Journal of Experimental Marine Biology and Ecology 426–427 (2012) 97–108

Contents lists available at SciVerse ScienceDirect

Journal of Experimental Marine Biology and Ecology

journal homepage: www.elsevier.com/locate/jembe

Strengthening statistical usage in marine ecology

Peter G. Beninger [a,*], Inna Boldina [a], Stelios Katsanevakis [b]

[a] Laboratoire de Biologie Marine, Faculté des Sciences, Université de Nantes, 2 rue de la Houssinière, 44322 Nantes Cedex 1, France
[b] European Commission, Joint Research Centre, Institute for Environment and Sustainability, Ispra, Italy

ARTICLE INFO

Article history:
Received 9 March 2012
Received in revised form 23 May 2012
Accepted 24 May 2012
Available online 19 June 2012

Keywords:
Alternate approaches
Data treatment/presentation
NHST
p-Values
Statistics misuse

ABSTRACT

Although within their own disciplines, the statistical, social science, medical, and terrestrial ecology literatures are replete with accounts of the widespread misapplication and misuse of statistical testing and interpretation, awareness of these issues is weak among marine scientists who are not statisticians, but whose work is nonetheless situated within the expanse of marine ecology. Moreover, the major recent developments in statistical approaches in these fields are, as yet, poorly-represented in the marine ecological literature. We present a non-technical review of (1) the most fundamental, yet pervasive, problems concerning classical statistics, with suggestions for improved practice, (2) alternate, often more appropriate and intuitive, approaches to statistical design and interpretation, and (3) the crucial roles of reviewers, and especially of editors and editorial boards. It is hoped that increasing the awareness of these issues will strengthen statistical usage in marine ecology.

Contents

* Corresponding author.
E-mail address: Peter.Beninger@univ-nantes.fr (P.G. Beninger).

0022-0981/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.jembe.2012.05.020

1. Introduction

The capacity of the human brain being what it is, no researcher can fully master all areas of science touching upon marine ecology. Hence, scientists who use statistics in their research relating more or less directly to marine ecology may be divided into three broad categories: (1) quantitative marine ecology specialists, i.e. those whose main level of expertise is in sampling and experimental design, data treatment and interpretation; these workers tend to focus upon the upper levels of ecological organization, e.g. populations and communities; (2) those working on biological questions related to marine ecology and who have upper-level familiarity with statistical procedures; they mainly focus upon the interaction of individual organisms with the environment, and are often unaware, or only vaguely so, of the numerous and lively debates on statistical methods in the statistical literature; (3) those whose main area of expertise is outside of statistics and who consider it as a kind of recipe book to be followed in order to be taken seriously by journal reviewers; they often work at the sub-individual level, e.g. in physiology, chemical ecology or cellular ultrastructure. The first category of scientists includes several notable 'giants' in the field of quantitative marine ecology, who have stressed the importance of constructing studies with the paramount objective of statistical robustness (e.g. Green, 1979,1989; Hurlbert, 1984; Peterson et al., 2001; Stewart-Oaten, 1995; Underwood, 1997; Underwood and Chapman, 2003). The present review is addressed primarily to the latter two categories of researchers, whom we here refer to as 'functional' ecologists (for lack of a better term), and who contribute the majority of papers touching upon ecology in marine science journals.

Assuming an adequate methodology, study results may be compromised at two critical, and interdependent, stages: planning the study, and analyzing/interpreting the subsequent data. While both of these areas have been abundantly discussed in the biological literature, statistical usage is subject to recurring problems of misuse and misinterpretation. Apart from run-of-the-mill quibbling over whether this or that test is most appropriate for this or that experimental design, there are much more fundamental problems which arise repeatedly, whether we are aware of them or not, in the use of statistics. Indeed, entire volumes and hundreds of important papers have been written concerning the past and present misconception, misuse, and misinterpretation of statistics in virtually all fields of research (e.g. Anderson et al., 2001; Cohen, 1994; Cumming and Finch, 2005; Hubbard and Bayarri, 2003; Huck, 2011; Morrison and Henkel, 1970; Sellke et al., 2001; Ziliak and McCloskey, 2008a), lending credence to Mark Twain's (1907) summary verdict that 'there are three kinds of lies: lies, damned lies, and statistics'. This may come as a surprise to some readers, who are accustomed to thinking of statistics as a tool to settle questions, not to raise them.

It may also come as a surprise to many marine ecologists that the social sciences (especially psychology) and medical sciences adopted statistical data analysis well before the field of ecology (mid 1950s vs mid-1960s), and that these fields have been on the forefront of new developments in statistical practice over the past two decades (a similar comment was made by Germano, 1999). Influenced by these developments, terrestrial ecologists have made appreciable inroads toward improved statistical procedures in recent years (Anderson et al., 2001; Fidler et al., 2006; Stephens et al., 2005; Yoccoz, 1991), but these considerations appear to be poorly-represented in the work of all but the most confirmed quantitative ecologists. Although specific suggestions are scattered throughout the marine ecological literature, we are aware of only two primarily-marine papers which have delved into major areas of the problem in the past 13 years, and only one of them was in a marine journal (Germano, 1999; Gerrodette, 2011).

The basic problem for biologists is the desire to establish incontrovertible 'facts' from numerical data gathered in the study of living organisms. Contrary to what most of us believe (and have usually been taught), the situations in which this may actually be done are very rare, and most ecological studies are far too complex in scope to permit such a thing, within the bounds of the statistical methodology employed. Furthermore, we would like to accomplish these objectives with *absolutely no subjective input*, ostensibly to show that our data interpretation is completely objective (Berger and Berry, 1988). There are several reasons why neither of these desires are realistic, as will be outlined below.

Here we present a brief, non-technical overview of the problems and perspectives regarding the foundations and interpretation of statistical analyses in marine ecology. To enhance focus, we will exclude spatial statistics, and concentrate on statistics of experimental or observational data and their component descriptors. We first review the common fundamental problems encountered with 'classical' statistical usage in marine biology/ecology, with suggestions for remediation, and then outline alternate approaches which allow us to more adequately design studies and analyze some types of ecological data. The pervasiveness of the problems encountered with statistical usage in marine biology/ecology are such that there is no justification for casting the first stone — and for this reason, we choose not to single out particular studies as examples not to be followed.

2. Epistemological fundamentals

2.1. Reasoned judgment vs blinkered mechanics

A mechanistic approach to data treatment has often replaced intelligent data interpretation, and this has been lamented by many statisticians in many fields. Scientists often feel that they must treat data in a certain stereotypical fashion in order to be taken seriously by their peers (Stewart-Oaten, 1995). There are two very important points to make on this subject: *(1) all statistical treatments rely on reasoned judgment, whether the scientist uses it or not, so it is impossible to think of statistics as a simple, blind, 'scale of justice'; (2) a failure to use reasoned judgment in statistical treatment of data is a fundamental abdication of responsibility which calls into question any subsequent conclusions.* Consider the following example: we wish to discover where the administration building is situated on a sloping university campus. Two methods are proposed: (1) draw random transects through the campus and fix sampling points on each transect; at each sampling point, ask if this is the administration building. (2) Our prior experience having shown that administration buildings are usually located at the dominant topographical feature, we decide to proceed to the building at the top of the hill and ask if it is the administration building. Both approaches will give us the correct answer, but the first approach is likely to require so much effort and cost that we may decide not to pursue the question at all.

2.2. Observational vs experimental studies

Although most statistics texts underscore the importance of randomization to the underpinnings of experimental study, this cumbersome requirement is often not satisfied in many such studies, and even less so in observational studies (e.g. seasonal variations in reproductive activity, biochemical composition, comparisons between different geographical locations, etc.). However, it must be remembered that randomization at the planning stage is a fundamental requirement for classical statistics, both parametric and non-parametric. Many observational studies apply classical hypothesis-testing statistics in their data treatment, and this generates seemingly meaningful numbers, but in reality such studies should rely much more on descriptive statistics and comparisons of effect sizes (Greenland, 1990; Rothman,

1990b), regardless of peer and reviewer pressure to the contrary. A second basic reason for eschewing classical statistics in observational studies is the requirement for hypothesis formulation based on a plausible theoretical framework, without which it is impossible to engage in precise interpretations of P-values or confidence intervals (Poole, 2001). Observational studies precede the first experimental studies, since it is impossible to formulate hypotheses when we know nothing at all about the systems studied.

2.3. Hypothesis testing vs significance testing

The majority of classical analyses are based upon a mixture of hypothesis testing and significance testing. Although these are today widely believed to be integral parts of a common approach, and presented this way in many statistics textbooks, Fisherian significance testing and Neyman–Pearson Type 1 error probability testing are derived from different foundations and intended for different objectives (Hubbard and Bayarri, 2003; Stefano et al., 2005). Indeed, the primitive genesis of the ensuing decades of confusion was Fisher's own misunderstanding of Gosset (=Student)'s precise meaning of 'statistical significance' (see Ziliak, 2011).

One of the basic, and far-reaching consequences of this, is the incongruity of associating Fisher's evidential P-value with Neyman–Pearson's Type 1 error rate (α); yet this is done routinely in many disciplines, including marine ecology, under the name of 'Null Hypothesis Significance Testing' (NHST), or simply 'Significance testing' (Fig. 1), and it has been said to render meaningless the extraordinarily numerous studies in which it has been performed (Hubbard and Bayarri, 2003). The situation is not helped by the ubiquitous, user-friendly statistical software which standardizes this approach, nor by recent papers which involuntarily muddy the waters by assigning a totally different meaning to α (see Christensen, 2005; Hubbard and Bayarri, 2005). A solution to this staggering problem has been proposed, by reporting observed P-values as lower bounds for Type 1 error probabilities (Sellke et al., 2001), but its complexity overshadows the actual tests originally performed. In the current state of affairs, we therefore present below the minimum precautions and guidelines for improved use of the Fisherian–Pearson NHST amalgam so prevalent today.

3. Problems encountered with null hypothesis significance testing (NHST)

3.1. Criticism of NHST

NHST has become a staple of ecological research since the 1960s, often in the form of Student t-tests or ANOVAs. Curiously, it has been severely criticized in hundreds of statistical papers throughout its rise to prominence (http://warnercnr.colostate.edu/~anderson/thompson1.html compiles 402 such papers up to 2001, and http://swfsc.noaa.gov/SignificanceTestRefs compiles 127 additional such papers up to 2010; of the recent papers, the following are particularly notable: Carver, 1978; Cohen, 1994; Fidler et al., 2006; Gelman and Stern, 2006; Germano, 1999; Gerrodette, 2011; Gigerenzer, 2004; Gigerenzer et al., 2004; Hubbard and Bayarri, 2003; Johnson, 1999; Martínez-Abraín, 2007; Sellke et al., 2001; Silva-Ayçaguer et al., 2010; Stang et al., 2010; Yoccoz, 1991). Following the lead of medical, economic, and social science journals, such papers have appeared in theoretical and terrestrial ecology journals, along with illustrative biometric studies (e.g. Fidler et al., 2006; Johnson, 1999; Stephens et al., 2007; Yoccoz, 1991). The comment that NHST is the 'most common and flagrant misuse of statistics' is one of the milder conclusions many statisticians have drawn (Johnson, 1999). Indeed, it has even been said, in many ways, that if NHST has had such a long and prolific history of misuse and misinterpretation, this may be more due to its contorted (or at least counter-intuitive) logic than to widespread deficiencies among its users (e.g. Beyth-Marom et al., 2008; Goodman, 1999; Sterne and Smith, 2001). The logic has been summarized as 'If *A* is true, *B* will happen sometimes; therefore if B has been found to happen, *A* can be considered disproved" (Berkson, 2003), which indeed seems contradictory and counter-intuitive! To paraphrase Berkson (2003), when confronted with a corpse, we do not say 'this is evidence against the hypothesis that no one is dead'! Rather, we say 'this person is dead', which is much more in line with our natural pattern of cognition. Yet for all its convoluted reasoning, NHST is the paradigm most researchers in marine ecology use today, and the plethora of problems it engenders, and recommendations for improved usage, must be highlighted as long as it remains current.

Much of the criticism of NHST has centered upon P-values and their interpretation (Fisherian component), and the inadequacy of

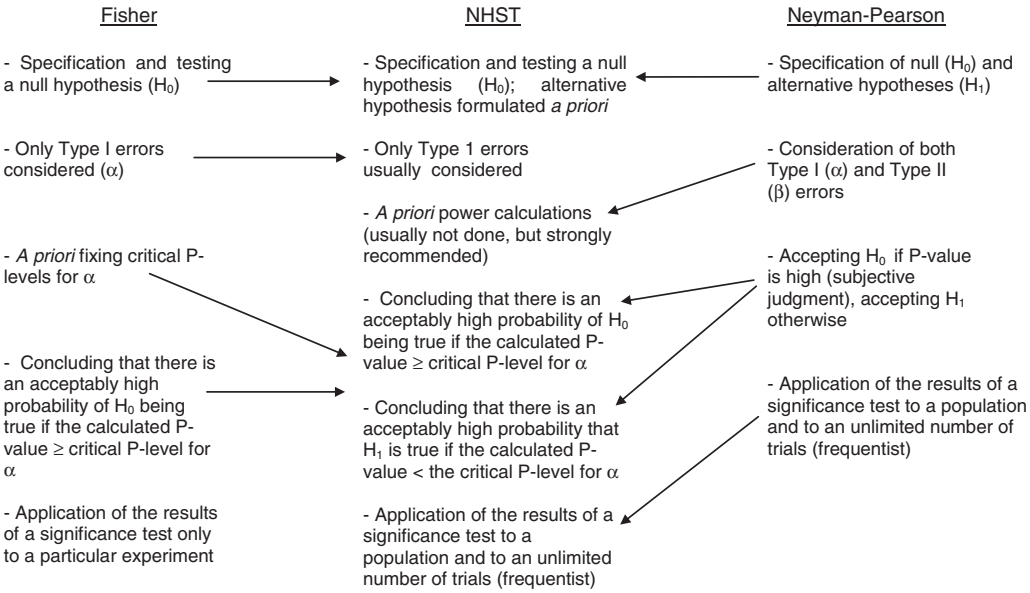

Fig. 1. Characteristics of Fisherian, Neyman–Pearson, and NHST approaches to evaluating differences.

mechanical cut-off levels for α (Neyman–Pearson component). These being the most frequent problems encountered in marine ecology papers, we will outline them briefly.

3.2. α levels and P-values

In the Fisherian–Pearson amalgam, P is the probability that a given effect could arise if the null hypothesis, or any hypothesis not envisaged in the alternative hypothesis, were true (Type I error), whereas α is the probability value above which we reject our alternative hypothesis, also called the significance level. Although most statistics textbooks do vaguely mention that the significance levels of 0.05, 0.01 etc. are subjective, workers in marine ecology often recognize the 0.05 level as a sort of mechanical Occam's razor: <0.05, reject null hypothesis, ≥ 0.05, accept null hypothesis. Statisticians have decried this reasoning for decades (Cohen, 1994; Fisher, 1959; Gelman and Stern, 2006; Gigerenzer, 2004; Gigerenzer et al., 2004; Hubbard and Bayarri, 2003; Stephens et al., 2005; and the numerous references cited in these works). Quite apart from the fact that we cannot *accept* a null hypothesis (this is tantamount to 'proving' a null hypothesis, which is simply not possible), it is argued that the onus is on the researcher to establish an α in line with his/her 'evidence and ideas' (Fisher, 1959), or at least in line with the 'risk posed to science or society of false positive or false negative results' (Mapstone, 1995; Stephens et al., 2005). While this may be possible in some research fields (especially medicine), it either does not make much sense, or is impossible to accomplish, in most marine ecological studies. We must therefore be open to the possibility of significance levels different from 0.05, and to the justification for such levels, especially with respect to the minimization of Type 2 errors (i.e., the probability of accepting H_0 when it is in fact false; see below). An $\alpha = 0.1$ might be a sufficient Type 1 error level for concluding a difference in coloration of reef fish, for example, while 0.05 may be required to conclude that there is a difference in heavy metal concentration in species of fish consumed by humans.

To have any meaning, significance levels must be decided upon prior to testing, even in the case of post-hoc significance testing.

3.3. Misinterpretation of P-values

A conventional expression in the medical statistical world is that misinterpretation of P-values (for α, obviously) has killed more people than any other type of scientific misconduct. In marine ecology, it has doubtless been responsible for many errors in ecological interpretation, which have certainly unduly influenced environmental policy, and brought about negative economic consequences. In fact, there is only one thing we may conclude from a P-value: *the probability of obtaining the result (or test statistic generically termed Evidence, E) if H_0 were, in fact, true*: $P(E|H_0)$.

In other words, P-values merely specify whether or not an effect is likely to exist (Stefano et al., 2005); they give no information at all on the magnitude of the effect (effect size), and *are usually much less informative than graphic presentations of statistical descriptors and their confidence intervals* (Lang et al., 1998). Unfortunately, in many ecological studies, a variety of unjustified conclusions are drawn from P-values. Twelve such fallacies have been tabulated by Goodman (2008); the following is a list of the most egregious misinterpretations, and their mis-applications in marine ecological studies, compiled over years of reviewing:

A. Statistical significance.
   Assuming an $\alpha = 0.05$, $P \leq 0.05$, there is a 5% or less probability that we would obtain the stated result (i.e. Evidence) if H_0 were true. The most widespread of the misunderstandings about P-values is that most workers interpret them as $P(H_0|E)$, or the probability that the null hypothesis is true, given the test statistic or Evidence (the Odds-Against-Chance Fantasy — Carver, 1978). The difference between these two interpretations is not semantic, it is hugely important. An example of this type of error in reasoning would be: (1) Most people who face a firing squad die from bullet wounds, and (2) Most people who die from bullet wounds have received them from a firing squad!
   Assuming that we correctly interpret the P-value ≤ 0.05 as $P(E|H_0)$, we may thus conclude that H_0 is probably not true. Note that this is not the same as saying that the alternative hypothesis is true; indeed, many workers assume that if $P \leq 0.05$, then the complement, 0.95, must mean there is a 95% probability that the alternative hypothesis is true (the Valid Research Hypothesis Fantasy — Carver, 1978). Once again, this is not a correct conclusion: there may be other, more compelling hypotheses to explain the observed data, but we have either not controlled for them, or thought of them, or we are epistemologically unable to imagine them! *At best, then, a significant P-value can only tell us that there is a weak probability that the null hypothesis is true. We may suggest one or several alternate hypotheses, but in no way are any of them 'proven' by our significance test*.

Two additional types of erroneous conclusions are often drawn from statistical significance and associated P-values:

I. The observed difference is biologically significant. A statistically significant difference may be obtained from data entirely bereft of biological significance. It is absolutely fundamental that we distinguish between *statistical* significance and *biological* significance (Johnson, 1999; Mapstone, 1995; Stefano et al., 2005; Yoccoz, 1991). This is one of the most important, and most-neglected, concepts in marine biology/ecology.
   - Example 1: A comparison of gastropod sizes at two sites in the same bay. Since the probability of attaining the predetermined α level increases with sample size, even the most minimal, biologically insignificant differences may register as statistically-significant differences. Conceivably, a mean size difference of less than 1 mm for animals measuring approx. 60 mm could thus register as a statistically significant difference, but it probably has no biological significance.
   - Example 2: A statistically-significant difference in the distribution of a non-biologically significant character. In Beninger et al. (1995a), the particular shape of marine bivalve cilia previously considered to be sensory was shown to be artefactual; however, ANOVAs and Tukey tests on the distribution of these cilia revealed numerous, and inconsistent, significant differences (unpublished because meaningless). Similarly, statistically significant differences in the densities of certain marine bivalve mucocyte types were obtained for an anatomical structure in which the mucocytes were vestigial (Beninger et al., 1995b).

II. The smaller the P-value, the greater the effect (the Significance Fallacy). This misconception is perhaps somewhat less widespread than the previous one, but nonetheless quite common. The fact that the probability of Type 1 error is very small does not mean that the magnitude of the effect (macrophyte growth, photosynthetic response, etc.) is large. The magnitude of the observed difference (effect size) is neither greater nor lesser simply because the P value is lower or higher than the α level set at the design stage of the study. A P-value does not give any information at all on how strong or reliable a result may be, merely the probability that such a result could have arisen, had the null hypothesis been true.

B. Statistical non-significance
   When a calculated P-value exceeds that of the pre-determined α (taking $\alpha = 0.05$ for the sake of an example), it means that there is $>5\%$ probability that the observed result could be due to

something other than the effect we have tested. There are several common, erroneous interpretations of what this really means:

I. The null hypothesis is true. As stated previously, it is impossible to *prove* a null hypothesis. The lack of statistical significance could be due to the null hypothesis being true, but it could also be due to a myriad of other causes which we have not identified. So even a very general null hypothesis such as 'there is no effect' may not be meaningful — there may not have been an effect because of an interfering variable for which we have not controlled because we could not imagine its existence, and without which the effect would be manifest at the pre-determined α level. There is no reason to suppose that there is NO effect when P exceeds the chosen significance level, particularly when the data do point in the direction of the effect. *The automatic acceptance of the null hypothesis when P exceeds the significance level is a widespread form of 'corrupt science' in marine ecology*, which, paradoxically, its practitioners often consider one of the highest and most rigorous forms of science (Carver, 1978).

II. There is no statistically-significant difference between groups, so it is possible to combine the groups. This is one of the most common errors of interpretation. It is often encountered at the review stage by referees, and unfortunately it gets past this stage and into print all too often. We have all read the expression "As there was no significant difference between the two (or more!) groups, we therefore combined (pooled) the data …". It should be obvious from the foregoing that lack of a statistically-significant difference does *not* allow groups to be treated as though they were the same with respect to the effect measured. They should be treated as separate groups, regardless of statistical insignificance at the chosen α level. For example, at $\alpha > 0.05$, P-value 0.06, there is only a 6% probability that this evidence would arise if the null hypothesis were true, so it is very likely that the difference between the groups is due to something other than the null hypothesis. If the experiment is carefully controlled, this is still, in fact, evidence in favor of the effect we measure (see above), so there is no reason to treat the groups as though they were the same! It has been suggested that a useful guidepost for combining groups would be a very high P-value of ≥ 0.25 (Underwood, 1997), but one must keep in mind that this does not mean that the groups are equivalent with respect to the effect tested, only that there is an acceptably large probability that any difference is due to factors other than the alternate hypothesis.

III. In the absence of a statistically-significant difference for a given effect between groups, we may conclude that the groups are equal with respect to this effect. This is perhaps the most egregious misinterpretation of statistical non-significance. Effect equality may only be statistically ascertained, within specified limits of probability, using tests of equivalence and noninferiority. These may be relatively straightforward, such as confidence interval construction (Garrett, 1997; http://www.graphpad.com), or very sophisticated (Wellek, 2010). Moreover, such tests rely on reasoned judgment (e.g. in deciding what a significant effect size is), so they are not the objective statistical razors which many biologists so earnestly — and naïvely — seek.

Given the long history of misinterpretation and misuse of P-values in the biomedical fields (as in all others), many journals have placed severe constraints on their use. For example, the instructions to authors in the journal Epidemiology include the following: 'We strongly discourage the use of P-values…' (http://edmgr.ovid.com/epid/accounts/ifauth.htm).

C. Multiple comparisons

Most of us have dutifully followed what we were taught in our statistics classes, that is, multiple comparisons (e.g. t-tests) increase the risk of a Type 1 error with each comparison — so we incorporate a compensation for this (i.e. progressive decrease of α levels). In reality, 'to correct or not to correct' is a subject of debate within the statistical community, with the extreme views being blinkered correction at all times (current standard practice), or never correction (Cohen, 1994; Pernerger, 1998; Rothman, 1990a). Correction only under certain circumstances (which are rather rare in marine biology/ecology) has also been advocated (Cook and Farewell, 1996), as has been downplaying the importance of correction, in favor of effect size, study design, and prior judgment-determined outcome measures (Feise, 2002).

The arguments against correction run as follows: In marine biology/ecology the null hypothesis is usually really a 'nil' hypothesis, and therefore often patently false, so that the real Type 1 error rate is 0%, and only Type 2 errors can be made (see below). Decreasing the α levels is therefore self-defeating, because it reinforces the safeguards against a nonexistent error, while at the same time automatically increasing the probability of a Type 2 error (Cohen, 1994; Feise, 2002). As will be seen below, Type 2 errors may be even more serious than Type 1 errors in marine ecology, so increasing their probability is not a desirable outcome.

Additional cogent arguments have been made against adjusting for Type 1 errors when making multiple comparisons, based on the relative risks of not doing so, assuming real null hypotheses are formulated, and the risk of loss of true information by doing so (the 'penalty for peeking'). Even assuming real null hypotheses are formulated, the argument in favor of adjusting for Type 1 errors only applies to random distributions, which is seldom the case when studying living systems (Rothman, 1990a). Applying such adjustments to living systems makes it more difficult to perceive patterns which exist at the heart of data clouded by individual, experimental, or observational variability. The basic debate here is, as often, reasoned judgment vs. the mechanical application of classical procedure based on normal distributions; many non-biologically-trained statisticians are likely to prefer the latter approach, while biologists *should* prefer the former! Proceeding with multiple comparisons, and even adjusting our hypotheses mid-way, is really proceeding as in the alternate, non-classical approaches (Greenland and Robins, 1991; Gelman, 2009 and see below), which appears to be much closer to how the human mind naturally goes about investigating the world (Dienes, 2011).

Beyond the problems associated with misunderstanding, misinterpretation, and misuse of α and P-values, the problems with NHST have been most succinctly and elegantly portrayed (within a literature singularly graced with eloquence), by Stephens et al. (2007). To sum up, NHST is, in most cases, a very inappropriate tool used in very inappropriate ways, to achieve a misinterpreted result. The driving force behind its use is the belief that it is a totally objective, mechanical procedure which will reveal objective truth precisely because we use it in this fashion. Not only is this obviously not the case, but there is no alternative, totally objective, mechanical procedure which will reveal objective truth in any classical approach, as has been eloquently underscored by several statistical luminaries, notably Jacob Cohen (1994). As a consequence, some biomedical journals have not only proscribed the use of P-values, but also any reference to statistical significance. The complete sentence extracted from the instructions to authors in the journal Epidemiology (see above) reads 'We strongly discourage the use of P-values *and language referring to statistical significance*' (http://edmgr.ovid.com/epid/accounts/ifauth.htm)!

The debate about NHST has recently received a great deal of attention with the publication of Ziliak and McCloskey's (2008a) *The cult of statistical significance: how the standard error costs us jobs, justice, and lives* and earlier papers, and the reaction to them from many fields of research (e.g. Hoover and Siegler, 2008; Miettinen, 2009; Spanos, 2008; but see Ziliak and McCloskey,

2008b). As is true throughout the history of statistics, the affective level in these debates often approaches that usually associated with politics or religion.

3.4. What about β?

β is the probability of a Type 2 error. In the ecological sciences, as in most disciplines, we overwhelmingly concentrate on limiting the probability of Type 1 error (α), but rarely that of Type 2 (β). We invite functional ecologists to consider how often they have explicitly incorporated β in their statistical planning and analysis, and to check how often this is done in the papers of the current issue of this journal or other marine biology/ecology journals. Despite the lack of preoccupation with this error source, the consequences of insufficient attention to β may be very important; a striking example in the medical field was given by Streiner (1990), and in the field of marine environmental research, it has been argued that the consequences of a Type 2 error are usually even more serious, and certainly more pernicious, than those of a Type 1 error (Fairweather, 1991; Mapstone, 1995; Peterman, 1990), yet almost all NHST papers in fisheries and aquatic sciences lack any reference to this aspect of statistical data treatment.

As is obvious from the foregoing, Type 1 and 2 errors vary reciprocally, such that decreasing the probability of a Type 1 error automatically (but non-linearly) increases the probability of a Type 2 error. We should also bear in mind that, as was the case for α, there is no intrinsic biological meaning in β (biological vs statistical meaning).

The real problem with β is that it is intrinsically unknowable. Whereas we can determine the probability that a given result may occur by chance (Type 1 error), we cannot determine the probability that it may *not* occur, if the two probabilities are not the only ones possible (Type 2 error). However, this is not a reason to ignore β — we must strive to reduce it, just as we attempt to reduce α. Reducing β is called increasing the *power* of the statistical test, defined as 1 − β (note that this still does not allow us to calculate the value of β). Without increasing α, there are only 2 avenues available for increasing power: reducing the variability of the data (e.g. re-doing the experiment with more efficient instrumentation or methodology, when possible), or increasing sample size (see e.g. Green, 1989 for a discussion of the determination of the necessary *n* to achieve a desired power level for the detection of a given response magnitude in pollution impact studies). Both options are usually associated with increased material costs. However, since β is inversely proportional to √N, relatively large sample size increments translate to much more modest gains in power (reductions in β). Keeping in mind the potential gravity of Type 2 errors in marine ecology, and the difficulty of increasing power by either reducing data variability or increasing sample size, it is therefore clear that in many cases the optimal compromise would be to increase the level of α, e.g. doubling it to 0.1 (and therefore increasing the risk of a Type 1 error), as this will automatically increase statistical power (Peterman, 1990), without incurring any additional material costs. Obviously, such decisions can only be made if we have some knowledge of the relative consequences of Type 1 and 2 errors for each particular study (informed judgment once again).

A power test may be used prospectively, in order to determine the sample size necessary to achieve a desired power level, or retrospectively, in order to calculate the power of the test we effected. By extension of the 'conventional' 0.05 α level, the usual target set for β is 3 or 4 times the α level, or 0.15–0.20. The reasoning for accepting a higher level for β than for α is that increasing power is usually costly (see above), and that in any event, these β levels are a considerable improvement over the majority of studies (Cohen, 1977). Prospectively, a power test is used to determine what sample size is necessary to obtain a β of 0.20; retrospectively, it is used to determine whether the sample size of a study was sufficiently large to achieve the desired β level. In the latter case, it is obviously too late to modify the study if we are not satisfied with the power achieved! Note that here again, the β level may be set higher or lower, depending on the anticipated consequences of a Type 2 error; in marine biology/ecology, these consequences are usually unknown, so the conventional 0.15–0.20 levels may be used, with the understanding that this is merely a convention.

Retrospective power tests may be used, as in the examples above, to determine whether or not our sample size was sufficient to achieve a desired power level. However, retrospective finding of insufficient power must not be used as a justification for deciding that a result is 'inconclusive' due to a small sample size, as this would nearly always guarantee such a finding when sample sizes and effect sizes are small (Nakagawa and Foster, 2004). The correct interpretation is that the study design does not permit any conclusions to be drawn concerning the effect (and this is not a very laudable conclusion!). There are other, more theoretical objections to retrospective power calculations (Hanley, 1994; Smith and Bates, 1992).

A final note on power testing: this procedure, whether used prospectively or retrospectively, relies on three 'judgment-determined' parameters: α, β, and effect size. The potential for (involuntary) experimenter bias is thus thrice that of the most frequent use of NHST, where only the probabilities of Type 1 errors are considered, and this has led several statisticians to declare it 'misleading', in the sense that the perceived risk level of a Type 2 error ascertained by this procedure is negated by the subjectivity or uncertainty involved in fixing the three necessary parameters (Johnson, 1999). However, reasoned attempts to guard against excessive probability of a Type 2 error are surely better than no precaution whatsoever! Since power is exclusively an NHST concern, this important but difficult issue could simply cease to exist if one of the alternate approaches, described below, were adopted (Hanley, 2004).

3.5. The expanding role of confidence intervals (CI)

In the contemporary move toward statistical renewal, heavy emphasis has been placed on an enhanced role for confidence intervals (Cumming and Finch, 2005; Nakagawa and Foster, 2004). Before summarizing the various dimensions of this development, it is necessary to refresh our thinking about data presentation, as it has recently been shown that even this seemingly basic set of concepts is poorly-understood by many leading researchers (Bella et al., 2005). What are loosely referred to as 'error bars' in graphs may be either *ranges*, *standard deviations (SD)*, *standard errors (SE)*, or *confidence intervals (CI)*. Although these terms are taught in introductory statistics classes, it is crucial to understand both the differences in definition *and in purpose* of these statistics. *Ranges* give the extreme upper and lower values of a measured effect. Although they are quite uncommon in marine ecological literature, they should be used more frequently, in situations where N is very small (e.g. ≤5), and it is meaningless to calculate a measure of dispersion about the mean; this type of situation is common when the measurements are extremely costly or difficult to obtain.

The *standard deviation* (SD) is a measure of dispersion of values about the sample mean. It is not correlated with sample size — adding observations does not necessarily reduce SD. Examination of standard deviations on graphs yields no information other than a quantitative appreciation of data variability at each sampling. A common error is to assume that SD's provide information on the proportion of data values within ±a given number of SD, e.g. ±1.96. This is only true for a large sample size characterized by a normal distribution, so no such inferences can be made in any other context. *Where N is patently small*, e.g. *<5, SD, which is a measure of dispersion about the sample mean, has such vanishingly small signification that it should neither be calculated nor presented*, regardless of the fact that most statistical software packages will blindly do so, even with N = 2!

The *standard error of the mean* (SE) relates the variability summarized in the SD to the sample size, i.e. SE = SD/√N, and since it will decrease as sample size increases, its boundaries will move closer to the population mean at the same time. A small SD at a small sample size

allows us to be fairly certain that the population mean lies within a small range of values. Especially at large sample sizes, SE is obviously much smaller than either SD or the half-width of the confidence interval (see below), and hence it is the measure researchers are most tempted to put on graphs of their mean values. In itself, the SE does not allow hypothesis testing or even informative comparisons between treatments. The reasoning given above for the non-use of SD at very small sample sizes obviously also applies to SE.

The *confidence interval* (CI) is a range of values, calculated from the sample observations, that is believed, with a particular probability, to contain the true parameter value. It is commonly estimated as the mean $\pm w$, the *margin of error*, calculated as $w = t_{(n-1),\,C} \cdot SE$, where C is the desired level of confidence (traditionally 95%, but see above), and t is the critical t-value for this C, at $(n-1)$ degrees of freedom. There are other options for CI estimations such as profile likelihood intervals or log-based intervals or bootstrap procedures, which often have better coverage properties, especially when the sampling distribution is non-normal and the CI might be asymmetric (Efron and Tibshirani, 1993; Royall, 1997). It is of the utmost importance to note that *this does not signify that the mean is the true effect, and the CI is the variability of the data about this effect* (this is, unfortunately, one of the most frequent misinterpretations). It signifies that the mean is a *point estimate* of the true effect, and that the corresponding CI is a range of plausible values for the true effect (μ). Values outside the CI are relatively implausible.

There are several important advantages of reporting data as point estimates (e.g. means), accompanied by the corresponding CI (Cumming and Finch, 2005; Stefano et al., 2005). These are:

- o 'Double duty' with NHST. Obviously, values outside a 95% CI correspond to a two-tailed $P \leq 0.05$, if these values reflect the null hypothesis. Conversely, values inside a 95% CI correspond to a two-tailed $P > 0.05$, if these values reflect the null hypothesis. Even more conveniently, *a ≤50% overlap of independent CI bars, which differ in width by ≤ a factor of 2, corresponds to* $P \leq 0.05$ (Cumming, 2009).
- o They are visually informative, compared to simple specification of means and P-values. Worked-through examples may be found in Cumming (2009), Cumming and Finch (2005), and Wolfe and Hanley (2002).
- o The extremities of the CI indicate the extreme possible values of effect size within the specified probability limits.

Editorial recognition of the importance of CI's has prompted some journals to specifically require that all point estimates be reported along with CIs (e.g. Canadian Journal of Psychiatry, http://publications.cpa-apc.org/browse/documents/6).

A specialized use of the CI is the *CI function* (or P-value function), which depicts all possible CIs around a point estimate. This function is especially applicable to meta-analyses, long popular in the medical sciences, and increasingly so in marine ecology (Lang et al., 1999; Sullivan and Foster, 1990).

3.6. A fresh start for Fisher?

Hurlbert and Lombardi (2009) advocate the use of 'Neo-Fisherian Significance Assessments (NFSA)' to overcome the problems of, and replace, the 'paleo-Fisherian and Neyman–Pearsonian paradigms' (i.e. NHST). These authors argue that NFSA more adequately detects the existence, direction, and magnitude of differences in statistical descriptors. They further argue that we should not be overly concerned with the 'bottom of the class' (sic) who have misused NHST to date (*authors' note: and may well continue to do so for NFSA*); rather, we should keep these techniques because they are or can be powerful when used properly. The approach requires a great deal of — once again — reasoned judgment. P-values are reported but no significance level is assigned. Support for rejecting a null hypothesis is presented as a function of the P-value, the power, and the design of the study. Obviously, all of the caveats associated with statistical power, interpretation of P-values, and the meaning of null hypotheses, and their rejection, remain, as does the problem of P-values and researcher intent (see below). NFSA is a modified way of proceeding with classical statistics, which does not require that they be abandoned, but rather that their use and interpretation be made more congruent with the task of presenting and judging evidence. Besides presenting this interesting advantage, NFSA does not require the prior selection of 'models', as is the case with the alternate approaches outlined below. To be sure, in many situations, the information necessary to formulate such models is lacking, so it is important to have at hand an approach which will simply reject a null hypothesis, as well as a carefully-controlled study in which the experimental hypothesis is the only likely alternate under the study conditions. It is early yet to judge whether NFSA will actually 'rise', but studies incorporating this statistical approach have begun to appear in the literature (French et al., 2011). As might have been predicted, since this approach does *not* propose a set, mechanical procedure, different researchers may be more or less rigorous in applying the reasoning; French et al. (2011) use the term 'significance' but not 'significance level', while Stoner (2011) simply avoided choosing an α level without explicitly using the 'reasoned judgment' outlined above.

3.7. Final word of caution: the P-value Achilles heel

When computers calculate exact P-values for a given frequentist statistic, the software makes assumptions about the *intent* of the researcher with respect to data collection. This is not at all a trivial or arcane point. As shown quite lucidly by Kruschke (2010a), the intent of the researcher may alter the critical value of the test statistic very substantially, and hence the probability of obtaining that value, were the experiment or analysis to be repeated many times (the basis for calculating the exact P-value). However, the actual intent of the researcher may not be, and usually is not, the intent assumed by the software. Calculation of CIs is prone to the same problem for the same reason. There are four ways to circumvent this: (1) do not calculate exact probability values (only use a well-informed critical α level and eliminate all reference to 'very significant' differences), (2) wait for software which will ask the pertinent questions in order to ascertain researcher intent concerning data collection, and in the meantime do (1) only, (3) use a variety of classical methods based principally on *effect size*, rather than P-values (ANOVA, regression, correlation), or (4) adopt one of the alternate approaches to investigation and data analysis outlined below.

Before leaving the vast domain of classical statistics, it is useful to recapitulate the most common misconceptions/misapplications encountered in marine biology/ecology, presented in Table 1. In all cases presented in Table 1, suggested remediation is 'brain on, computer off'.

4. Alternate approaches to investigation and data analysis

The alternate methods of providing evidence, increasingly used in the medical and social sciences, and more recently in terrestrial and aquatic ecology, are all based on the comparison of models or model components (parameters), and subsequent selection of the model(s) which is (are) best supported by the data. Selection criteria may be *likelihood* (Likelihood approach), *information content and model complexity* (Information-theoretic approach), or *credibility* (Bayesian approach). Another common point in these approaches is the enhanced role of *informed judgment* in model or parameter selection. Emphasis is placed on the careful a priori definition of a set of candidate models, based on the science of the problem (insofar as it is known at the time of the study). This is conceptually more difficult than estimating the model parameters and their precision, and this is

Table 1
A rogue's gallery of common classical statistical errors.

1) Tests are mechanistic and involve no subjectivity or reasoned judgment. The output is only S/NS
2) Statistical significance = biological significance
3) Minimum $\alpha = 0.05$ for statistical significance (cutoff value)
4) $P > 0.05$ = no effect, or groups are the same/identical with respect to the effect
5) Whenever $P > 0.05$ for an effect, groups can be combined
6) P-value reflects effect size
7) It is necessary to adjust the critical P-value in multiple comparisons
8) H_0: $\mu 1 = \mu 2$, or null hypothesis of no difference, is a meaningful statement
9) Rejecting H_0 affirms the experimental or chosen alternate hypothesis
10) Minimization of β is unimportant or infeasible
11) All 'error bars' give different versions of the same information
12) 95% CIs must not overlap for there to be an α P of ≤ 0.05
13) 95% CIs which touch but do not overlap show an α P of 0.05
14) It is possible to compare error bars on a series of sampling dates (inter-date comparison)
15) As long as the software will calculate a statistic, it can and should be reported
16) Replication of a previously-published study is not worthy of publication

where the deepest available understanding of ecological processes and critical thinking are needed. We refer to these approaches collectively as 'ITLB' (Information theoretic–Likelihood–Bayesian).

The *likelihood function* and the term *likelihood* are used in all three alternate approaches, so it is useful to define them here, and especially to differentiate likelihood and probability (P-values). The *likelihood function* is the set of probabilities for various outcomes, given specified parameter values. *Likelihood* is defined as the probability of obtaining the exact data observed, D, given the hypothesis (outcome) being considered [P(D|H)]. Likelihoods are values which correspond to the height of the probability distribution at a particular point, whereas P-values are areas of the probability distribution.

4.1. Likelihood analysis

In the likelihood approach, different models, including the null hypothesis, can be compared according to their likelihood (Edwards, 1992; Royall, 1997). One form of the approach has been hybridized to fit the classical type of statistical analysis, the *likelihood ratio* test, in which we calculate how many times more likely the data are under one model than the other i.e. the null model (hypothesis) vs the alternate model (hypothesis). P-values and critical levels for rejection of the null model may be used, but are not necessarily part of likelihood analysis.

Although P-values and critical levels have the superficial advantage of reassuring the user that he or she is operating within the 'secure' perimeter of classical statistics, decisions based on likelihood ratios alone are actually more well-informed (Pernerger and Courvoisier, 2010). The graded estimation of likelihood, given several possible models, each with their own likelihood ratios, more closely resembles many real-world situations, and contrasts with the (at best) 'mechanical razor', black-or-white approach typically taken in the use of classical statistics. It has the immense advantage of being intuitive; correct interpretation of likelihood analyses, and consequent decision-making, is much more probable, even for students not trained in the technique (Pernerger and Courvoisier, 2010). The understanding and interpretation of statistical tests and procedures are obviously the foundation of the entire statistical enterprise, and it is to be hoped that research in the recent field of statistical cognition will play a major role in the optimization of 'cognition-friendly' statistical procedure (Beyth-Marom et al., 2008; Cumming et al., 2004).

4.2. Information-theoretic analysis

Ecologists frequently adapt and adopt concepts from information-thermodynamic theory, as tools to quantify essential characteristics of complex systems such as ecosystems; perhaps the best-known is the archetypical biodiversity measure, the Shannon–Wiener index. The information-theoretic approach to data treatment (Anderson, 2008; Burnham and Anderson, 2002) is an integrated process of a priori specification of a set of candidate models (based on the science of the problem), model selection based on the principle of parsimony, and the estimation of parameters and their precision. The principle of parsimony implies the selection of a model with the smallest possible number of parameters for adequate representation of the data, i.e. a trade-off between model fit (likelihood) and model complexity. The most commonly-used measure to apply the principle of parsimony is Akaike's Information Criterion (AIC — Akaike, 1973).

Under the information-theoretic approach, it is not assumed that truth is included in the set of candidate models and the issue is not which model is true, but rather which model, when fitted to the data, is the one which best represents the finite information contained in the data. The concept of a 'true' model seems to be of little utility in marine ecology, as biological systems are quite complex with many small effects (tapering effects), individual heterogeneity, and interactions that are generally unknown. In the information-theoretic approach, 'information' about the biological system under study is assumed to exist in the data, and the goal is to express this information in a coherent and compact way, which may then be interpreted in the light of whatever other relevant information also exists. There is, of course, no predetermined cut-off level for acceptance or rejection of hypotheses.

Since larger data sets usually contain more information, more complicated models may be supported by larger data sets. The information-theoretic approach allows formal inference to be based on several or even all the candidate models rather than on only the 'best' model. This procedure is termed multi-model inference (MMI) and has several theoretical and practical advantages (Burnham and Anderson, 2002; Katsanevakis, 2006).

One of the earliest uses of this approach in marine biology was in a series of papers, which quietly revolutionized the modeling of growth and allometry in marine ectotherms (Katsanevakis, 2006; Katsanevakis and Maravelias, 2008; Katsanevakis et al., 2007a), and also contained concise explanations of information theory and its particular relevance to growth and allometric modeling. This approach has since been extended to other species (e.g. Griffiths et al., 2010; Harry et al., 2011; Lin and Tzeng, 2009; Mercier et al., 2011; Rabaoui et al., 2007, 2011; Yokoyama and Amaral, 2011). The information-theoretic approach has also been applied to other areas of marine biology and ecology, such as respiration studies (Katsanevakis et al., 2007b), investigations of the effect of exploitation pattern on the status of fish stocks (Vasilakopoulos et al., 2011), investigations of stock-recruitment relationships of fish (Galindo-Cortes et al., 2010), spatial distribution and habitat use (Katsanevakis et al., 2010), modeling detectability in underwater visual surveys (Katsanevakis and Thessalou-Legaki, 2007; Katsanevakis et al., 2011), and estimating occupancy patterns of marine species (Katsanevakis et al., 2011). Advantages and caveats concerning the use of Information Theory have been succinctly outlined in Anderson et al. (2001); it is clear that this is a promising tool for analyzing and understanding data, not only in experimental work, but also in observational studies, where the classical hypothesis-testing approaches seem to have no theoretical justification and often perform poorly (Burnham and Anderson, 2002).

4.3. Bayesian analysis

The Bayesian approach to investigation is based on the successive re-allocation of *credibility*. Credibility is defined as *the likelihood of models or explanations which are repeatedly modified as increasing amounts of information are gathered about the models*. Some models may be weakened and rejected in this process (diminished

credibility), and others may be strengthened and eventually adopted (enhanced credibility). Bayesian analysis goes against the grain of classical statistics because, obviously, the term 'credibility' is anathema to those who have been taught that science (and in particular statistics) has nothing to do with belief. The epistemological foundation for this approach is the recognition that natural phenomena are inherently complex, and that considering multiple variables (i.e. multiple models or explanations) brings us closer to the truth than just analyzing isolated variables. Furthermore, whether they will admit it or not, scientists engage in credibility evaluation of their (and other's) data virtually every day, consciously or unconsciously. And finally, the term 'credibility' can be replaced with the more reassuring term 'factual congruency', where congruency is contingent upon observation or experimentation; some authors simply use the term 'likelihood' (Dienes, 2011), while others use the inverse: greater or lesser 'doubt' as a function of the evidence presented (Goodman, 2001).

In the biological world, Bayesian techniques are familiar to all who have done, or read papers about, cladistic phylogeny (e.g. Dufour et al., 2006; Mikkelson et al., 2006). The same process of successive strengthening of belief or factual congruency can be applied to ecological problems, and it is often far more appropriate than classical approaches (Dienes, 2011). It can be applied to many different types of ecological problems, from hypothesis comparisons to determination of development times in natural planktonic populations (Gould and Kimmerer, 2010), spatial distribution (e.g. Palmer et al., 2011), fisheries stock assessment (Jiao et al., 2011; Punt and Hilborn, 1997), and, most recently, trophic ecology (Beninger et al., 2011; Moore and Semmens, 2008). Bayesian analysis requires a starting-point, or *prior hypothesis*, characterized by a probability distribution, which is then either strengthened or weakened by addition of data. Although it is often said that Bayesian 'priors' need not be especially likely or even informative, their distributions must be appropriate for the hypothesis (Christensen, 2005). A simplified procedure, using the classical null hypothesis as the minimum Bayes factor (the change in probability of the hypothesis), has been proposed, and although it suffers from the same confounding of effect size and probability of occurrence as NHST, it circumvents most of the other NHST problems (Goodman, 2001). It should be viewed as an attempt to lead classically-trained researchers to alternate approaches via a stepping-stone, although those workers who really contemplate testing the Bayesian waters are probably not the ones who will avail themselves of such a prop.

5. Classical and alternate approaches: informed use is critical

Different methods of statistical inference (i.e. classical and alternative approaches) can give quantitatively and qualitatively different results (Gerrodette, 2011). In marine conservation and management this can have serious implications, as the requirement to disprove a null hypothesis of no effect (or no impact or no decline) can lead to a non-precautionary or non-action attitude. Against the backdrop of vociferous reformation, counter-reformation, and cognitive inertia, several recent papers have called for the obvious: better understanding of the foundations and limitations of both classical and alternative approaches to data treatment, and informed use of one or the other, or, despite some objections, both in concert (Stephens et al., 2005).

Although classical and ITLB approaches differ fundamentally in philosophy and method, they have a common denominator: both rely heavily on a priori hypothesis formulation. In the case of classical statistics, this is due to the desire for an ostensibly 'arms-length' relation to the data, whereas in ITLB it is due to the necessity of a scientifically justified starting-point. In both approaches, however, the data are not intended to be repeatedly re-analyzed to reveal unsuspected proximities or differences, as in principal component analysis and all post-hoc methods. However, since our understanding of biological phenomena is so patently limited, *post-hoc* comparisons can be an extremely useful tool for discovering new relationships and proposing new hypotheses — which may help to inform the starting and competing models in ITLB.

The constant, and insidiously tempting danger of excess is 'data dredging' (Classical) or 'model dredging' (ITLB), where we have no idea at all about the phenomena, and attempt, *a posteriori*, to 'discover' any seemingly plausible scenario to either fit the statistical results (Classical) or serve as models (ITLB) (Stephens et al., 2005). This practice has been called HARKing: Hypothesizing After the Results are Known, and has been shown to be quite widespread in most disciplines (Kerr, 1998). Although no quantitative study is available on its prevalence in marine ecology, personal contact with many marine scientists over several decades indicates that this field is no exception.

Notwithstanding the foregoing, scientists whose work is at the extreme edge of their scientific field (and which therefore precedes human knowledge) are all aware of the converse danger of insistence on a priori hypothesis formulation: our limited, or inexistent, understanding of the phenomena may prevent us from formulating hypotheses congruent with the underlying reality. In these situations, scientists must indeed HARK, and although a seemingly preponderant negative effect of HARKing may be concluded by counting the number of HARKing studies (Kerr, 1998), a most emphatic exception must be made for studies beyond which there is only a vast unknown. Once again, we see that proper statistical treatment is really about informed judgment and constant self-examination and criticism.

6. The crucial role of editors and reviewers

6.1. Study replication

In both the classical and the ITLB approaches, many authors who have detected statistically-significant differences or who have substantially supported models, are tempted to conclude that their results are highly unlikely to be due to anything other than the proposed alternative hypothesis or best model. Readers of such papers are also tempted to conclude that repeating the experiment is both useless, and a waste of resources. Nothing could be further from the truth. As noted by Kruschke (2010b), even statistically-significant findings with a high power in the classical approach do not give any indication of repeatability of the results. This point was most forcefully argued by Johnson (1999), who concluded that *the only path to increased certainty is through true study replication* (and even then, this does not *guarantee* that the proposed alternative hypothesis is true, since we could obtain the same results consistently for reasons other than the alternative hypothesis). Here, the responsibility of editors and referees is crucial: *replicative studies should be welcomed, not discouraged as being 'repetitive'*, as per current journal practice. The counter-argument from editors, that there is already not enough room for original papers, does not change the necessity for such studies, if approaching truth is the goal of these journals. The systemic reward for originality, as well as the costs and often the practical impossibility of replication in marine biology/ecology (e.g. studies of seasonal effects — it is not possible to control climatic conditions), while perfectly justifiable, are perhaps the greatest barriers to replication.

6.2. Systematic publication bias

Another serious problem arising from unofficial journal/referee policy is the practice of rejecting manuscripts which report non-statistically significant results (the 'file-drawer effect' — Carver, 1978; Rosenthal, 1979). The consequence is, of course, an over-representation of studies which report statistically-significant results concerning a given question or phenomenon (publication bias), which may even lead to the reporting of effects which do not exist

(e.g. if only a small minority of studies, published, were able to detect an effect, while the vast majority, non-published, could not).

Publication bias has been known for decades, has most recently been highlighted in the journal Nature (Sarewitz, 2012), and is even described in the popular Web site Wikipedia (http://en.wikipedia.org/wiki/Publication_bias). Besides the inevitable consequent misrepresentation of reality, such bias will be immeasurably amplified in the increasingly popular 'meta-analyses' which dot the marine ecological landscape. On the other hand, since there are many more negative or non-statistically-significant possible results than there are positive ones (see Comments in Sarewitz, 2012), editors must somehow select, without bias, studies which bring something new to the story, in addition to a negative or non-significant result. This requires a very vast and deep understanding of the fields in which they specialize.

Detection of publication bias in a particular field can be achieved using funnel-plot asymmetry (Dubben and Beck-Bornholdt, 2005; Egger et al., 1997), or, once again, a Bayesian modeling approach (Givens et al., 1997). Editors should welcome submission of high-quality studies of this type for the fields covered by their journals (e.g., Song et al., 2010).

6.3. Statistical overkill

In the same vein as publication bias, the practice of 'statistical overkill' is equally unfortunate. Under pressure from peers, reviewers and editors, researchers often feel compelled to 'pump up' the statistical treatment and presentation of their results. In many cases, the resulting ill-founded, poorly-chosen, inadequately presented statistical treatment not only adds nothing to the original data, it actually obscures and detracts from it (Beninger et al., 2010). *Researchers, especially those who report the results of observational studies or of field work in which some components could not be randomized, should not feel obligated to embellish their data with statistical treatment developed for randomized, experimental studies working within a strong theoretical framework*. Their studies should be recognized by reviewers and editors as perfectly legitimate and publishable, if within the scope of the journal and of a sufficiently high caliber.

6.4. The roles of statisticians

Statisticians often distinguish between themselves and 'subject specialists', with whom they are occasionally invited to collaborate. While statisticians argue convincingly for their implication at the earliest stages of any project (Murray, 1988; Strasak et al., 2007), this is not always possible, due to a shortage of statisticians willing to engage in such enterprises, and also to the problem of discipline insularity, or lack of sufficient training in each other's field to allow effective collaboration.

Whether or not statisticians are involved in projects at an early stage, or not at all, quality control of data treatment is as crucial as for all other aspects of a manuscript. Editors of scientific journals are the ultimate 'gatekeepers' of newly-generated knowledge, and as such they should have the necessary mechanisms at their disposal to impose a final, and sometimes only, quality control on statistical treatment of data. In concrete terms, this means the association of statistical consultants with the editorial team.

Although to our knowledge, no marine biology/ecology journal currently includes statistical consultants on their editorial staff, the medical sciences have a fairly long track record of such practice. As early as 1990, the Canadian Journal of Psychiatry referred to their regular statistical consultant in an editorial on research methods (Bland, 1990), who also published a paper on sample size and power in the same issue (Streiner, 1990). Whereas in 1981, the proportion of manuscripts reviewed by statistical consultants at any stage after submission was less than 15% (George, 1985), in 1995, 52% of the manuscripts submitted to medical journals in the top-ranking quartile were so reviewed, versus 27% in the bottom-ranking quartile (Goodman et al., 1998). Indeed, the top-ranking medical journals commonly maintain several such consultants (e.g. four at the New England Journal of Medicine). The probability of having a staff statistical consultant on the editorial boards of the top 25%-ranked medical journals was 82% in 1995 (Goodman et al., 1998), and is probably close to 100% today. Obviously, within the medical specialties, there are fields which rely much more heavily on statistics than others, e.g. epidemiology on the high end, and dermatology on the low end, and this is reflected in the statistical review practices of their specialty journals (Katz et al., 2004).

The overall positive effect of statistical review on manuscript quality has been demonstrated (Altman, 1998). On the other hand, the negative effect of ill-informed, yet frequent, statistical critique by 'subject specialists' has been deplored (Bacchetti, 2002). These observations lead to the conclusion that at the very least, whenever statistical treatment is criticized by a 'subject specialist', the manuscript should be referred to a statistical consultant. An example of a statistical reviewing checklist is presented in Houle and Penzien (2009), and although it is not totally without reproach, it does provide a starting point for journals not yet up to speed.

The role of marine biology/ecology journal editors, acting through a policy of statistical review, is thus absolutely crucial to the improvement of data treatment in manuscripts submitted, or even in preparation. Merely stating that researchers should strengthen statistical procedure (as we do here), while necessary, is not sufficient to effect real change; stricter editorial policy and clearer author guidelines are needed (Fidler et al., 2005). Indeed, a statement that manuscripts will be or even *may be* subject to review by a staff statistician (e.g. American Journal of Kidney Disease, www.ajkd.org/content/edpolicies), will probably in itself result in greater care during data treatment, manuscript preparation, and probably also in project planning. This is all the more paramount in view of the fact that most submitted manuscripts are eventually published *somewhere*. Although we are unaware of a comparative bibliometric study in marine biology/ecology, in the medical field it has been determined that only 15% of submitted manuscripts remained inactive following rejection; the rest were either published elsewhere (75%), under review elsewhere (3%), or being prepared for re-submission elsewhere (7%) (Hall and Wilcox, 2007).

7. Conclusion

Just as the techniques of chemical analyses or nucleic acid sequencing constantly improve, so do the techniques of study design and statistical treatment of data. While the nature and context of ecological data are often quite different from those of the medical or social sciences, in which numerous, seemingly esoteric techniques abound, the basic approaches to experimental data have indeed evolved quite significantly in the past few decades, greatly expanding the biologist's statistical toolbox. Beyond the burgeoning additions and refinements to classical statistics which fill the pages of the major statistical journals, the increasing use of ITBL approaches has opened promising new paths in our progress toward more complete phenomenological understanding, and rather than either 'going with the flow' (inertia) or trying to be at the 'cutting edge' (fashionista), biologists should strive to use the appropriate statistical tool for each project (Anderson et al., 2001; Stephens et al., 2005). For this to become more reflexive, it is obvious that the foregoing considerations must be more widely disseminated in university biology/ecology curricula at the undergraduate level. This means that, in most instances, at least three, and probably four, semester statistics courses, taught by biologically-familiar statisticians, will be required in order to achieve a comprehensible introduction, and hopefully the curiosity and motivation to begin selecting the best tools for each job.

Acknowledgments

We are grateful to two anonymous reviewers for comments, and to S.E. Shumway for very fruitful discussions. IB was supported by a PhD scholarship from the French Ministère de l'enseignement supérieur et de la recherche, and this work emerged in the course of a project funded by the Région Pays de la Loire. **[SS]**

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 267–281.

Altman, D.G., 1998. Statistical reviewing for medical journals. Stat. Med. 17, 2661–2674.

Anderson, D.R., 2008. Model Based Inference in the Life Sciences: A Primer on Evidence. Springer Science + Business Media, New York.

Anderson, D.R., Link, W.A., Johnson, D.H., Burnham, K.P., 2001. Suggestions for presenting the results of data analyses. J. Wildl. Manage. 65, 373–378.

Bacchetti, P., 2002. Peer review of statistics in medical research: the other problem. Br. Med. J. 324, 1271–1273.

Bella, S., Fidler, F., Williams, J., Cumming, G., 2005. Researchers misunderstand confidence intervals and standard error bars. Psychol. Method. 10, 389–396.

Beninger, P.G., Potter, T.M., St-Jean, S., 1995a. Paddle cilia fixation artefacts in pallial organs of adult *Mytilus edulis* and *Placopecten magellanicus* (Mollusca, Bivalvia). Can. J. Zool. 73, 610–614.

Beninger, P.G., St-Jean, S., Poussart, Y., 1995b. Labial palps of the blue mussel *Mytilus edulis* (Bivalvia: Mytilidae). Mar. Biol. 123, 293–303.

Beninger, P.G., Valdizan, A., Decottignies, P., Cognie, B., 2010. Field reproductive dynamics of the invasive slipper limpet, *Crepidula fornicata*. J. Exp. Mar. Biol. Ecol. 390, 179–187.

Beninger, P.G., Elner, R.W., Morançais, M., Decottignies, P., 2011. Downward trophic shift during breeding migration in the shorebird *Calidris mauri* (western sandpiper). Mar. Ecol. Prog. Ser. 428, 259–269.

Berger, J.O., Berry, D.A., 1988. Statistical analysis and the illusion of objectivity. Am. Sci. 76, 159–165.

Berkson, J., 2003. Tests of significance considered as evidence. Int. J. Epidemiol. 32, 687–691.

Beyth-Marom, R., Fidler, F., Cumming, G., 2008. Statistical cognition: towards evidence-based practice in statistics and statistics education. Stat. Educ. Res. J. 7, 20–39.

Bland, R.C., 1990. Research methods in psychiatry. Can. J. Psychiatry 35, 614–615.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference: A Pratical Information-theoretic Approach, second ed. Springer-Verlag, New York.

Carver, R.P., 1978. The case against statistical significance testing. Harv. Educ. Rev. 48, 378–399.

Christensen, R., 2005. Testing Fisher, Neyman, Pearson, and Bayes. Am. Stat. 59, 121–126.

Cohen, J., 1977. Statistical Power Analysis for the Social Sciences. Academic Press, New York.

Cohen, J., 1994. The Earth is round ($p<.05$). Am. Psychol. 49, 997–1003.

Cook, R.J., Farewell, V.T., 1996. Multiplicity considerations in the design and analysis of clinical trials. J. R. Stat. Soc. A 159, 93–110.

Cumming, G., 2009. Inference by eye: reading the overlap of independent confidence intervals. Stat. Med. 28, 205–220.

Cumming, G., Finch, S., 2005. Inference by eye: confidence intervals and how to read pictures of data. Am. Psychol. 60, 170–180.

Cumming, G., Williams, J., Fidler, F., 2004. Replication, and researchers' understanding of confidence intervals and standard error bars. Underst. Stat. 3, 299–311.

Dienes, Z., 2011. Bayesian vs orthodox statistics — which side are you on? Perspect. Psychol. Sci. 6, 274–290.

Dubben, H.H., Beck-Bornholdt, H.P., 2005. Systematic review of publication bias in studies on publication bias. Br. Med. J. 331, 433–434.

Dufour, S.C., Steiner, G., Beninger, P.G., 2006. Phylogenetic analysis of the perihydrothermal vent bivalve *Bathypecten vulcani* based on 18s rRNA. Malacologia 48, 35–42.

Edwards, A.W.F., 1992. Likelihood. expanded edition Johns Hopkins University Press, Baltimore, MD.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall, New York.

Egger, M., Smith, G.D., Schneider, M., Minder, C., 1997. Bias in meta-analysis detected by a simple, graphical test. Br. Med. J. 315, 629–634.

Fairweather, P.G., 1991. Statistical power and design requirements for environmental monitoring. Aust. J. Mar. Fresh. Res. 42, 555–567.

Feise, R.J., 2002. Do multiple outcome measures require p-value adjustment? BMC Med. Res. Methodol. 2, 8.

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., Edmonds, H., Harrington, C., Schmitt, R., 2005. Evaluating the effectiveness of editorial policy to improve statistical practice: the case of the *Journal of Consulting and Clinical Psychology*. J. Consult. Clin. Psychol. 73, 136–143.

Fidler, F., Burgman, M.A., Cumming, G., Buttrose, R., Thomason, N., 2006. Impact of criticism of null hypothesis significance testing on statistical reporting practices in conservation biology. Conserv. Biol. 20, 1539–1544.

Fisher, R.A., 1959. Statistical Methods and Scientific Inference, second ed. Oliver and Boyd, Edinburgh.

French, S.F., González-Suárez, M., Julie, K., Young, J.K., Durham, S., Gerber, L.R., 2011. Human disturbance influences reproductive success and growth rate in California sea lions (*Zalophus californianus*). PLoS One 6 (3), e17686.

Galindo-Cortes, G., De Anda-Montanez, J.A., Arreguín-Sánchez, F., Salas, S., Balarta, E.F., 2010. How do environmental factors affect the stock–recruitment relationship? The case of the Pacific sardine (*Sardinops sagax*) of the northeastern Pacific Ocean. Fish. Res. 102, 173–183.

Garrett, K.A., 1997. Use of statistical tests of equivalence (bioequivalence tests) in plant pathology. Phytopathology 87, 372–374.

Gelman, A., 2009. Why we (usually) don't have to worry about multiple comparisons. CPRC Working Paper No. 09-12, 1–31.

Gelman, A., Stern, H., 2006. The difference between 'significant' and 'not significant' is not itself statistically significant. Am. Stat. 60, 328–331.

George, S.L., 1985. Statistics in medical journals: a survey of current policies and proposals for editors. Med. Pediatr. Oncol. 13, 109–112.

Germano, J.D., 1999. Ecology, statistics, and the art of misdiagnosis: the need for a paradigm shift. Environ. Rev. 7, 167–190.

Gerrodette, T., 2011. Inference without significance: measuring support for hypotheses rather than rejecting them. Mar. Ecol. 32, 404–418.

Gigerenzer, G., 2004. Mindless statistics. J. Socio-Econom. 33, 587–606.

Gigerenzer, G., Krauss, S., Vitouch, O., 2004. The null ritual: what you always wanted to know about significance testing but were afraid to ask. In: Kaplan, D. (Ed.), The Sage Handbook of Quantitative Methodology for the Social Sciences, Sage Ltd., Thousand Oaks, CA, pp. 391–408.

Givens, G.H., Smith, D.D., Tweedie, R.L., 1997. Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. Stat. Sci. 12, 221–250.

Goodman, S.N., 1999. Toward evidence-based medical statistics. 1: the P value fallacy. Ann. Intern. Med. 130, 995–1004.

Goodman, S.N., 2001. Of P-values and Bayes: a modest proposal. Epidemiology 12, 295–297.

Goodman, S.N., 2008. A dirty dozen: twelve p-value misconceptions. Semin. Hematol. 45, 135–140.

Goodman, S.N., Altman, D.G., George, S.L., 1998. Statistical reviewing policies of medical journals: caveat lector? J. Gen. Intern. Med. 13, 753–756.

Gould, A.L., Kimmerer, W.J., 2010. Development, growth, and reproduction of the cyclopoid copepod *Limnoithona tetraspina* in the upper San Francisco Estuary. Mar. Ecol. Prog. Ser. 412, 163–177.

Green, R.H., 1979. Sampling Design and Statistical Methods for Environmental Biologists. Wiley, Chichester.

Green, R.H., 1989. Power analysis and practical strategies for environmental monitoring. Environ. Res. 50, 195–205.

Greenland, S., 1990. Randomization, statistics, and causal inference. Epidemiology 1, 421–429.

Greenland, S., Robins, J.M., 1991. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. Epidemiology 2, 244–251.

Griffiths, S.P., Fry, G.C., Manson, F.J., Lou, D.C., 2010. Age and growth of longtail tuna (*Thunnus tonggol*) in tropical and temperate waters of the central Indo-Pacific. ICES J. Mar. Sci. 67, 125–134.

Hall, S.A., Wilcox, A.J., 2007. The fate of epidemiologic manuscripts: a study of papers submitted to Epidemiology. Epidemiology 18, 262–265.

Hanley, J.A., 2004. Confidence limits vs power calculations. Epidemiology 5, 264–266.

Harry, A.V., Macbeth, W.G., Gutteridge, A.N., Simpfendorfer, C.A., 2011. The life histories of endangered hammerhead sharks (Carcharhiniformes, Sphyrnidae) from the east coast of Australia. J. Fish Biol. 78, 2026–2051.

Hoover, K.D., Siegler, M.V., 2008. The rhetoric of 'signifying nothing': a rejoinder to Ziliak and McCloskey. J. Econ. Methodol. 15, 57–68.

Houle, T.B., Penzien, D.B., 2009. Statistical reviewing for *Headache*. Headache 49, 159–161.

Hubbard, R., Bayarri, M.J., 2003. Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. Am. Stat. 57, 171–182.

Hubbard, R., Bayarri, M.J., 2005. Christensen, R. 'Testing Fisher, Neyman, Pearson, and Bayes', in The American Statistician 59: 121–126: comment by Hubbard and Bayarri and response. Am. Stat. 59, 353.

Huck, S.W., 2011. Reading Statistics and Research. Addison Wesley Longman, Boston.

Hurlbert, S.H., 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 54, 187–211.

Hurlbert, S.H., Lombardi, C.M., 2009. Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. Ann. Zool. Fennici. 46, 311–349.

Jiao, Y., Cortés, E., Andrews, K., Guo, F., 2011. Poor-data and data-poor species stock assessment using a Bayesian hierarchical approach. Ecol. Appl. 21, 2691–2708.

Johnson, D.H., 1999. The insignificance of statistical significance testing. J. Wildl. Manage. 63, 763–772.

Katsanevakis, S., 2006. Modelling fish growth: model selection, multi-model inference and model selection uncertainty. Fish. Res. 81, 229–235.

Katsanevakis, S., Maravelias, C.D., 2008. Modelling fish growth: multi-model inference as a better alternative to *a priori* using von Bertalanffy equation. Fish Fish. 9, 178–187.

Katsanevakis, S., Thessalou-Legaki, M., 2007. First record of *Alicia mirabilis* (Anthozoa: Actiniaria) from the Aegean Sea and density assessment with distance sampling in a site of high abundance. Mar. Biol. Res. 3, 468–472.

Katsanevakis, S., Thessalou-Legaki, M., Karlou-Riga, C., Lefkaditou, E., Dimitriou, E., Verriopoulos, G., 2007a. Information-theory approach to allometric growth of marine organisms. Mar. Biol. 151, 949–959.

Katsanevakis, S., Xanthopoulos, J., Protopapas, N., Verriopoulos, G., 2007b. Oxygen consumption of the semi-terrestrial crab *Pachygrapsus marmoratus* in relation to body mass and temperature: an information theory approach. Mar. Biol. 151, 343–352.
Katsanevakis, S., Salomidi, M., Panou, A., 2010. Modelling distribution patterns and habitat preference of the invasive green alga *Caulerpa racemosa* in the Saronikos Gulf (Eastern Mediterranean). Aquat. Biol. 10, 57–67.
Katsanevakis, S., Zenetos, A., Mačić, V., Beqiraj, S., Poursanidis, D., Kashta, L., 2011. Invading the Adriatic: spatial patterns of marine alien species across the Ionian–Adriatic boundary. Aquat. Biol. 13, 107–118.
Katz, K.A., Crawford, G.H., Lu, D.W., Kantor, J., Margolis, D.J., 2004. Statistical reviewing policies in dermatology journals: results of a questionnaire survey of editors. J. Am. Acad. Dermatol. 51, 234–240.
Kerr, N.K., 1998. HARKing: hypothesizing after the results are known. Pers. Soc. Psychol. Rev. 2, 196–217.
Kruschke, J.K., 2010a. Bayesian data analysis. Wiley Interdiscip. Rev. Cogn. Sci. 1, 658–676.
Kruschke, J.K., 2010b. What to believe: Bayesian methods for data analysis. Trends Cogn. Sci. 14, 293–300.
Lang, J.M., Rothman, K.J., Cann, C.L., 1998. That confounded P-value. Epidemiology 9, 7–8.
Lang, J.M., Rothman, K.J., Cann, C.L., 1999. The P-value and P-value function. Epidemiology 10, 345–346.
Lin, Y.J., Tzeng, W.N., 2009. Modelling the growth of Japanese eel *Anguilla japonica* in the lower reach of the Kao-Ping River, southern Taiwan: an information theory approach. J. Fish Biol. 75, 100–112.
Mapstone, B.D., 1995. Scalable decision criteria in environmental impact assessment: effect size, type I, and type II errors. In: Schmitt, R.J., Osenberg, C.W. (Eds.), Detecting Ecological Impacts: Concepts and Applications in Coastal Habitats. Academic Press, New York, pp. 67–80.
Martínez-Abraín, A., 2007. Are there any differences? A non-sensical question in ecology. Acta Oecol. 32, 203–206.
Mercier, L., Panfili, J., Paillon, C., Ndiaye, A., Mouillot, D., Darnaude, A.M., 2011. Otolith reading and multi-model inference for improved estimation of age and growth in the gilthead sea bream *Sparus aurata* (L.). Estuarine Coastal Shelf Sci. 92, 534–545.
Miettinen, O.S., 2009. Book review: Ziliak S T, McCloskey D N. The cult of statistical significance: how the standard error costs us jobs, justice, and lives. Eur. J. Epidemiol. 24, 111–114.
Mikkelson, P.M., Bieler, R., Kappner, I., Rawlings, T.A., 2006. Phylogeny of Veneroidea (Mollusca: Bivalvia) based on morphology and molecules. Zool. J. Linn. Soc. 148, 439–521.
Moore, J.W., Semmens, B.X., 2008. Incorporating uncertainty and prior information into stable isotope mixing models. Ecol. Lett. 11 (470), 480.
Morrison, D.E., Henkel, R.E., 1970. The Significance Test Controversy — A Reader. Aldine Publishing Co, Chicago.
Murray, G.D., 1988. The task of a statistical referee. Br. J. Surg. 75, 664–667.
Nakagawa, S., Foster, T., 2004. The case against retrospective statistical power analyses with an introduction to power analysis. Acta. Ethol. 7, 103–108.
Palmer, M., Balle, S., March, D., Alós, J., Linde, M., 2011. Size estimation of circular home range from fish mark-release-(single)-recapture data: case study of a small labrid targeted by recreational fishing. Mar. Ecol. Prog. Ser. 430, 87–97.
Pernerger, T.V., 1998. What's wrong with Bonferroni adjustments. Br. Med. J. 316, 1236–1238.
Pernerger, T.V., Courvoisier, D.S., 2010. Interpretation of evidence in data by untrained medical students: a scenario-based study. BMC Med. Res. Methodol. 10, 78.
Peterman, R., 1990. Statistical power analysis can improve fisheries research and management. Can. J. Fish. Aquat. Sci. 47, 2–15.
Peterson, C.H., McDonald, L.L., Green, R.H., Erickson, W.P., 2001. Sampling design begets conclusions: the statistical basis for detection of injury to and recovery of shoreline communities after the 'Exxon Valdez' oil spill. Mar. Ecol. Prog. Ser. 210, 255–283.
Poole, C., 2001. Low P-values or narrow confidence intervals: which are more durable? Epidemiology 12, 291–294.
Punt, A.E., Hilborn, R., 1997. Fisheries stock assessment and decision analysis: the Bayesian approach. Rev. Fish Biol. Fish. 7, 35–63.
Rabaoui, L., Tlig-Zouari, S., Katsanevakis, S., Ben Hassine, O.K., 2007. Comparison of absolute and relative growth patterns among five *Pinna nobilis* populations along the Tunisian coastline: an information theory approach. Mar. Biol. 152, 537–548.
Rabaoui, L., Tlig-Zouari, S., Katsanevakis, S., Belgacem, W., Ben Hassine, O.K., 2011. Differences in absolute and relative growth between two shell forms of *Pinna nobilis* (Mollusca: Bivalvia) along the Tunisian coastline. J. Sea Res. 66, 95–103.
Rosenthal, R., 1979. The "File Drawer Problem" and the tolerance for null results. Psychol. Bull. 86, 638–641.
Rothman, K.J., 1990a. No adjustments are needed for multiple comparisons. Epidemiology 1, 43–46.
Rothman, K.J., 1990b. Statistics in nonrandomized studies. Epidemiology 1, 417–418.
Royall, R., 1997. Statistical Evidence — A Likelihood Paradigm. Champan & Hall, London.
Sarewitz, D., 2012. Beware the creeping cracks of bias. Nature 485, 149.
Sellke, T., Bayarri, M.J., Berger, J.O., 2001. Calibration of P-values for testing precise null hypotheses. Am. Stat. 55, 62–71.
Silva-Ayçaguer, L.C., Suárez-Gil, P., Fernández-Somoano, A., 2010. The null hypothesis significance test in health sciences research (1995–2006): statistical analysis and interpretation. BMC Med. Res. Methodol. 10, 1–9.
Smith, A.H., Bates, M.N., 1992. Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. Epidemiology 3, 449–452.
Song, F., Parekh, S., Hooper, L., Loke, Y.K., Ryder, J., Sutton, A.J., Hing, C., Kwok, C.S., Pang, C., Harvey, I., 2010. Dissemination and publication of research findings: an updated review of related biases. Health Technol. Assess. 14, 1–236.
Spanos, A., 2008. Review of S. T. Ziliak and D. N. McCloskey's, The cult of statistical significance. Eras. J. Phil. Econ. 1, 154–164.
Stang, A., Poole, C., Kuss, O., 2010. The ongoing tyranny of statistical significance testing in biomedical research. Eur. J. Epidemiol. 25, 225–230.
Stefano, J.D., Fidler, F., Cumming, G., 2005. Effect size estimates and confidence intervals: an alternative focus for the presentation and interpretation of ecological data. In: Burke, A.R. (Ed.), New Trends in Ecology Research. Nova Science Publishers, New York, pp. 71–102.
Stephens, P.A., Buskirk, S.W., Hayward, G.D., Del Rio, C.M., 2005. Information theory and hypothesis testing: a call for pluralism. J. Appl. Ecol. 42, 4–12.
Stephens, P.A., Buskirk, S.W., Hayward, G.D., Del Rio, C.M., 2007. Inference in ecology and evolution. Trends Ecol. Evol. 22, 192–197.
Sterne, J.A.C., Smith, G.D., 2001. Br. Med. J. 322, 226–231.
Stewart-Oaten, A., 1995. Rules and judgments in statistics: three examples. Ecology 76, 2001–2009.
Stoner, D.C., 2011. Ecology and conservation of cougars in the Eastern Great Basin: effects of urbanization, habitat fragmentation, and exploitation. All Graduate Theses and Dissertations. Paper 989. http://digitalcommons.usu.edu/etd/989.
Strasak, A.M., Zaman, Q., Pfeiffer, K.P., Göbel, G., Ulmer, H., 2007. Statistical errors in medical research — a review of common pitfalls. Swiss Med. Wkly. 137, 44–49.
Streiner, D.L., 1990. Sample size and power in psychiatric research. Can. J. Psychiatry 35, 616–620.
Sullivan, K.M., Foster, D.A., 1990. Use of the confidence interval function. Epidemiology 1, 39–42.
Twain, M., 1907. Chapters from my autobiography. In: Fishkin, S.F. (Ed.), 1996, The Oxford Mark Twain. Oxford University Press.
Underwood, A.J., 1997. Experiments in Ecology: Their Logical Design and Interpretation Using Analysis of Variance. Cambridge Univ. Press, Cambridge.
Underwood, A.J., Chapman, M.G., 2003. Power, precaution, type II error and sampling design in assessment of environmental impacts. J. Exp. Mar. Biol. Ecol. 296, 49–70.
Vasilakopoulos, P., O'Neill, F.G., Marshall, C.T., 2011. Misspent youth: does catching immature fish affect fisheries sustainability? ICES J Mar. Sci. 68, 1525–1534.
Wellek, S., 2010. Testing Statistical Hypotheses of Equivalence and Noninferiority, second ed. Chapman & Hall/CRC, Boca Raton.
Wolfe, R., Hanley, J., 2002. If we're so different, why do we keep overlapping? When 1 plus 1 doesn't make 2. Can. Med. Assoc. J. 166, 65–66.
Yoccoz, N.G., 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. Bull. Ecol. Soc. Am. 72, 106–111.
Yokoyama, L.Q., Amaral, A.C.Z., 2011. Allometric growth of a common Nassariidae (Gastropoda) in south-east Brazil. J. Mar. Biol. Ass. UK 91, 1095–1105.
Ziliak, S.T., 2011. Matrixx v. Siracusano and Student v. Fisher: statistical significance on trial. Significance 8, 131–134.
Ziliak, S.T., McCloskey, D.N., 2008a. The Cult of Statistical Significance: How the Standard Error Costs us Jobs, Justice, and Lives. University of Michigan Press, Ann Arbor.
Ziliak, S.T., McCloskey, D.N., 2008b. Science is judgment, not only calculation: a reply to Aris Spanos's review of 'The cult of statistical significance'. Eras. J. Phil. Econ. 1, 165–170.