailed *t*-tests and one one-tailed *t*-test and a third reporting one one-tailed and one two-tailed Fisher's exact test. For these, the per article frequency of usage of one-tailed *t*-tests would be 67%, and the per procedure frequency of *one-tailed* tests would be 50% for the *t*-test and 100% for Fisher's exact test.

The results of the survey are given in Table 1 with separate tabulations, by journal and year, for the 12 most frequently used statistical procedures where one-tailed testing was possible. Table 2 summarizes some of that information, grouping procedures into two categories, parametric and non-parametric, in order to illuminate some of the major contrasts of interest.

Of the 248 and 301 articles published in AB in 1989 and 2005, 193 and 243 articles respectively, carried out statistical tests on at least one data set for which one-tailed tests were possible. Of the 336 and 284 articles in OE in 1989 and 2005, 187 and 160 articles, respectively, carried out at least one statistical test on data for which a one-tailed test was possible.

For assessing difference between journals or years we have occasionally used, in the following sections, χ^2-tests to compare two percentages. These were all two-tailed tests, involved 1 degree of freedom, and only their resultant *P* values are reported. Fastidious minds might object to our rustic statistical approach on grounds that we have, after all, censused four volumes, not sampled them, and that we have engaged in extensive 'pooling' in reducing data sets to 2×2 contingency tables. They are welcome to try alternative approaches using the the raw data in our Table 1.

Failure to report 'tailedness'

We first note the dismal, though improving, reporting rate for tailedness. Very often, and in OE more than in AB, the authors did not indicate, directly or indirectly, for *any* of the tests used whether they were one- or two-tailed. For 1989 and 2005 the respective percentages of articles lacking this information were were 34 and 8 for AB and 65 and 44 for OE. Perhaps animal behaviourists are more attentive to this detail because they are more likely than ecologists to regard one-tailed tests as a valid option.

Rates calculated on a per procedure basis, using the information given under 'All uses' in Table 1, rather than that under 'Any test', also present a grim but improving picture. For 1989 and 2005 respectively, tailedness was not determinable for 44% and 29% of the procedures reported in AB, or for 74% and 57% of procedures reported in OE. All χ^2-tests for both journals, whether based on numbers of articles or number of procedures, gave strong evidence the improvement between 1989 and 2005 was real ($P < 0.0001$).

We suspect that many of the non-reporting authors share our point of view, routinely use only two-tailed tests, and therefore consider specification of that fact superfluous. Nevertheless, their silence does somewhat constrain our analysis, not to mention precise interpretation of their own results. In many cases, however, supplementary information was provided – such as critical *t*-values for a specified α – that allowed us to determine 'tailedness' even though the author was not explicit on the point.

Overall frequency of use of one-tailed tests

We first consider the overall frequency of use for the two journals taken together. Of the 193 and 312 articles reporting, for 1989 and 2005 respectively, at least one test where 'tailedness' could be determined, 41% and 21% used one-tailed tests at least once (Table 1). On a per procedure basis the difference between use rates is similar, 36% *versus* 17%, based on sample sizes of 325 and 427 for the 2 years (Table 2, bottom). The decline in use seems real by both measures ($P < 0.001$).

What if we make the generous assumption that, of the 187 and 91 articles giving in each year no information on the 'tailedness' of any of their tests, not one used a one-tailed test? The usage rates for 1989 and 2005 then drop to 21% and 16%. If we make the even less reasonable assumption that all of the 187 articles silent on tailedness *did* use one-tailed tests, the usage rates jumps to 70% and 39%. Any figure in this range (16–70%) is discouragingly high if these tests are as inappropriate as we believe them to be.

All the reports were of basic research, though especially in the case of OE, the research often was intended to contribute to solution of practical problems. In no case of one-tailed test usage could the investigator have been expected to ignore a result strongly in the direction not tested for. But, no such results seem to have occurred! Only in a few instances did authors give a rationale for use of one-tailed tests. In both 1989 and 2005, there were exactly 13 articles in which the justification given was that a prediction had been made. In one other case each year the reason given was that a difference in one of the possible directions would have been either uninterpretable or of no interest. Users of one-tailed tests most often select such tests on the grounds that they have predicted *a priori* the direction of a result. With a prediction

doi:10.1111/j.1442-9993.2009.01946.x

Table 1. Use of one-tailed tests in the 1989 and 2005 volumes of *Animal Behaviour* and *Oecologia*

	Animal behaviour				*Oecologia*			
Statistical procedure	A No. of articles using test	B No. of articles where 'tail' is clear	C No. of articles using 1-tailed test	D (C/B) × 100	E No. of articles using test	F No. of articles where 'tail' is clear	G No. of articles using 1-tailed test	H (G/F) × 100
1989		(n = 248 articles)				(n = 336 articles)		
Correlation and regression methods								
r (Pearson)	46	18	6	**33**	61	14	4	**29**
r_s (Spearman)	36	16	9	**56**	24	13	8	**62**
τ (Kendall)	4	1	0	**0**	3	0	–	–
Linear regression	19	2	0	**0**	63	2	2	**100**
Methods for continuous variables								
t-tests[ab]	71	38	10	**26**	82	23	9	**39**
Wilcoxon[a]	116	65	28	**43**	48	14	7	**50**
Z-test	7	5	2	**40**	2	0	–	–
Sign test	18	11	3	**27**	3	1	1	**100**
Methods for categorical variables								
Binomial	24	13	4	**31**	4	2	2	**100**
χ^2-test	60	32	4	**13**	36	14	4	**29**
Fisher exact	30	11	5	**45**	9	1	1	**100**
G-test	22	13	5	**38**	9	6	2	**33**
Any test	193	128	50	**39**	187	65	29	**45**
All uses[c]	453	225	76	**34**	344	90	40	**44**
2005		(n = 301 articles)				(n = 284 articles)		
Correlation and regression methods								
r (Pearson)	40	22	9	**41**	44	12	0	**0**
r_s (Spearman)	49	39	11	**28**	15	5	1	**20**
τ (Kendall)	3	3	2	**67**	2	0	–	–
Linear regression	17	17	0	**0**	66	29	0	**0**
Methods for continuous variables								
t-tests[†‡]	125	104	16	**15**	71	36	9	**25**
Wilcoxon[†]	99	64	11	**17**	28	10	1	**10**
Z-test	2	2	1	**50**	1	0	–	–
Sign test	8	2	1	**50**	0	0	–	–
Methods for categorical variables								
Binomial	27	11	6	**55**	1	0	–	–
χ^2-test	41	33	2	**6**	26	19	2	**11**
Fisher exact	15	5	0	**0**	6	1	0	**0**
G-test	11	10	0	**0**	7	3	1	**33**
Any test	243	222	52	**23**	160	90	14	**16**
All uses[§]	437	312	59	**19**	267	115	14	**12**

[†]Include tests for both paired and unpaired data. [‡]Or equivalent, as when ANOVA is used to compare two treatments. [§]Data in this row are on a per procedure basis. Thus if an AB article used the t-test, linear regression and Fisher's exact test each one or more times, it would contribute 3 to column A and 0–3 to column B and to column C. Key results shown in bold.

doi:10.1111/j.1442-9993.2009.01946.x

Table 2. Summary of changes in frequency of use of one-tailed tests in the1989 and 2005 volumes of *Animal Behaviour* and *Oecologia*, by general category of procedure

Procedure type, journal, year & $n^{\dagger}$	% of diagnosable uses that were one-tailed	% reduction in frequency of use (P-value, fr χ^2-test)
Parametric procedures (r, regression, t, Z, binomial)		
Animal Behaviour		
1989 ($n = 76$)	29	
2005 ($n = 156$)	21	28 (0.21)
Oecologia		
1989 ($n = 51$)	33	
2005 ($n = 77$)	12	64 (0.006)
Non-parametric procedures (r_s, tau, Wilcoxon, sign, χ^2, Fisher's exact, G)		
Animal Behaviour		
1989 ($n = 149$)	36	
2005 ($n = 156$)	17	53 (0.0002)
Oecologia		
1989 ($n = 49$)	47	
2005 ($n = 38$)	13	72 (0.002)
All procedures, both journals		
1989 ($n = 325$)	36	
2005 ($n = 427$)	17	53 (<0.0001)

†Number of articles with at least one diagnosable use of a given procedure, as given in columns B and F in Table 1. An article with diagnosable uses of three different procedures would contribute an n of 3 to the base from which frequencies in the first column were calculated.

success rate of apparently 100%, these AB and OE authors must be clairvoyant, be working in areas where theory is very 'mature', or be asking questions the answers to which are already known.

As one of us is not entirely without sin, having occasionally during a wild and reckless youth used one-tailed tests for predicted results (Lombardi & Curio 1985a,b), we of course cannot be too indignant.

Behaviour *versus* ecology

Results reported in the rows of Table 1 labeled 'Any test' suggest that, contrary to our initial idea, differences between ecologists and behaviourists in frequency of use of one-tailed tests are slight or non-existent. In 1989, frequencies for OE and AB were 45% and 39%, respectively ($P = 0.56$), whereas in 2005 they were 16% and 23% ($P = 0.66$). However, if we make the not too unrealistic assumption that only two-tailed tests were used in those papers silent on 'tailedness', usage is greater by behaviourists. Calculated as column C/column A and column G/column E, the percentage of papers using one-tailed tests in 1989 was 26% for AB and 16% for OE ($P = 0.018$), and in 2005 was 21% for AB and 9% for OE ($P = 0.0013$). These numbers suggest that, to the extent that these journals are representative of their fields, behaviourists might be using one-tailed tests on the order of 43–62% more often than are ecologists.

Similar trends are documented when we calculate rates on a per procedure basis (rows of Table 1 labeled 'All uses'). No consistent difference is found between AB and OE in rates of one-tailed test use based solely on diagnosable uses – 34% *versus* 44% in 1989 ($P = 0.076$), and 19% *versus* 12% in 2005 ($P = 0.10$). But if we make the not unrealistic assumption that only two-tailed tests were used in those procedures where 'tailedness' was not indicated, then AB authors seem to be consistently higher users. Again calculating them as column C/column A and column G/column E, one-tailed use rates in 1989 become 17% and 12% ($P = 0.041$) for AB and OE respectively, and in 2005 become 14% and 5% ($P = 0.0005$) for AB and OE, respectively. These might suggest that behaviourists are using one-tailed tests 45–160% more frequently than are ecologists.

We note that Hurlbert and White (1993) found that only four of 95 experimental papers by zooplankton ecologists made use of one-tailed tests. They did not record, however, the number of these 95 papers where the 'tailedness' of tests was left unspecified.

In both years there were some striking differences between AB and OE in the frequency of usage of particular types of tests. Regression methods were used more often in OE. Methods for categorical data were used more often in AB. We believe such differences simply reflect substantive differences between animal behaviour and ecology in the types of studies conducted and response variables measured.

doi:10.1111/j.1442-9993.2009.01946.x

Table 3. Statistics books most frequently cited by articles in *Animal Behaviour* and *Oecologia* in 1989 and 2005

	Number of articles citing book			
	1989		2005	
Book	*Animal Behaviour* (n = 248)	*Oecologia* (n = 336)	*Animal Behaviour* (n = 301)	*Oecologia* (n = 284)
Sokal and Rohlf (1969, 1981, 1995)	37	50	25	20
Siegel (1956), Siegel and Castellan (1988)	23	4	13	1
Zar (1974, 1984, 1999)	6	13	17	15
Snedecor (1956), Snedecor and Cochran (1967, 1980)	6	8	0	0
Conover (1980)	6	2	1	1
Winer *et al.* (1962, 1971)	2	6	0	0
Hollander and Wolfe (1973)	4	1	0	0
Steel and Torrie (1960, 1980), Steel *et al.* (1997)	1	4	0	1
Martin and Bateson (1986, 1993)	3	0	11	0
Cohen (1977, 1988)	0	0	9	1
Scheiner and Gurevitch (1993)	–	–	1	6
Crawley (1993)	–	–	4	1
Underwood (1997)	–	–	1	4

Only books listed are those which had at least four entries in at least one cell of this table.

More puzzling is the observation that, for the same general type of analysis, AB articles were much more likely to use non-parametric procedures than were OE articles (Table 1). Over both years, AB authors used Wilcoxon tests as frequently as *t*-tests, whereas OE authors used *t*-tests about twice as often as Wilcoxon tests. AB authors used *r* about as frequently as r_s and τ, but OE authors showed a distinct preference for *r*. These differences between AB and OE authors in frequency of usage of specific types of tests seem real and not reasonably attributed to sampling error (all tests, $P < 0.05$).

Parametric *versus* nonparametric tests

In 1989, there was a striking tendency in both journals for the per article frequency of use of one-tailed tests to be greater for nonparametric methods than for parametric ones (Table 1, columns D and H). When data for the two journals are combined, we find one-tailed tests being used 31% of the time with *r* and 57% of the time with r_s and τ. We find one-tailed tests being used 31% of the time with *t*-tests and 44% of the time with Wilcoxon tests. Calculated over only these two categories of tests, the frequencies of one-tailed usage are 31% for the parametric procedures and 48% for the nonparametric procedures ($P = 0.018$).

This pattern was not observed in 2005 (Table 1, columns D and H). Combining data for both journals, one-tailed tests were used 26% of the time with *r* and 30% with r_s and τ ($P = 0.74$). Also we find one-tailed tests being used 18% of the time with *t*-tests and 16% of the time with Wilcoxon tests ($P = 0.76$). The difference between the overall frequencies of one-tailed usage for these parametric procedures (20%) and the corresponding nonparametric ones (21%) now seems to disappear.

A similar contrast between 1989 and 2005 is seen when rates on a per procedure basis are examined, using data in Table 2. When data for the two journals are combined, one-tailed tests were used 39% of the time with non-parametric procedures and 31% of the time with parametric ones ($P = 0.13$) in 1989 whereas the corresponding rates were 16% and 18% ($P = 0.31$) in 2005.

A *rational* explanation of why one-tailed tests would be considered appropriate perhaps 27–55% more often with nonparametric procedures than with parametric ones in 1989, but were considered appropriate at about the same rate in 2005 eludes us. If almost all use of one-tailed tests in basic research is irrational; however, perhaps it is not surprising to find, in the patterns of that use, further evidence of irrationality. The historical roots of this problem are identified in the bibliographies of the AB and OE authors, as we now discuss.

Proximate sources of error

Most of the AB and OE articles list no statistics book in their bibliographies. Those that do, cite three books – Siegel (1956; Siegel & Castellan 1988), Sokal and Rohlf (1969, 1981, 1995) and Zar (1974, 1984, 1999) – far more than they do any others (Table 3). The last two of these are among the statistics books most widely used by biologists generally. The misprescriptions in all these books likely are a major cause not only of the misuse of one-tailed tests but also of the particular

doi:10.1111/j.1442-9993.2009.01946.x

Table 4. Comparison of three statistics books with regard to erroneous recommendations and statements concerning one-tailed tests using procedures for categorical data

A		B	C	D	E
Reference & test	Pages	Tailedness of example	Contrary result in 1-tailed example likely to be of interest?	Authors imply region of rejection should be 2-tailed for 2-tailed test?	Total number of errors
Siegel (1956)					
Binomial (1×2 tables)	36–40	1-tailed	yes	yes†	
Fisher's exact	96–104	1-tailed	yes	yes†	
χ^2 (2×2 tables)	107–109	1-tailed	yes	yes	8
Median (χ^2)	111–116	1-tailed	yes	yes	
McNemar (χ^2)	63–67	1-tailed	yes	yes	
Siegel & Castellan (1988)					
Binomial (1×2 tables)	39–42	1-tailed	yes	yes‡	
Fisher's exact	103–111	1-& 2-tailed	yes	yes†	
χ^2 (2×2 tables)	116–117	2-tailed	NA	no	4
Median (χ^2)	124–128	1-tailed	yes	yes	
McNemar (χ^2)	75–80	2-tailed	NA	no?	
Sokal and Rohlf (1981)					
Binomial (1×2 tables)	77–79	1-tailed	yes‡	yes†	
G (2×2 tables)	732–738	2-tailed	NA	no	
χ^2 (1×2 tables)	700–701	2-tailed	NA	no	1‡
χ^2 (2×2 tables)	743	2-tailed	NA	no	
Fisher's exact	738–743	2-tailed	NA	yes†	
McNemar (*G*)	769–770	2-tailed	NA	no	

Except where otherwise noted, each 'yes' in columns C and D of this table signifies an error. †The implication is correct in this case. ‡Same error present in 1995 edition. NA, not applicable.

patterns of use we have documented in Tables 1 and 2. If the blame falls on few shoulders, this is not to say the poor, bewildered behaviourists and ecologists would easily have found better advice elsewhere. Of 52 statistics books examined by us, 40 give vague, inconsistent, or simply bad advice as to when one-tailed tests are permissible (Appendix S1). The primary literature on the topic is equally confused.

With respect to one-tailed tests, the flaws of these three books cited above are easily summarized. All suggest the use of one-tailed tests when the investigator predicts or thinks they know in what direction a result will lie (Table 4; Appendix S1). All except Siegel (1956) additionally accept their use when there is 'interest only' in a result in one particular direction. They thus ignore the critical conflict between these two criteria: prediction of the direction of a result in no way implies lack of interest, collective or individual, in a result in the contrary direction. All apply one-tailed tests to many data sets where results in the unpredicted direction would have been of definite interest to science (e.g. Table 4, columns B and C). Zar pushes one-tailed tests especially hard. In his 2004 edition, though he has no main entry in his index for one-tailed tests, he in fact discusses and gives examples of them in at least 23 different places. Few other modern books so effectively confer unmerited legitimacy on use of one-tailed tests in biology.

None of these three books discusses what should or could be carried out when, after deciding to use a one-tailed test, a result in the direction opposite that expected is obtained. Should the investigator select Option 1, 2, 3, or 4?

The widespread reliance on these books seems to explain rather fully the widespread misuse of one-tailed tests, especially by animal behaviourists for whom Siegel (1956) has long been a favourite reference (Table 3; Martin & Bateson 1986, 1993). And the fact that Siegel (1956) was the first compendium of non-parametric methods to be published and has been the most widely used one for a few decades, clearly can account for the more frequent use of one-tailed tests with non-parametric than with parametric procedures. There is no statistics book that suggests the use of one-tailed parametric procedures as insistently as Siegel (1956) suggests the use of one-tailed non-parametric procedures, although Zar (1999, 2004) comes close, presenting approximately 15 examples using one-tailed parametric procedures.

1989 *versus* 2005: on the road to recovery?

The labor of having added data for 2005 to this article has been rewarded, finally, by some good news! Use of one-tailed tests decreased between 1989 and 2005 by

doi:10.1111/j.1442-9993.2009.01946.x

something on the order of 50%. In Table 1 this can be seen by comparing the 1989 rows labeled 'Any test' or 'All uses' with those for 2005. In Table 2, it is shown that the decrease was found in both AB and OE for both parametric and non-parametric procedures. This large decrease might help explain why certain contrasts (e.g. AB *vs*. OE, parametric *vs*. non-parametric procedures) that were clear in the 1989 data were different or less clear in the 2005 data.

There was also some improvement in the frequency with which authors made clear whether they were using a one- or two-tailed procedure. See the column B/column A and column F/column E ratios in Table 1. As we stated at the beginning of this article, accurate assessment of frequency of use of one-tailed tests is hindered when 'tailedness' for any fraction of uses is unclear. Thus some of the apparent decrease in use of one-tailed tests could be an artefact resulting from two-tailed procedures being more frequently identified as such in 2005 than in 1989. There is no way to know.

Most AB and OE authors continued to cite no statistics text in 2005, just as in 1989. There were some interesting shifts in which texts were cited in the 2 years, including a large decrease in citations of Siegel (1956); Siegel and Castellan (1988), despite the availability of the second edition. But these shifts represented a shifting among unreliable texts on the issue, not a shift from use of unreliable to reliable ones.

Some of the 1989–2005 reduction in use of one-tailed tests might have been the result of pre-publication circulation of this article. Different versions of it have been in circulation since 1994. Numerous editors, editorial board members, statisticians and other scientists have received copies or reviewed it in official or unofficial capacities, many offering helpful suggestions, several offering praise and a few offering signs of extreme consternation! These include: Patrick Bateson, Emili García-Berthou, C. Ray Chandler, Peter Chapman, Boyd Collier, Eberhard Curio, Richard B. Darlington, Juan D. Delius, Fred C. Dyer, Joseph Fleiss, Janneke Hoekstra, Douglas H. Johnson, Sam K. Kachigan, Manfred Milinski, Karl S. Peace, Thorsten Reusch, Michael Riggs, F. James Rohlf, Jennifer L. Shaw, Stephen M. Smith, Robert R. Sokal, Clayton L. Stunkard, Sandra L. Vehrenkamp, Meredith J. West and several anonymous reviewers. It is also possible, of course, that even though they were not citing them in their articles, authors during the 1989–2005 time period were increasingly influenced by those few texts giving good advice on this issue (e.g. Fleiss 1986; Altman 1991; Welkowitz *et al.* 1991; Schulman 1992; Bart *et al.* 1998; see below).

RELIABILITY OF MODERN TEXTS

Change in texts must come more slowly. It is evident that the overall quality of advice on one-tailed testing remains exceedingly poor in both older and newer statistics books. This is documented in Appendix S1 which presents verbatim the advice given by 52 statistics books as to when one-tailed tests should or might be used. In that table, these books are categorized into *sombreros blancos* (white hats, $n = 12$) which give reasonable advice, *sombreros negros* (black hats, $n = 15$) which give bad advice and *sombreros grises* (gray hats, $n = 25$) which give vague, inconsistent or mixed advice.

Advice aimed at behavioural scientists is particularly unhelpful. We already have discussed the first three books in Table 3. The widely used and, in many ways, excellent primer on methodology for behaviour studies by Martin and Bateson (1986, 1993) was a strong force for inappropriate use of one-tailed tests. It praised Siegel (1956) as a 'bible' for behaviourists, and championed prediction of result as justification for one-tailed testing. It also recommended five other advanced statistics books, all of which either accept prediction of result as justification for one-tailed testing (e.g. Snedecor & Cochran 1980; Sokal & Rohlf 1981; Zar 1984) or are vague on the matter (Conover 1980; Meddis 1984).

In their second edition Martin and Bateson (1993) noted that 'The inappropriate use of one-tailed tests is one of the most common statistical malpractices in the behavioural literature'. This statement apparently referred, however, only to the *a posteriori* selection of one-tailed tests. They continued to recommend use of one-tailed tests when the direction of a result is predicted beforehand. In their third edition, Martin and Bateson (2007) no longer refer to Siegel (1956) as the 'bible'. But unfortunately they do refer to Siegel and Castellan (1988) as 'Probably still the best single text on non-parametric methods; clear, concise, reliable and indispensable', and still recommend one-tailed tests 'when knowledge or theory predicts the direction of the difference'.

Several other texts aimed in part at behavioural scientists (Ferguson 1971; Hays 1981; Kirk 1982; Glass & Hopkins 1984; Darlington & Carlson 1987) also give advice on this topic that conflicts with the collective interest criterion (Appendix S1). Moreover, this is true of the journal AB itself. Its 'Instructions to authors' (Animal Behaviour 2004) refer authors to Kimmel (1957) as an authority on the matter of one-tailed tests. They then negate that useful advice by also stating that one-tailed tests are justified anytime 'there are strong *a priori* reasons for predicting the direction of a difference . . .' – which, of course, is much of the time and a direct contradiction of Kimmel's advice.

Of the 12 books we have classified, with some liberality, as *sombreros blancos*, only Welkowitz *et al.* (1971, 1991) is bold enough to state that one-tailed tests should never be used *and* gives the full rationale for that position. Fleiss (1981, 1986), Schulman (1992)

doi:10.1111/j.1442-9993.2009.01946.x

and Bart *et al.* (1998) do the next best thing: they suggest the 'interest only' criterion and also discuss the rationale underlying it (Appendix S1). Another positive model for textbooks of the future would be that of Altman (1991) who states, 'One-sided tests are rarely appropriate. . . . Two-sided *P* values will be used throughout this book, and I recommend that they are used routinely'.

Most other books in the *sombrero blanco* category state the 'interest only' criterion but with the implication that only the 'interest' of the individual investigator need be considered. Some state the criterion clearly but then muddy the water with questionable examples. Helsel and Hirsch (1992), for example, suggest the 'interest only' criterion (Appendix S1), and then list several situations where they believe a one-tailed test would be appropriate. These situations, however, are ones where a result in either direction would be of definite interest, even if hopes, predictions, or expectations might lie in a single direction. For example, they say a one-tailed test would be appropriate to test whether a new sewage treatment plant reduced nutrient loads in plant effluent. Certainly every party concerned with such a plant would be hopeful that it would accomplish such a reduction; but they also would be highly interested if it increased nutrient loads. Those responsible for the design and construction of the plant might prefer only the optimistic one-tailed test, but the taxpayers and city council should insist on the two-tailed one!

Hawkins (2005) likewise recommends 'avoiding one-tailed tests like the plague' but then gives as an example of an 'exceptional reason' justifying a one-tailed test, a situation where 'you were testing a cancer drug and were absolutely sure the drug would not decrease survival . . .' (Appendix S1). But severe side effects of experimental cancer drugs are common, and surely if the facts in such a case turned out to contradict the *a priori* 'absolute' certainty, there would be a sudden shift in interest and focus.

APPROPRIATE USAGE

If, by the collective interest criterion, a one-tailed test is appropriate, then we recommend that by all means such a test be used. In contrast, a necessarily *post hoc* declaration in a manuscript that 'interest' was 'only' in a difference in the observed direction is not itself sufficient to establish 'appropriateness'. The general subject matter and nature of the study must themselves make it evident to a rational skeptic that a strong result in the unpredicted direction, if it had occurred, would not have resulted in a sudden change of 'interest' *and* that a revised testing procedure would not have been of interest to the scientific community or to society at large.

If the probability distributions of all test statistics were symmetric, like those for t or like those for symmetric contingency tables, and if exact P values always were reported, there would be no need for explicitly one-tailed tests. Their P values, when desired, could always be determined by dividing in two those yielded by two-tailed tests. These two conditions often do not obtain, however. So if a one-tailed test is justified, then the best general practice will be to report the P values specific to such a test.

Drug-testing

Are one-tailed tests generally appropriate for applied research, as many writers have implied? We believe not. That books aimed at such researchers make heavy use of one-tailed examples, we regard as unfortunate. For most applied research, as for virtually all basic research, a specific P value by itself is not the basis for taking or not taking some specific action. It is only used in a subjective fashion to help us decide what we might conclude about δ and the phenomena it bears on. We might be primarily interested in testing for the superiority of a new drug, educational technology, or manufacturing process. But if an experiment shows that this new candidate is actually inferior to the old standard, then that is not information we want to ignore. As Fleiss (1986, 1987) points out in connection with drug testing, such a result might suggest that something was wrong with the theory or logic that led up to the experiment and that a new question or phenomenon needs to be investigated. Publication of such an unexpected result can also serve as a warning or idea-generator for other scientists.

The domain of research where one-tailed tests are truly appropriate and useful consists of those testing situations where the P value or confidence interval obtained will directly influence an action to be taken. These usually will be situations where small differences, however real, will not be of practical significance. 'Smallness', of course, would be defined only in relation to each particular context. Thus we will be interested in testing null hypotheses of the form $\delta \geq c$ or $\delta \leq c$, where $c \neq 0$. As long ago suggested by Hodges and Lehmann (1954), 'we would test that . . . means do not differ by more than an amount specified to represent the smallest difference of practical interest'. If they are found to do so, with some specified level of confidence $(1 - \alpha)$, then one action is taken. If this condition is not met, another action is taken. In practice, the 'action' usually is not taken by the scientist(s) who carried out the research and statistical analysis. Rather it is taken by a government regulator, a vice-president for product development, or some other such decision-maker. Let us consider two types of situation.

doi:10.1111/j.1442-9993.2009.01946.x

When a new drug is developed and tested, it could be argued that the drug should not be approved or marketed unless it shows some minimum degree of efficacy or of superiority over a placebo or over an existing drug. Medically, the argument might be that this serves as a counterweight to negative side effects, both known and unknown; or it might be that public health in general is negatively affected by having many drugs of similar efficacy on the market. For a pharmaceutical company, if a new drug is only 10% more efficacious than its old one, it might not make economic sense to set up a new production operation and marketing programme for that drug; better, perhaps, to continue research on new compounds until one giving greater improvement in efficacy is found.

Let's say that a government agency decrees that new drugs for viral influenza must bring about, in clinical trials, a cure rate (μ_T) at least 20% greater than that (μ_C) in a placebo treatment. Two one-tailed testing procedures are available for this situation. In one, we set $H_0{:}\mu_T < \mu_C$ and require that $P < \alpha$ and that $d_{min} \geq 0.2$, where $d_{min} = (\mu_T - \mu_C)/\mu_C$, in order for approval to be given. In the other, we set $H_0{:}\mu_T < 1.2\mu_C$ and require only that $P < \alpha$ in order for approval to be given. The second approach is more conservative in that it makes approval practically impossible if the drug increases the cure rate by only 20% or slightly more. Using a much higher α in the second than in the first approach will cause some convergence in the results of the two approaches.

The two approaches are similar in that the burden of demonstrating sufficient efficacy is placed on the company. This is the conventional situation. If the company scientists do not use sufficient replication or do not adequately control extraneous variables, the null hypothesis will not be rejected even though $\mu_T >> 1.2\ \mu_C$.

The appropriateness of a one-tailed test derives from the fact that, with either approach, the decision maker's interest, though possibly not that of the scientists, does truly lie in only one direction. With the second approach it also derives from the fact that the 'opposite direction', that specified by the null hypothesis, includes not only $\mu_T = \mu_C$ and $\mu_T < \mu_C$ but also $1.2\ \mu_C > \mu_T > \mu_C$. Thus a two-tailed test of $H_0{:}\mu_T = 1.2\ \mu_C$ would not yield a P value pertinent to the unexpected observation that $m_T < m_C$, a result of possible interest at least to the scientists. If further exploration of such an unexpected observation is desired, a two-tailed test of $H_0{:}\mu_T = \mu_C$ can be applied to the same data and will yield a P value helpful to that exploration.

As indicated by complaints (e.g. Peace 1988) about the U.S. Federal Drug Administration's insistence on two-tailed tests, the above application of one-tailed tests in medical or pharmaceutical research is uncommon. Both researchers and regulators focus on the null hypothesis of no effect. As Overall (1991) notes, 'the issue of magnitude has been resolved in favour of statistical significance, because it is essentially impossible to achieve agreement concerning how small a treatment effect must be to be inconsequential'.

Noninferiority and equivalence trials

One tailed tests also prove useful in a relatively new area of statistical interest called 'noninferiority' and 'equivalence' trials. Good recent discussions of concepts and problems in this area include those of Pocock (2000), Barker *et al.* (2001), D'Agostino *et al.* (2003), FDA (2003), Welleck (2003), Tempelman (2004), Tamayo-Sarver *et al.* (2005) and Piaggio *et al.* (2006). Such trials have been of interest primarily to medical researchers, but can have applications in the environmental sciences and other fields as well (e.g. EPA 1989; McBride 1999; Manly 2003; Welleck 2003).

A new drug is proposed for a condition or disease for which there is already an approved and effective drug on the market. The manufacturer of the new drug can reasonably request that it be approved by the regulatory authorities if it can be showed that it is 'at least not inferior' to the older drug by more than some predetermined margin of error or margin of inferiority. This older drug might be produced by the same manufacturer or by a different one; it might also be simply a new formulation or method of delivery of the old drug. If effectiveness of the new drug or formulation and the old one are compared in a randomized trial, testing for non-inferiority could, in principle, be carried out via a one-tailed test of a null hypothesis of the form $H_0{:}\delta \geq c$ where c is positive and where δ is the margin of inferiority, the maximum allowable amount by which outcomes for patients under the new drug could be worse than those for patients under the old drug. Use of a placebo control group in such a trial usually would be unethical and not allowed, given that patients in the placebo treatment would be denied the proven benefits of the older drug.

However, at present the margin of inferiority is usually defined in terms not of the difference in patient outcomes between the new and old drug treatments but rather in terms of the difference between (or ratio of) the actual respective *effect sizes* of the new and old drugs relative to placebo controls. Non-inferiority is determined by a significance test or by whether a confidence interval about that difference or ratio extends below an arbitrarily set value. The actual effect sizes can only be estimated – with much uncertainty and trepidation – using patient response data from a true placebo or control group in an earlier study. That study naturally will have been one carried out in a different time and place and with at least a slightly different

doi:10.1111/j.1442-9993.2009.01946.x

spectrum of patients and protocols. As usual, the arithmetic is easy, but the conclusions reached are heavily determined by subjective decisions concerning the values of 'margin of inferiority' and α that are used in calculations and by how different response or patient outcome variables are weighted relative to each other.

Equivalence or bioequivalence trials involve considerations, decisions and complications similar to those of non-inferiority trials. Equivalence trials aim at demonstrating that a new drug or a new preparation or formula of an old one is medically or pharmaceutically neither superior nor inferior to an old one. Analysis is by way of two different one-tailed tests or a confidence interval about a ratio or difference between two effect sizes. Equivalence trials can be more useful in pharmacokinetic studies than in clinical ones. The FDA (2003) defines bioequivalence as, 'the absence of a significant difference in the rate and extent to which the active ingredient or active moiety in pharmaceutical equivalents or pharmaceutical alternatives becomes available at the site of drug action when administered at the same molar doses under similar conditions in an appropriately designed study'. It is important that the active ingredient reach the liver, for example, not only in adequate quantities but also not in excessive, possibly damaging quantities.

Given their complexities, it is not surprising the non-inferiority and equivalence trials are not yet common, are often inadequately described, and inspire reviewers to comments, such as 'It is not our intent to promote noninferiority or equivalence trials: the design should be appropriate to the question to be answered' (Piaggio *et al.* 2006).

Ecotoxicology

Toxicology is another field where ingenious use of one-tailed tests or one-sided confidence intervals has begun to be made by scientists and regulatory agencies. However, the statistical concern in testing the effects of, say, a pesticide, food additive, or pollutant is rather different than in the case of a new drug: we are now concerned that an effect *not* exceed some specified magnitude.

Two common testing situations are (i) where one wishes to estimate the effect of a specified dose or concentration, for example, of a pesticide; and (ii) where one wishes to estimate the dose that will produce an effect of some specified magnitude. The role that one-tailed testing can play in these contexts is a topic of a growing literature. Important papers include those of Crump (1984), Hoekstra and van Ewijk (1993a,b) and Stunkard (1994).

The first of the above types of testing situations might arise when it has been determined, for example, that a new insecticide, applied at χ kg ha^{-1} will achieve a satisfactorily high kill of mosquito larvae in standing water habitats. Before use of the insecticide at that rate can be officially approved, however, it is necessary to show that it will not cause unacceptably adverse effects on selected non-target elements of the biota or standard test species. Depending on the particular species and response variable (e.g. population size, reproductive rate, survival rate, etc.), it might be appropriate to specify that the proposed application rate (χ kg ha^{-1}, or comparable concentration appropriate for a beaker or microcosm experiment) not cause more than a 20% reduction in the response variable relative to its value in a control treatment. That is, we require that $\mu_T > 0.80\ \mu_C$.

There are two approaches to testing whether that proposition is true (Stunkard 1994). The classical approach would be to let $H_0{:}\mu_T \geq 0.8\ \mu_C$ and $H_1{:}\mu_T < 0.8\ \mu_C$ and to reject H_0 only if $P \leq \alpha$. This approach is counterproductive relative to the regulatory objective, however. It encourages weak experimental design and increases the chance that the insecticide will be approved even if $\mu_T < 0.8\ \mu_C$. The pesticide manufacturer, which is often in charge of conducting the test, has an incentive to use few replicate experimental units per treatment and to exercise little control over extraneous variables. This will favour high P values and conclusions, such as 'the insecticide did not cause survival rate to decrease significantly more than 20%' – even when the data might show that $m_T << 0.8\ m_C$.

The simple solution to the above is to simply reverse the null and alternative hypotheses. Then we have $H_0{:}\mu_T \leq 0.8\ \mu_C$ and $H_1{:}\mu_T > 0.8\ \mu_C$. That is we start with the presumption that there will be an unacceptably adverse effect, and require the manufacturer to provide strong evidence ($P \leq \alpha$) that the presumption is false before granting approval to the insecticide. If it is true that $\mu_T > 0.8\ \mu_C$, then the manufacturer will be rewarded by setting up an experiment with a strong design and high power: If, in contrast, $\mu_T < 0.8\ \mu_C$, the chance of approval is extremely low and society is afforded the protection the regulations are aimed at providing. Moreover, although in this case greater power is of no value to the manufacturer, neither is it disadvantageous. There is no disincentive to strong design and high precision. Be the experiment strong or weak, P will be high and the insecticide will not be approved.

Rigid and unimaginative use of this approach can be unproductive for both the manufacturer and society, however. As a colleague reminds us, large field experiments on pesticide effects can be well-designed, extremely expensive, and still have low precision (P. Chapman 1993, pers. comm.). Thus even when $\mu_T >> 0.8\ \mu_C$, these experiments can fail to secure official approval of the pesticide, just as when the conventional test procedure is used and $\mu_T >> \mu_C$, they can fail

doi:10.1111/j.1442-9993.2009.01946.x

to detect the adverse effect (Shaw *et al.* 1994). Three possible solutions are evident: increase precision by the 'brute force' and expensive approach of markedly increasing replication; improve designs and protocols in ways that might increase precision with no or little additional expense; or set α at a higher value than the bureaucratically overfavoured one of 0.05. This last possibility merits discussion among ecotoxicologists. Why should it not be sufficient, that $m_T > 0.8\ m_C$ and $P < 0.25$ when H_0: $\mu_T \leq 0.8\ \mu_C$?

In the other testing situation mentioned, we wish to estimate the dose or concentration of the substance likely to cause a particular magnitude of effect. Classic examples are the estimation of LD_{50} (lethal dose) and EC_{50} (effective concentration) values, the concentrations which, during a specified time interval, cause a 50% reduction in survival or in some other characteristic of test population. These values are useful mainly as indicators of the relative toxicity of different substances, or the relative sensitivity of different species. One-tailed tests and one-sided confidence intervals have no special use relative to them.

More recently, scientists and regulatory agencies have begun using experimentally determined dose response curves to estimate the dose or concentration of a substance which will have either no adverse effect or an adverse effect not greater than some specified, acceptably small (e.g. reduction of 1% or 10%) magnitude. These are known under various acronyms, such as LOEC or LOEL (lowest observed effect level), NOEC or NOEL (no observed effect level), EC_{10} (concentration causing a 10% reduction in the response variable), etc., and are already used for decision making by regulatory agencies. Yet some of these have undesirable statistical properties equivalent to those described above for testing the acceptability of an insecticide's effect with H_0: $\mu_T \geq 0.8\ \mu_C$ rather than with H_0:$\mu_T \leq 0.8\ \mu_C$. The weaker the experimental design, for example, the higher will be the estimated NOEL and the more likely the NOEL will be a concentration at which actual effects will be appreciable (Crump 1984; Hoekstra & Van Ewijk 1993a,b; Crane & Newman 2000).

A solution to this problem is found in Crump's (1984) Benchmark Dose (BD) and Hoekstra and Van Ewijk's (1993a,b), Bounded Effect Concentration (BECx). These are identical conceptually and differ only in their manner of calculation. They are defined as the concentration that will cause, at most, an X% reduction in the variable of interest. They are calculated from dose-response data by first estimating the ECx and then calculating a confidence interval about that value. The BD or BECx is defined as the lower bound of that interval. This BECx, usually divided by some safety factor, then becomes the maximum concentration that will be considered acceptable in a given environmental situation. This is essentially a one-sided procedure, regardless of whether the BECx is defined as the lower bound of a two-sided $(1-\alpha)$ confidence interval or that of a one-sided $(1-2\alpha)$ confidence interval.

There will be continued discussion as to the best way to calculate both ECx and its confidence interval. Conceptually, the general approach seems flawless, however. From the point of view of a pesticide manufacturer, for example, it rewards strong design and precision. These raise the value of BD or BECx to as close to ECx as is desired. Greater precision thus increases the maximal environmental concentration that will be considered acceptable and favours approval by the regulatory agency. This assumes, of course, that the regulatory agency does not arbitrarily respond to rising BECx values by lowering X or α or by increasing the safety factor!

The one-tailed test or one-sided confidence interval clearly has genuine utility in pharmaceutical and ecotoxicological decision-making situations, such as those described above. It is apparent, however, that these situations are atypical not only of basic research but also of most sorts of applied research as well. Certainly in the fields of ecology and animal behaviour, the original focus of our investigation, situations where one-tailed tests will be appropriate and useful are vanishingly rare.

A NOTE ON BAYESIAN APPROACHES

Given their increasing use, we offer a few comments on how one-tailed tests would be handled by Bayesian methods. The Bayesian version of a one-tailed test would proceed as follows. First, one would have to define a prior probability of zero for all possible results in the unpredicted direction. Then one would state a prior for H_0, and typically one much higher than 0.0 is recommended despite near universal agreement that in almost all testing situations the null hypothesis of 'no difference' is false. The remainder of the prior probability distribution is then defined in a subjective and largely arbitrary manner for the set of conceivable results. And finally the posterior probabilities of H_0 and H_1 are calculated.

By some reports, Bayesian methods would seem about to replace more classical frequentist procedures. For example, it is stated that 30% of the 2001–2005 articles in the *Journal of the American Statistical Association* concern Bayesian statistics (Wagenmakers & Grunwald 2006) and that 'Bayesian inference is fast becoming an accepted statistical tool among ecologists' (Ellison 2004).

We believe, however, that such claims exaggerate the value of Bayesian methods for the commonest data analysis situations. The literature on Bayesian statistics has a surfeit of inclarities and errors well beyond the

doi:10.1111/j.1442-9993.2009.01946.x

scope of the present article to analyse and correct. Problems include: the usual widespread confusion over the distinction between statistical hypotheses and scientific hypotheses; failure to recognize that inferential methods in most disciplines are useful mainly, albeit implicitly, for purposes of estimation; a focus on uncommon types of testing situations, such as a comparison of a point null and a point alternative hypothesis; claims of 'objectivity' based on highly biased prior probabilities, as when the H_0 of 'no effect' is assigned a prior of 0.5; and, most generally, the injection of more subjectivity into statistical analyses before interpretation of results of such analyses.

In the face of such problems, persistence of strong advocacy of Bayesian methods as a substitute for standard sorts of significance testing might be due in part to 'a recurring cycle of criticism of classical [frequentist] statistical inference' (Harris 2005), in which critics have routinely mistaken misuse of the classical methods by scientists and statisticians for flaws in the methods themselves (as documented, e.g. by Abelson 1995, 1997; Hagen 1997; Harris 1997, 2005; Mulaik *et al.* 1997; Reichardt & Gollob 1997; Rossi 1997; Nickerson 2000; Balluerka *et al.* 2005). That is, Bayesian methods might be perceived as a refuge from problems that either do not exist or have been exaggerated.

A good introduction to the diversity of opinion on Bayesian statistics is provided by Berger and Sellke (1987), Berger (2003), Hubbard and Bayari (2003) and Dennis (2004), and the commentaries by 15 other statisticians accompanying those articles, together with Spanos (1999), Spiegelhalter *et al.* (2004), Christensen (2005) and Cox (2006). Recent works on their use for multi-model inference in ecology include Dorazio and Johnson (2003), Ellison (2004) and Hobbs and Hilborn (2006).

CONCLUSIONS

This investigation was undertaken with the simple objective of calling attention to the misuse of one-tailed tests in animal behaviour and ecology and of offering corrective advice. Curiosity as to the origins and history of the problem soon led us into a sociostatistical labyrinth, full analysis of which is reserved for companion articles (Hurlbert & Lombardi 2009 & 2008, upubl.) Delving into the psychological literature, we found that the simple technical and philosophical issues involved had been resolved by 1960. Unfortunately the message of key papers by Kimmel (1957), Goldfried (1959) and Eysenck (1960) rarely was incorporated into statistics texts or reference works. Instead the contrary and poorly conceived advice of certain early statistics texts (e.g. Edwards 1954; McNemar 1955; Siegel 1956) became, in many disciplines, the conventional wisdom: if the direction of a result is predicted, use a one-tailed test. Even most modern statistics books list this 'prediction' criterion as a valid one (Appendix S1). Our analysis, in contrast, supports those few authors who have argued that prediction is never a valid justification for use of one-tailed tests. The claim that there is 'interest only' in results in a particular direction should be acceptable only if exceptional circumstances make it clear that the investigator truly would have been willing to disregard results strongly in the direction supposedly 'not of any interest' *and* only if such a contrary result would have been of no interest to science or society as a whole.

In the end, our animal behaviourist and ecologist colleagues stand somewhat exculpated, given the quality of advice available to them (Appendix S1). They must take more individual responsibility, however, for evaluating the logical foundations underlying the quantitative methodologies they use. The apparent marked decrease between 1989 and 2005 in use of one-tailed tests in the two journals analysed suggests that, on this topic, they have already begun to do so.

ACKNOWLEDGEMENTS

This work was supported in part by a grant from the Association for the Study of Animal Behaviour to CML and a grant from the National Science Foundation (BSR-8509535) to SHH. C. M. L. also thanks San Diego State University for its hospitality and use of its facilities during her study visit there. For helpful comments on the manuscript, we especially thank P. Bateson, C.R. Chandler, P. Chapman, B.D. Collier, R.B. Darlington, J. Delius, D.A. Farris, E. García-Berthou, J. Hoekstra, S.K. Kachigan, G. Matt, M. Milinski, R.J. Harris, M. Riggs, J.L. Shaw, S.M. Smith, C. Stunkard, M.J. West and so on. *We dedicate this paper to Michael Riggs for his incisive criticisms of early versions of the manuscript and for forcing us to read more widely, and for more years, than we desired.*

REFERENCES

Abelson R. P. (1995) *Statistics as Principled Argument*. Laurence Erlbaum, Hillsdale.

Abelson R. P. (1997) A retrospective on the significance test ban of 1999 (If there were no signficance tests, they would be invented). In: *What If There Were No Significance Tests?* (eds L. L. Harlow, S. A. Mulaik & J. H. Steiger) pp. 117–44. Laurence Erlbaum Associates, Mahwah.

Altman D. G. (1991) *Practical Statistics for Medical Research*. Chapman and Hall, New York.

Animal Behaviour (2004) Instructions to authors. *Anim. Behav.* **68,** ii–vii.

doi:10.1111/j.1442-9993.2009.01946.x

Bakan D. (1966) The test of significance in psychological research. *Psychol. Bull.* **66,** 423–37.
Balluerka N., Gomez J. & Hidalgo D. (2005) The controversy over null hypothesis significance testing revisted. *Methodology* **1,** 55–70.
Barker L., Rolka H., Rolka D. & Brown C. (2001) Equivalence testing for binomial random variables: which test to use? *Am. Stat* **55,** 279–87.
Bart J., Fligner M. A. & Notz W. I. (1998) *Sampling and Statistical Methods for Behavioral Ecologists*. Cambridge University Press, Cambridge.
Berger J. O. (2003) Could Fisher, Jeffreys and Neyman have agreed on testing? *Stat. Sci.* **18,** 1–32.
Berger J. O. & Sellke T. (1987) Testing a point null hypothesis: the irreconciliability of *P* values and evidence (with comments). *J. Am. Stat. Assoc.* **82,** 112–39.
Braver S. L. (1975) On splitting the tails unequally: a new perspective on one- versus two-tailed tests. *Educ. Psychol. Meas.* **35,** 283–301.
Burke C. J. (1953) A brief note on one-tailed tests. *Psychol. Bull.* **50,** 384–7.
Casella G. & Berger R. L. (1987) Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Am. Stat. Assoc.* **82,** 106–11.
Christensen R. (2005) Testing Fisher, Neyman, Pearson, and Bayes. *Am. Stat.* **59,** 121–6.
Cohen J. (1977) *Statistical Power Analysis for the Behavioral Sciences*, 1st edn. Academic Press, New Cork.
Cohen J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Academic Press, New Cork.
Conover W. J. (1980) *Practical Nonparametric Statistics*. Wiley, New York.
Cox D. R. (2006) *Principles of Statistical Inference*. Cambridge University Press, Cambridge.
Crane M. & Newman M. C. (2000) What level of effect is a no observed effect? *Environ. Toxicol. Chem.* **19,** 516–9.
Crawley M. J. (1993) *GLIM for Ecologists*. Blackwell, Oxford.
Crump K. S. (1984) A new method for determining allowable daily intakes. *Fund. Appl. Toxicol.* **4,** 854–71.
D'Agostino R. B. Sr, Massaro J. M. & Sullivan L. M. (2003) Non-inferiority trials: design concepts and issues: the encounters of academic consultants in statistics. *Stat. Med.* **22,** 169–86.
Darlington R. B. & Carlson P. M. (1987) *Behavioral Statistics*. The Free Press, New York.
Dennis R. (2004) Statistics and the scientific method in ecology. In: *The Nature of Scientific Evidence* (eds M. L. Taper & S. R. Lele) pp. 327–78. University of Chicago Press, Chicago.
Dorazio R. M. & Johnson F. A. (2003) Bayesian inference and decision theory – a framework for decision making in natural resource management. *Ecol. Appl.* **13,** 556–63.
Edwards A. L. (1954) *Statistical Methods for the Behavioral Sciences*. Rinehart, New York.
Ellison A. M. (2004) Bayesian inference in ecology. *Ecol. Lett.* **7,** 509–20.
EPA (1989) *Methods for Evaluating the Attainment of Cleanup Standards, Volume 1: Soils and Solid Media*. Report 230/02-89/042, Office of Policy, Planning and Evaluation, U.S. Environmental Protection Agency, Washington DC.
Eysenck H. J. (1960) The concept of statistical significance and the controversy about one-tailed tests. *Psychol. Rev.* **67,** 269–71.
FDA (2003) *Bioavailability and Bioequivalence Studies for Orally Administered Drug Products – General Considerations*. Federal Drug Administration, Rockville. [Cited 13 March 2009.] Available from URL: http://www.fda/gov/cder/guidance/5356fnl.pdf
Feinstein A. R. (1974) Clinical biostatistics, XXV: a survey of the statistical procedurees in general medical journals. *Clin. Pharmacol. Ther.* **15,** 97–107.
Ferguson G. A. (1971) *Statistical Analysis in Psychology and Education*, 3rd edn. McGraw-Hill Book Company, New York.
Fisher R. A. (1960) *The Design of Experiments*, 7th edn. Oliver and Boyd, Edinburgh.
Fisher R. A. (1971) *The Design of Experiments*, 9th edn. Hafner Press, New York.
Fleiss J. L. (1981) *Statistical Methods for Rates and Proportions*, 2nd edn. Wiley, New York.
Fleiss J. L. (1986) *The Design and Analysis of Clinical Experiments*. Wiley and Sons, New York.
Fleiss J. L. (1987) Some thoughts on two-tailed tests. *J. Control Clin. Trials* **8,** 394.
Freedman D., Pisani R. & Purves R. (1991) *Statistics*, 2nd edn. Norton, New York.
Freedman D., Pisani R. & Purves R. (1998) *Statistics*, 3rd edn. Norton, New York.
Gaines S. D. & Rice W. R. (1990) Analysis of biological data when there are ordered expectations. *Am. Nat.* **135,** 310–7.
Gibbons J. D. & Pratt J. W. (1975) *P*-values: interpretation and methodology. *Am. Stat.* **29,** 20–5.
Glass G. V. & Hopkins K. D. (1984) *Statistical Methods in Education and Psychology*, 2nd edn. Englewood Cliffs, Prentice-Hall.
Goldfried M. R. (1959) One-tailed tests and 'unexpected' results. *Psychol. Rev.* **66,** 79–80.
Hagen R. L. (1997) In praise of the null hypothesis statistical test. *Am. Psychol.* **52,** 15–24.
Hand J. & McCarter R. E. (1985) The precedures and justification of a two-tailed directional test of significance. *Psychol. Rep.* **56,** 495–8.
Harris R. J. (1997) Reforming significance testing via three-valued logic. In: *What If There Were No Significance Tests?* (eds L. L. Harlow, S. A. Mulaik & J. H. Steiger) pp. 145–74. Laurence Erlbaum Associates, Mahwah.
Harris R. J. (2005) Classical statistical inference: practice versus presentation. In: *Encyclopedia of Statistics in Behavioral Science*, Vol. 1 (eds B. S. Everitt & D. C. Howell) pp. 268–78. Wiley, Chichester.
Hawkins D. (2005) *Biomeasurement*. Oxford Univ. Press, New York.
Hays W. L. (1981) *Statistics*. Holt, Rinehart, & Winston, Inc., New York.
Helsel D. R. & Hirsch R. M. (1992) *Statistical Methods in Water Resources*. Elsevier, Amsterdam.
Hobbs N. T. & Hilborn R. (2006) Alternatives to statistical hypothesis testing in ecology: a guide to self teaching. *Ecol. Appl.* **16,** 5–19.
Hodges J. L. & Lehmann E. L. (1954) Testing the approximate validity of statistical hypotheses. *J. R. Stat. Soc. B* **16,** 261–8.
Hoekstra J. A. & Van Ewijk P. H. (1993a) Alternatives for the no-observed-effect level. *Env. Toxicol. Chem.* **12,** 187–94.
Hoekstra J. A. & Van Ewijk P. H. (1993b) The bounded effect concentration as an alternative to the NOEC. *Sci. Total Environ.* Suppl, 705–11.
Hollander M. & Wolfe D. A. (1973) *Nonparametric Statistical Methods*. Wiley, New York.
Hubbard R. & Bayari M. J. (2003) Confusion over measures of evidence (*P*'s) versus errors (α's) in classical statistical testing. *Am. Stat* **57,** 171–82.

doi:10.1111/j.1442-9993.2009.01946.x

Hurlbert S. H. & Lombardi C. M. (2009) Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Ann. Zool. Fenn.* (in press).
Hurlbert S. H. & White M. D. (1993) Experiments with freshwater invertebrate zooplanktivores: quality of statistical analyses. *Bull. Mar. Sci.* **53,** 128–53.
Jeffreys H. (1939) *Theory of probability*. Oxford University Press, London.
Jones L. V. (1952) Tests of hypotheses: one-sided vs. two-sided alternatives. *Psychol. Bull.* **49,** 43–6.
Kaiser H. F. (1960) Directional statistical decisions. *Psychol. Rev.* **67,** 160–7.
Kimmel H. D. (1957) Three criteria for the use of one-tailed tests. *Psychol. Bull.* **54,** 351–3.
Kirk R. E. (1982) *Experimental Design*, 2nd edn. Brooks/Cole Publishing Company, Pacific Grove.
Koch G. G. & Gillings D. B. (1988) Tests, one-sided versus two-sided. In: *Encyclopedia of Statistical Sciences*, Vol. 9 (eds S. Kotz & N. L. Johnson) pp. 218–22. John Wiley and Sons, New York.
Leventhal L. & Huynh C.-L. (1996a) Directional decisions for two-tailed tests: power, error rates, and sample size. *Psychol. Methods.* **1,** 278–92.
Leventhal L. & Huynh C.-L. (1996b) Analyzing listening tests with the directional two-tailed test. *J. Audio Eng. Soc.* **44,** 850–63.
Lombardi C. M. & Curio E. (1985a) Influence of environment on mobbing by Zebra finches. *Bird Behav.* **6,** 28–33.
Lombardi C. M. & Curio E. (1985b) Social facilitation of mobbing in the Zebra finch *Taeniopygia guttata*. *Bird Behav.* **6,** 34–40.
McBride G. B. (1999) Equivalence tests can enhance environmental science and management. *Aust. N. Z. J. Stat.* **41,** 19–29.
McKinney W. P., Young M. J., Hartz A. & Lee M. B. (1989) The inexact use of Fisher's exact test in six major medical journals. *J. Am. Med. Assoc.* **261,** 3430–3.
McNemar Q. (1955) *Psychological Statistics*, 2nd edn. Wiley, New York.
Manly B. F. J. (2003) One-sided tests of bioequivalence with nonnormal distributions and unequal variances. *J. Agric. Biol. Environ. Stat.* **9,** 1–14.
Marks M. R. (1951) Two kinds of experiment distinguished in terms of statistical operations. *Psychol. Rev.* **58,** 179–84.
Martin P. & Bateson P. (1986) *Measuring Behaviour: An Introductory Guide*, 1st edn. Cambridge University Press, Cambridge.
Martin P. & Bateson P. (1993) *Measuring Behaviour: An Introductory Guide*, 2nd edn. Cambridge University Press, Cambridge.
Martin P. & Bateson P. (2007) *Measuring Behaviour: An Introductory Guide*, 3rd edn. Cambridge University Press, Cambridge.
Meddis R. (1984) *Statistics Using Ranks*. Blackwell Scientific Publications, Oxford.
Meehl P. E. (1967) Theory-testing in psychology and physics: A methodological paradox. *Philos. Sci.* **34,** 103–15.
Mulaik S. A., Raju N. S. & Harshman R. A. (1997) Reforming significance testing via three-valued logic. In: *What If There Were No Significance Tests?* (eds L. L. Harlow, S. A. Mulaik & J. H. Steiger) pp. 65–116. Laurence Erlbaum Associates, Mahwah.
Neyman J. (1937) Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Trans. R. Soc. (A)* **236,** 333–80.
Neyman J. (1950) *First Course in Probability and Statistics*. Henry Holt, New York.
Neyman J. (1976) The emergence of mathematical statistics. In: *On the History of Statistics and Probability* (ed. D. B. Owen) pp. 149–93. Dekker, New York.
Neyman J. & Pearson E. S. (1933) On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. London, A* **231,** 289–337.
Nickerson R. S. (2000) Null hypothesis significance testing: a reviewof an old and continuing controversy. *Psychol. Meth.* **5,** 241–301.
Oakes M. (1986) *Statistical Inference: A Commentary for the Social and Behavioural Sciences*. Wiley, New York.
Overall J. E. (1991) A comment concerning one-sided tests of significance in new drug applications. *J. Biopharm. Stat.* **1,** 157–60.
Peace K. E. (1989) The alternative hypothesis: one-sided or two-sided? *J. Clin. Epidemiol.* **42,** 473–6.
Peace K. E. (1988) Some thoughts on one-tailed tests. *Biometrics* **44,** 911–2.
Peace K. E. (1991) One-sided or two-sided *P* values: which most appropriately address the question of drug efficacy? *J. Biopharm. Stat.* **1,** 133–8.
Peizer D. B. (1967) A note on directional inference. *Psychol. Bull.* **68,** 448.
Piaggio G., Elbourne D. R., Altman D. G., Pocock S. J. & Evans S. J. W. (2006) Reporting of noninferiority and equivalence randomized trials. *J. Am. Med. Assoc.* **295,** 1152–61.
Pillemer D. B. (1991) One- versus two-tailed hypothesis tests in contemporary educational research. *Educ. Res.* **20,** 13–7.
Pocock S. J. (2000) The pros and cons of non-inferiority (equivalence) trials. In: *The Science of Placebo: Towards an Interdisciplinary Research Agenda* (eds H. A. Guess, A. Kleinman, J. W. Kusek & L. W. Engel) pp. 236–48. BMJ Books, London.
Reichardt C. S. & Gollob H. F. (1997) When confidence intervals should be used instead of statistical significance tests, and vice versa. In: *What If There Were No Significance Tests?* (eds L. L. Harlow, S. A. Mulaik & J. H. Steiger) pp. 259–86. Laurence Erlbaum Associates, Mahwah.
Rice W. R. & Gaines S. D. (1994) Extending nondirectional heterogeneity tests to evaluate simply ordered alternative hypotheses. *Proc. Natl. Acad. Sci., USA* **91,** 225–6.
Rossi J. S. (1997) A case study in the failure of psychology as a cumulative science: the spontaneous recovery of verbal learning. In: *What If There Were No Significance Tests?* (eds L. L. Harlow, S. A. Mulaik & J. H. Steiger) pp. 175–98. Laurence Erlbaum Associates, Mahwah.
Royall R. M. (1997) *Statistical Evidence: A Likelihood Paradigm*. Chapman and Hall, New York.
Salsburg D. S. (1992) *The Use of Restricted Significance Tests in Clinical Trials*. Springer, New York.
Scheiner S. M. & Gurevitch J., eds (1993) *Design and Analysis of Ecological Experiments*. Chapman & Hall, New York.
Schulman R. S. (1992) *Statistics in Plain English*. Van Nostrand Reinhold, New York.
Shaffer J. P. (1972) Directional statistical hypotheses and comparisons among means. *Psychol. Bull.* **77,** 195–7.
Shaw J. L., Moore M., Kennedy J. H. & Hill I. R. (1994) Design and statistical analysis of field aquatic mesocosm studies. In: *Aquatic Mesocosm Studies in Ecological Risk Assessment* (eds R. L. Graney, J. H. Kennedy & J. H. Rodgers) pp. 85–103. Lewis Publ., Boca Raton, Florida.
Siegel S. (1956) *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York.

doi:10.1111/j.1442-9993.2009.01946.x

Siegel S. & Castellan, N. J. Jr (1988) *Nonparametric Statistics for the Behavioral Sciences*, 2nd edn. McGraw-Hill, New York.
Snedecor G. W. (1956) *Statistical Methods*, 5th edn. Iowa State University, Ames.
Snedecor G. W. & Cochran W. G. (1967) *Statistical Methods*, 6th edn. Iowa State University, Ames.
Snedecor G. W. & Cochran W. G. (1980) *Statistical Methods*, 7th edn. Iowa State University, Ames.
Snedecor G. W. & Cochran W. G. (1989) *Statistical Methods*, 8th edn. Iowa State University, Ames.
Sokal R. R. & Rohlf F. J. (1969) *Biometry*, 1st edn. Freeman, San Francisco.
Sokal R. R. & Rohlf F. J. (1981) *Biometry*, 2nd edn. Freeman, San Francisco.
Sokal R. R. & Rohlf F. J. (1995) *Biometry*, 3rd edn. Freeman, San Francisco.
Spanos A. (1999) *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge University Press, Cambridge.
Spiegelhalter D. J., Abrams K. R. & Myles J. P. (2004) *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, New York.
Steel R. G. D. & Torrie J. H. (1960) *Principles and Procedures of Statistics*, 1st edn. McGraw-Hill, New York.
Steel R. G. D. & Torrie J. H. (1980) *Principles and Procedures of Statistics*, 2nd edn. McGraw-Hill, New York.
Steel R. G. D., Torrie J. H. & Dickey D. A. (1997) *Principles and Procedures of Statistics*, 3rd edn. McGraw-Hill, New York.
Stunkard C. L. (1994) Tests of proportional means for mesocosm studies. In: *Aquatic Mesocosm Studies in Ecological Risk Assessment* (eds R. L. Graney, J. H. Kennedy & J. H Rodgers) pp. 71–83. Lewis Publ., Boca Raton.
Tamayo-Sarver J. H., Albert J. M. & Tamayo-Sarver M. (2005) Advanced statistics: how to determine whether your intervention is different, at least as effective as, or equivalent: a basic introduction. *Acad. Emerg. Med.* **12,** 536–42.
Tempelman R. J. (2004) Experimental design and statistical methods for classical and bioequivalence hypothesis testing with an application to dairy nutrition studies. *J. Anim. Sci.* **82** (E. Suppl), E162–72.
Tryon W. W. (2001) Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. *Psychol. Meth.* **6,** 371–86.
Tukey J. W. (1991) The philosophy of multiple comparisons. *Stat. Sci.* **6,** 100–16.
Underwood A. J. (1990) Experiments in ecology and their management: their logics, functions, and interpretations. *Aust. J. Ecol.* **15,** 365–89.
Underwood A. J. (1997) *Experiments in Ecology*. Blackwell, London.
Upton G. J. G. (1992) Fisher's exact test. *J. R. Stat. Soc. B* **155,** 395–402.
Wagenmakers E.-J. & Grunwald P. (2006) A Bayesian perspective on hypothesis testing: a comment on Killeen (2005). *Psychol. Sci.* **71,** 641–2.
Welkowitz J., Ewen R. B. & Cohen J. (1971) *Introductory Statistics for the Behavioral Sciences*, 1st edn. Harcourt Brace Jovanovich, New York.
Welkowitz J., Ewen R. B. & Cohen J. (1991) *Introductory Statistics for the Behavioral Sciences*, 4th edn. Harcourt Brace Jovanovich, New York.
Welkowitz J., Ewen R. B. & Cohen J. (1999) *Introductory Statistics for the Behavioral Sciences*, 5th edn. Harcourt Brace Jovanovich, New York.
Welkowitz J., Cohen B. H. & Ewen R. B. (2006) *Introductory Statistics for the Behavioral Sciences*, 6th edn. Wiley, New York.
Welleck S. (2003) *Testing Statistical Hypotheses of Equivalence*. Chapman and Hall/CRC, Boca Raton.
Winer B. J., Brown D. R. & Michels K. M. (1962) *Statistical Principles in Experimental Design*, 1st edn. McGraw-Hill, New York.
Winer B. J., Brown D. R. & Michels K. M. (1971) *Statistical Principles in Experimental Design*, 2nd edn. McGraw-Hill, New York.
Winer B. J., Brown D. R. & Michels K. M. (1991) *Statistical Principles in Experimental Design*, 3rd edn. McGraw-Hill, New York.
Zar J. H. (1974) *Biostatistical Analysis*, 1st edn. Prentice-Hall, Inc., New York.
Zar J. H. (1984) *Biostatistical Analysis*, 2nd edn. Prentice-Hall, Inc., New York.
Zar J. H. (1999) *Biostatistical Analysis*, 4th edn. Prentice-Hall, Inc., New York.
Zar J. H. (2004) *Biostatistical Analysis*, 5th edn. Prentice-Hall, Inc., New York.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Advice given by 52 statistics books on when one-tailed tests are appropriate (PDF)

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

doi:10.1111/j.1442-9993.2009.01946.x