2008

what exactly are you inferring? A closer look at hypothesis testing

MC Newman
Virginia Institute of Marine Science

SETAC PRESS

Environmental Toxicology and Chemistry, Vol. 27, No. 5, pp. 1013–1019, 2008
© 2008 SETAC
Printed in the USA
0730-7268/08 $12.00 + .00

Critical Review

''WHAT EXACTLY ARE YOU INFERRING?'' A CLOSER LOOK AT HYPOTHESIS TESTING

MICHAEL C. NEWMAN*
College of William & Mary–Virginia Institute of Marine Science, P.O. Box 1346, Route 1208 Greate Road,
Gloucester Point, Virginia 23062, USA

(*Received* 19 *June* 2007; *Accepted* 15 *October* 2007)

Abstract—This critical review describes the confused application of significance tests in environmental toxicology and chemistry that often produces incorrect inferences and indefensible regulatory decisions. Following a brief review of statistical testing theory, nine recommendations are put forward. The first is that confidence intervals be used instead of hypothesis tests whenever possible. The remaining recommendations are relevant if hypothesis tests are used. They are as follows: Define and justify Type I and II error rates a priori; set and justify an effect size a priori; do not confuse $p(E \mid H_0)$ and $p(H_0 \mid E)$; design tests permitting Positive Predictive Value estimation; publish negative results; estimate a priori, not post hoc, power; as warranted by study goals, favor null hypotheses that are not conventional nil hypotheses; and avoid definitive inferences from isolated tests.

Keywords—Statistical significance Experimental design Confidence intervals Power

INTRODUCTION

Scientists use accepted methods to generate new information, which is then organized around explanations. What constitutes accepted methods or favored explanations changes as experience and insight grow. It follows that all healthy sciences, including environmental toxicology and chemistry, require periodic review and revision of their practices and paradigms.

The central role of significance testing in assessing evidence suggests that associated statistical methods deserve critical evaluation. That is the specific goal of this review. Fundamental changes occurring in the application of statistics in health science and epidemiology [1–6], socioeconomics [7–9], psychology [10,11], and ecology [12,13] will be brought forward as being pertinent to environmental sciences. The overarching premise during this review is that significance tests should effectively guide rational transformation of observations into knowledge [1] about how contaminants act in or affect our environment.

Significance testing, especially null hypothesis–based significance testing, is arguably one of the most common ways in which scientific inferences are made by environmental chemists and toxicologists. Yet, its prominence and the unanimity about its soundness emerge more from custom than from scrutiny. The initial disagreements of Fisher, Pearson, and Neyman about key features remain unresolved and imperfectly integrated into present-day applications [2]. Common misinterpretations about the exact meanings [2,10,11,14–16] of Type I and II error rates confuse inferences, including those directly germane to regulatory activities [15]. Biologically trivial effects that are statistically significant are given unwarranted attention [8], and biologically crucial effect sizes are ignored [15]. Publication bias confuses literature interpretation, meta-analysis, and estimation of prior probabilities [5]. Realization of these shortcomings in what Gigerenzer calls statistical rituals [16] recently prompted fundamental shifts in other sciences, including discouragement or outright prohibition of significance testing in prominent medical [17], psychology [12], and conservation biology [13] journals.

FOUNDATION CONCEPTS

What is p?

No simple answer, save a careless one, exists for this question. The two most common explanations emphasize belief or relative frequency of occurrence. The original Bayesian context was that probability (p) suggests plausibility: p informs an investigator so that his or her degree of belief in a hypothesis can be adjusted based on evidence. For example, a forecast of a 95% chance of a blizzard suggests that one ought to remain home. One's degree of belief or level of certainty [18] changes as evidence accumulates and is used to generate p. Attempting to steer clear of Bayesian subjectivity, frequentists treat p as a probability for an observable event, outcome, or state, such as a 50% chance in the long run of heads resulting from fair coin tosses. One's state of belief about a certain hypothesis is irrelevant to the frequentist [19].

What does a ''significant'' test mean?

Again, no single answer exists. A set of contrasting answers, however, is commonly presented based on the pioneering works of Bayes, Fisher, Neyman, and Pearson.

Fisher discarded the Bayesian vantage as being too subjective and dependent on uncertain prior probabilities. He established significance testing as a more objective inferential approach. Strongly influenced by Popper's logic of falsification, Fisher asserted that sufficiently improbable events can be considered impossible: Statistical methods facilitate ''practical falsification'' or pseudofalsification [18–20]. A p value associated with a particular test statistic suggests whether a null hypothesis is sufficiently improbable to be considered practically falsified in the sense of a logical refutation [18]. The p value is the probability of getting the observed (e.g., 8

* newman@vims.edu.
Published on the Web 12/16/2007.

heads in 10 coin tosses) or more extreme outcomes (e.g., 9 or 10 heads in 10 tosses) under a null hypothesis [20]. Fisher did not advocate dogmatic application of any particular threshold *p*, although at one point, he did suggest the 0.05 convention as one of several standards for evidential strength [2,3,16]. He explained that a *p* value of 0.05 or less might be appropriate to make an inference in some instances but, as dictated by the researcher's understanding or goals, might only suggest the need for further experimentation in others. The prior probabilities of the Bayesian approach might be acceptable to Fisher only if derived in a clear, rigorous, and defensible manner [19]. The obvious fact that improbable events do occur remained a major shortcoming of Fisher's position. Sacrificing some objectivity, Popper countered that improbable outcomes in Fisher's approach might be better thought of as not being reproducible at will.

Neyman and Pearson judged Fisher's vantage to be untenable [18], introducing in its stead hypothesis testing, which defines primary and rival (alternate) hypotheses. Error rates are established for falsely deciding to reject the primary hypothesis (Type I error rate of α) or alternate hypothes(es) (Type II error rate of β). The seriousness of each error type determined which hypothesis was the primary hypothesis (i.e., that for which a decision error would have the most serious consequences) and the values for α and β [18]. The most serious error type determines the primary hypothesis, because selection of the most appropriate test statistic is based on the assumption that that hypothesis is true.

> When the errors can be distinguished by their gravity, the more serious of them is normally called a Type I error. . . suppose two alternative theories concerning a food additive were entertained, one that the substance is safe, the other that it is highly toxic. . . it would be less of a danger to assume that a safe additive was toxic than that a toxic one was safe [18].

Arguably, the most serious consequences if studying adverse effects most often would be associated with falsely deciding that an unsafe compound is safe; therefore, the null hypothesis should be that it is unsafe and the alternate hypothesis that it is safe. Oddly, the opposite is the more common practice. For example, a Dunnett's test might be applied to a sublethal effects data set to test the means of five nonreference treatments relative to a reference treatment mean under the null hypothesis that no difference exists. Adopting Fisher's terminology, the hypothesis associated with α was called the null hypothesis [18]. An effect size (ES) also is defined (i.e., what constitutes a meaningful effect in any particular test). For example, the ES might be a 25% decrease in reproduction in the above sublethal effects test if a toxicant-induced decrease in reproduction of that size would result in local extinction of a wild population. A test critical region is then established, and the observation-derived test statistic is compared to that region in a way that minimizes the chance of exceeding the specified error rates. Unlike Fisher's significance testing of a hypothesis, two hypotheses are incorporated, and the rates of each of the two error types are defined a priori based on judgment. The α and β are decision error rates associated with a particular test or experiment; they are not thresholds for deciding whether a hypothesis is plausible [16]. Unlike Fisher's approach, the Neyman–Pearson approach aims only to guide future behavior about the proposed hypotheses (i.e., to act as if one or another hypothesis were true), not to infer from the experiment that a null hypothesis was falsified [2,16]. According to the Neyman–Pearson line of reasoning, you are more likely to be correct in the long run if you behave toward a hypothesis in the manner suggested by the test results. A shortcoming of this context is that the in-the-long-run condition of such testing is a fiction [18] relative to actual scientific inquiry and decision making.

Fisher's approach focuses on inductive inference about a single hypothesis using pseudofalsification, whereas the Neyman–Pearson approach informs future behavior based on a test using two complementary hypotheses, associated decision error rates, and a specified ES. Both fail to completely avoid the subjectivity of Bayesian methods. Objective criteria are not possible for identifying a sufficiently improbable *p* value in Fisher's significance testing or for choosing the right combination of primary hypothesis, decision error rates, and ES in Neyman–Pearson hypothesis testing [18]. Depending on the test statistics applied to the same data, Fisher's null hypothesis might or might not be rejected. An objective way of defining how favorable a Neyman–Pearson hypothesis test result is relative to behaving as if a hypothesis were true is not congruent with the common practice of categorizing results with conventions such as accepted/rejected at $\alpha = 0.05$ or the ''roving α'' [2] classification of results as nonsignificant ($p > 0.05$), significant ($0.01 < p \leq 0.05$), or highly significant ($p \leq 0.01$).

The Bayesian approach dominated statistical thinking before Fisher, Neyman, and Pearson but was pushed aside in the 1920s as being too subjective. Bayesian methods currently enjoy much wider acceptance, primarily because the subjectivity in all approaches is more widely appreciated but also because convenient software now exists for its implementation. The Bayesian approach uses probabilities to gauge the belief in a particular hypothesis warranted by evidence; for example,

$$p(H_1|E) = \frac{p(H_1)p(E|H_1)}{p(E)} \tag{1}$$

where $p(H_1 | E)$ is the posterior probability of the hypothesis (H_1) given the evidence or data (E), $p(H_1)$ is the probability of H_1 prior to considering E, $p(E | H_1)$ is the probability of getting E if H_1 is true, and $p(E)$ is the probability of E regardless of whether H_1 is true. In this simplest form of Bayes' theorem, the prior probability (e.g., $p(H_1)$) is combined with a normalized likelihood ($p(E | H_1)/p(E)$) to estimate a posterior probability of a hypothesis based on the evidence (e.g., $p(H_1 | E)$). Here, the likelihood of the evidence given that the hypothesis is true, $p(E | H_1)$, is normalized to $p(E)$. As new evidence is gathered, the posterior *p* can be used as a prior *p* to produce a new posterior *p*. Several alternate hypotheses ($n - 1$) can be included in these Bayesian calculations to infer the degree of belief in a hypothesis (H_1) as warranted by evidence:

$$p(H_1|E) = \frac{p(H_1)p(E|H_1)}{\sum_{i=2}^{n} [p(H_i)p(E|H_i)]} \tag{2}$$

where $\Sigma\ p(H_n)$ sums to one. In this case, the prior *p* for H_1 is multiplied by the likelihood of the evidence given H_1 divided by the sum of the prior *p* for each H_i times the corresponding *p* for the evidence given a particular H_i. (The reader should note that Bayes factors [21] allow much more involved comparisons of competing models than illustrated here.) So *p* in Bayesian inference methods reflects an evidence-based belief in a particular hypothesis (among a specified set of hypotheses); for example, p(Fish kill | Copper discharge) = 0.98 warrants high, but not absolute, confidence that a fish kill will occur if a copper exposure of the specified qualities occurs.

SIGNIFICANCE TESTING PROBLEMS

That confusion exists about significance tests is unsurprising given both how recently this testing convention became established and the different interpretations of p. Emerging consensus from several sciences is that the resulting significance test malpractice now impedes as much as fosters progress [2,3,5,6,8–13,16].

One major difficulty involves the inconsistent combining of elements of the Fisher and Neyman–Pearson approaches [2] into what Cohen [10,11] calls the ''usual reject-H_0-confirm-the-theory'' approach. The following example using sublethal effects testing of a novel class of compounds is typical. For each in a class of new compounds, a series of concentration treatments is established within an experimental design prescribed by regulatory guidance or published by a reputable scientist. Observations of potential effect are taken and a Dunnett's test applied with a null hypothesis of no difference from a reference treatment ($\alpha = 0.05$). Next, the test statistic p value for each concentration treatment is used to classify that treatment concentration as either having or not having an adverse effect. The final research report does not discuss those compounds for which no significant adverse effect was noted in any treatment, because failure to reject the null hypothesis could have resulted from inadequate experimental design. The researcher decides to repeat those nonsignificant tests at some later date. The following problems emerge in this approach.

First, a misinterpretation of Neyman–Pearson hypothesis tests appears to be based on the Fisherian context. The α is one of two conditional probabilities of making a decision error during a specific hypothesis test, not a metric allowing one to decide if the null hypothesis is true. The Neyman–Pearson vantage cannot be taken to decide to act as if an effect exists, because two a priori decision rates were not established. On the other hand, no alternate hypothesis of an effect would exist if Fisher's vantage were taken. Second, the strict rejected versus not rejected interpretation of results is based merely on an arbitrarily selected convention ($p \leq 0.05$). Third, a pervasive misinterpretation exists that a low p value associated with the primary hypothesis (e.g., 0.04) indicates a high p of the secondary hypothesis being true (e.g., perhaps 0.96). Fourth, a pervasive inattention to power ($1 - \beta$) is present despite its essential role in Neyman–Pearson hypothesis testing. Fifth, judgment was not applied a priori to select the most appropriate Type I and II error rates. Sixth, a pervasive preoccupation with statistical significance and inattention to ES exists, including a failure to establish ES a priori. Seventh, the conventional no-effect (nil) hypothesis approach is applied such that the obligation to generate the most meaningful or discerning alternate hypotheses is ignored. Finally, a tendency exists to publish significant results more readily than nonsignificant results or to expand a study until a significant result is found based, incorrectly, on Fisher's pseudofalsification context of significance testing.

Although Fisher intended p to be a flexible inferential tool for rejection of a specified hypothesis and Neyman and Pearson intended p (to be assessed relative to the α decision error rate in obligatory combination with β and ES) to dictate behavior toward primary and alternate hypotheses, p values of less than 0.05 commonly are used to definitively reject a null hypothesis and to infer that the alternate hypothesis is true. For example, it is usual to conclude, using $\alpha = 0.05$ from a conventional sublethal effect test, that an effect exists at a treatment concentration, because that treatment's mean response was statistically significantly different from that of the reference mean. The false assertion that improbability of a primary hypothesis inferred from a p value means that the alternate hypothesis is probable is so prevalent that it has a name, the inverse probability error [3,10,11]. In actuality, a ''p value substantially overstates the evidence against a null hypothesis'' [2]. During the application of Neyman–Pearson hypothesis testing, it is likely that no or little time was spent balancing Type I and II error rates or determining what constituted a meaningful ES. The most important decision error might be unduly trivialized [8] or a toxicologically trivial, but statistically significant, effect elevated to the status of publishable finding [22]. Finally, the effect level (e.g., lowest-observed-effect concentration and associated no-observed-effect concentration) is approximated from a single test, and future testing is implied to be unnecessary. Results are not treated as conditional evidence subject to change as more evidence accrues.

These are the features of the presently confused blending of the decision-based approach of Neyman and Pearson with Fisher's context of pseudofalsification to produce what Ziliak and McCloskey [9] call mechanical testing. Gigerenzer [16] suggests that mechanical application of the ''null ritual'' is perpetuated by risk aversion associated with picking the wrong statistical tool from a diverse toolbox.

> Awareness of the origins of the [null] ritual and of its rejection could cause a virulent cognitive dissonance, in addition to dissonance with editors, reviewers, and dear colleagues. Suppression of conflicts and contradicting information is in the very nature of this social ritual [16].

Cognitive dissonance aside, evolving best practices are essential to the health of any science. Nine changes in current practices that can reduce some of these problems are suggested below.

Define and justify Type I and II error rates

The recent convention of applying an α of 0.05 in combination with an unspecified β and ES is inappropriate [23]. Hypothesis test α and β are chosen based on the seriousness of making each decision error, yet recent custom abrogates such judgments. Fixing α but allowing β to range within ill-defined limits set by experimental design and data variability implies that only one decision error is truly crucial (i.e., Fisher's vantage for judging the plausibility of a single hypothesis).

Quotients are convenient tools to balance the relative seriousness of the two decision errors [23]. Pairing error rates of $\alpha = 0.05$ and $\beta = 0.2$ implies that the consequences of making a Type I error is fourfold more serious than that of a Type II error, because $\alpha/\beta = 0.25 = 1/4$. Selecting $\alpha = \beta = 0.05$ for a toxicity test indicates that the seriousness of falsely rejecting the hypothesis of no effect is the same as that of falsely rejecting the hypothesis of an effect. The majority of sublethal effect tests fix α at 0.05 and, by virtue of standard design, produce β in the range of 0.2 (assuming an ES of ~20–30%) [24]. This creates the debatable default position that the consequences of Type I error (i.e., falsely rejecting the hypothesis of no toxic effect) are fourfold more serious than those of a Type II error (i.e., falsely rejecting the hypothesis of a toxic effect). Avoiding judgment does not eliminate decision error consequences: It simply obscures them, resulting in compromised judgments about the need for future scrutiny.

Define and justify the test ES

A p value is an unreliable indicator of whether a decision is being made about a meaningful effect—about what McCloskey calls the hypothesis test's ''oomph'' [8]. As an extreme illustration, if the size of an effect is treated as being irrelevant, a null hypothesis of no difference will always be rejected given enough observations. Establishing a test without defining a meaningful ES is inherently misleading. A hypothesis test should be designed with an a priori ES based on sound insight and knowledge [9–11].

First consider using ES confidence limits

Arguments to replace hypothesis testing with presentations of confidence limits are increasing as a consequence of the confusion surrounding ES, p, and error rates [17,25,26]. For 25 years, several key human health science journals have depended increasingly on confidence intervals to convey ES, precision, and statistical significance simultaneously [17].

It is important to note that the 95% value for confidence intervals, like the Type I error rate of 0.05, is a convention and that other values might be more appropriate depending on circumstances or goals. Regardless of which percentage is selected, care should be taken when interpreting intervals. For example, a 95% confidence interval defines the interval $\bar{x} - t_{n-1,95}\text{SE} \leq \mu \leq \bar{x} + t_{n-1,95}\text{SE}$, where SE is the standard error. If one were to generate many such intervals, 95% of the intervals would contain μ in the long run. It is incorrect to state the probability is 0.95 that a particular interval includes μ.

Cumming and Finch [25] suggest the following three general rules for confidence interval presentation: Select error bars associated directly with the relevant effect, presentation should be sensitive to the experimental design, and the confidence intervals should be thoroughly interpreted. More guidance can be found in Cummings and Finch [25] and in Di Stefano [26] for applying confidence intervals to a range of common situations. Altman et al. [27] describe detailed applications of confidence interval techniques, including those dealing with means, medians, proportions, regression analysis, time-to-event studies, and meta-analyses. Altman et al. [27] also provide convenient software to facilitate implementation of these methods. The SAS® software package [28] also has procedures (e.g., INTERVALS option in the PROC CAPABILITY) that make calculations convenient for a wide range of analyses.

Do not confuse p(E | H*₀*) *and* p(H*₀* | E*)*

This point can be introduced with an old joke. Walking down a city street, a woman passes a man who is jumping and waving his arms wildly. She asks him why he's doing this, and he responds, ''It scares away elephants.'' To her retort that there are no elephants in the city, the man exclaims, ''You see. It works!'' Put in more explicit, but equally absurd, terms, p(No Elephants | Behavior Scares Elephants) = p(Behavior Scares Elephants | No Elephants). Obviously, knowledge of other probabilities, such as p(Elephants), is required to judge the soundness of the gentleman's hypothesis.

Most conventional applications of null-hypothesis significance tests generate test statistics associated with the probability of getting the data or evidence if the null hypothesis is true (i.e., $p(E \mid H_0)$). Therefore, rejection of H_0 reflects the chance of getting the data if H_0 is true, not how likely it is that H_0 is true given the data (i.e., $p(H_0 \mid E)$). The distinctness of $p(E \mid H_0)$ and $p(H_0 \mid E)$ is obvious from Bayes' theorem above. More information ($p(E)$ and $p(H_0)$) than provided by the hypothesis test is needed to estimate the probability of H_0 given E. Continuing with the previous example of applying Dunnett's test to sublethal effects test data, rejection of the null hypothesis of equal means for the reference and a toxicant-spiked treatment does not lead directly to the conclusion that a sublethal effect exists at the treatment concentration. It indicates only that the observations have a low probability in the long run of having occurred if the null hypothesis is true.

Design tests allowing estimation of Positive Predictive Value

How does one estimate the probability of an alternate hypothesis being true given a significant hypothesis test? An estimate of this probability is the Positive Predictive Value (PPV):

$$\text{PPV} = \frac{(1 - \beta)R}{R - \beta R + \alpha} \tag{3}$$

where R is the ratio of ''true relationships'' to ''no relationships'' estimated prior to testing [4,5]. Calculation of PPV from the above equation requires informed estimation of R and judgment about the appropriate α and β based on the seriousness of making decision errors and ES. Otherwise, the probability cannot be established for this hypothesis being true given a positive test. The related probability that the null hypothesis is true given a significant test (False Positive Result Probability [FPRP]) is the following [4]:

$$\text{FPRP} = \frac{\alpha(1 - \pi)}{\alpha(1 - \pi) + \pi(1 - \beta)} \tag{4}$$

where π is the prior probability of association between treatment and effect (i.e., $R/(R + 1)$).

The previous example of a hypothetical sublethal effect data set evaluated with a one-way Dunnett's test can be used to illustrate that a test's Type I error rate of 0.05 is not a reliable indicator of PPV. Assume for purposes of illustration that most tests have five nonreference treatments and that most toxicologists design experiments so that the lowest-observed-effect concentration is one of the middle treatments. Then, R would be two or three significant treatments of a total of five treatments (i.e., 2/5 or 3/5). Also, let α and β be 0.05 and 0.2 [24], respectively:

$$\text{PPV}_{R=0.4} = \frac{(1 - \beta)R}{R - \beta R + \alpha} = \frac{(0.8)(0.4)}{0.4 - (0.2)(0.4) + 0.05} \approx 0.86$$

$$\text{PPV}_{R=0.6} = \frac{(1 - \beta)R}{R - \beta R + \alpha} = \frac{(0.8)(0.6)}{0.6 - (0.2)(0.6) + 0.05} \approx 0.91$$

The common assumption that p is minimally $1 - \alpha$ or 0.95 that an effect exists at a treatment concentration given a statistically significant test is clearly wrong. Here, 9 in 10 would be a better estimate than the presupposed 19 in 20, or better, chance. Similarly, it is untrue that 0.05 reflects the probability that the null hypothesis is true given a significant test (i.e., FPRP). The FPRP ranges in this example from 0.09 to 0.14, opening up the question of how small the FPRP, or the large PPV, must be to make a decision from this type of sublethal effect testing. The situation worsens for studies with higher β values, such as mesocosm [29] and epidemiology [4] studies.

Epidemiology studies reviewed by Wacholder et al. [4] had typical R and β values resulting in a PPV of 0.5. A statistically significant result in one of the reviewed epidemiology studies had only a 50:50 chance of correctly indicating a true effect. Equally pessimistic were Kraufvelin's comments about tests of mesocosm data [29].

It is recommended that the information needed to estimate PPV or FPRP be generated and included in discussions of environmental chemistry or toxicology studies.

Publish negative results

Studies showing no significant effects are judged to be of ambiguous value based on the historical pseudofalsification vantage point. The common practice of not setting or reporting Type II error rates should lead to caution when interpreting such studies, although similar caution, oddly, is not practiced when interpreting tests with significant effects.

Publication bias has two undesirable consequences. The underlying goal associated with most testing is to understand PPV or FPRP. Publication bias makes the associated estimation of R or π inaccurate. This compromises inferences from the literature, although methods for coping with this bias do exist [5]. A more subtle effect emerges as this publication bias combines with the publication time-lag bias. Ioannidis and Trikalinos [6] observe that the initial literature for a new research theme tends to have more reported significant studies than reported nonsignificant studies. Reports of significant findings tend to be more contrasting if emphasis is placed on Type I error rates. Researchers are attracted to contrasting reports, so these kinds of studies are more likely to catch the attention of editors and move quickly into publication. So, the preoccupation with Type I error rate and the neglect of PPV initially seeds the literature with contrasting significant studies. Ioannidis and Trikalinos [6] define the consequent Proteus phenomenon as being the appearance of highly contrasting studies during the onset of any new research theme, followed by a gradual movement toward more consistency among reports. The debates associated with the Proteus phenomenon can impede initial progress in a new area of research.

Estimate a priori, not post hoc, power

Test power ($1 - \beta$) is extremely important to define, because failure to reject H_0 might reflect either insufficient power or the high probability of the observations (i.e., the common nonrejected-null-hypothesis dilemma) [14]. Consequently, well-intended journals and agencies request inappropriate post hoc estimates of observed test power [14] to suggest the reason why H_0 was not rejected. Recalling the core roles of β, α, and ES in hypothesis testing, however, power makes sense only if established a priori. Unfortunately, the requirement of a pilot study or critical literature analysis seems to foster avoidance of a priori power estimation in favor of post hoc power estimation.

Hoenig and Heisey [14] argue forcefully against post hoc power estimation from observed test statistics. They describe the power approach paradox in which it is wrong to assume that the nonsignificant H_0 for a test with high power is more likely to be true than that for a second nonsignificant H_0 with lower associated power. Observed power adds no insight, because it is determined by the test's p value, which can vary widely for the two nonsignificant tests. Equally unhelpful are post hoc estimates of minimum significant difference or detectable ES. Instead, Hoenig and Heisey recommend inference from confidence intervals: "Once we have constructed a confidence interval, power calculations yield no additional insights" [14]. This is consistent with the third change suggested above.

Use null, not nil, hypotheses

Emerging from Fisher's initial vantage and the "usual reject-H_0-confirm-the-theory" approach [10,11], a bad habit of automatically using a hypothesis of no difference or correlation as the null hypothesis has become entrenched. Cohen [10] refers to this as the nil hypothesis approach, which stipulates an ES of zero and misinterprets Fisher's term "null" to mean "zero" instead of "to be nullified." As already mentioned, an ES of zero can always by rejected given enough observations, so this approach lacks merit as a reliable tool for informing decisions. It also is inconsistent with the Neyman–Pearson context of hypothesis testing, which informs decisions to act as if one or another of two (or more) hypotheses is true.

Null hypotheses should be established based on sound judgment. For example, a H_0 of the decrease in reproductive output is more than 25% under a certain exposure regime might be based on the demographic insight that the species population likely would go locally extinct if output dropped by more than 25%. This kind of null hypothesis selection and testing requires more thoughtfulness about decision error consequences and about error rate and ES magnitudes, but it rewards such effort by producing much more meaningful results [10,14,16].

Avoid definitive inferences from isolated tests

A review of the above materials should suggest that a single hypothesis test rarely is as useful as a series of inferentially linked experiments and associated tests. As evidence accumulates, the PPV or FPRP changes based on the changes to R or π. The most effective inferences emerge from carefully planned research programs or themes [5].

ENVIRONMENTAL CHEMISTRY AND TOXICOLOGY

How does the environmental chemistry and toxicology literature stand up to the issues presented above? Ten representative journals with good impact factors were reviewed to suggest an initial answer: *Aquatic Toxicology, Archives of Environmental Contamination and Toxicology, Chemosphere, Ecotoxicology, Ecotoxicology and Environmental Safety, Environmental Pollution, Environmental Science and Technology, Environmental Toxicology and Chemistry, Marine Pollution Bulletin*, and *The Science of The Total Environment.* For each journal, a random number generator was used to pick the volume and then the article number for 10 articles published between 1996 and 2006 inclusive. Features of each article were scored (Yes, No, or Not Applicable) as summarized below and in Figure 1. Ninety-seven of the 100 surveyed papers applied quantitative methods amenable to hypothesis testing. The 95% confidence intervals shown in parentheses were produced with the Wilson method [27] from frequencies first estimated with the SAS 9.1 software package [28] PROC FREQ.

In 57% (47–67) of the 97 quantitative publications, inferences were based on hypothesis testing. Notably, many of the surveyed environmental chemistry publications presented results graphically and compared them to predictions from theories instead of relying heavily on hypothesis testing. This 57% is lower than the percentage noted in a 2005 survey of two ecology journals (*Ecology* and *Journal of Ecology*) and

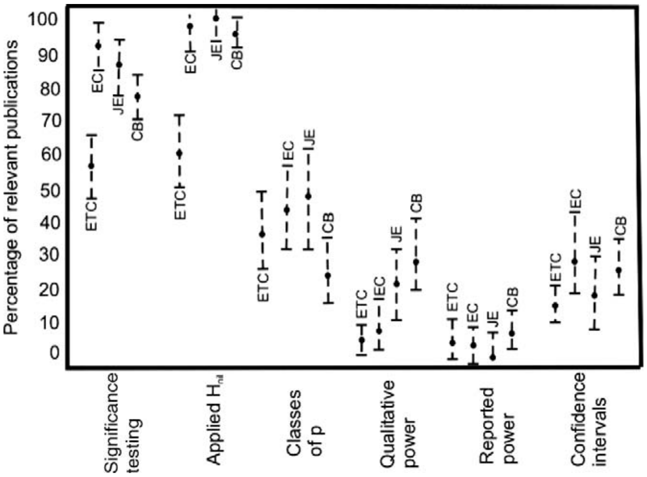

Fig. 1. A comparison of results of the environmental science and chemistry survey (ETC) to a 2005 survey [13] of two prominent ecology journals, *Ecology* (EC) and *Journal of Ecology* (JE), and two conservation biology journals, *Conservation Biology* and *Biological Conservation* (CB). Mean percentages are shown with corresponding 95% confidence intervals. Note that H_{nil} indicates use of the conventional nil hypothesis during statistical testing.

two conservation biology journals (*Conservation Biology* and *Biological Conservation*) [13]. Most of the publications using hypothesis tests applied the nil hypothesis. This nil hypothesis use rate was lower than that in the ecology/conservation biology survey, because the base rate of hypothesis testing also was lower, as just stated.

In 62% (49–74) of the publications employing hypothesis tests, test results were treated in a dichotomous significant-or-not-significant manner based on $\alpha = 0.05$. A set α chosen by convention before the test is incongruent with Fisher's context of significance testing. A priori selection of an α, but not a β, is inconsistent with the Neyman–Pearson vantage. As discussed already, this use of p values can lead to considerable confusion. Another 38% (26–51) of the publications used p values to categorize results within schema, such as not significant, significant, very significant, or highly significant. Such use is specifically incompatible with the Neyman–Pearson framework in which α and β are established a priori and p has no meaning outside the decision error context. In that context, either the null hypothesis or the alternate hypothesis is rejected with the specified error rates. Such use with an inferred alternate hypothesis is inconsistent with Fisher's vantage. The 2005 survey of ecology and conservation biology journals [13] showed similar levels of use for such schema.

No publication using hypothesis testing reported calculating power a priori, and a low 4% (1–13) of publications discussed power issues in qualitative terms only. Power or some metric of minimum ES was estimated post hoc in 6% (2–16) of the publications employing hypothesis testing. The 2005 survey of ecology and conservation biology journals [13] also had extremely low percentage reporting of power.

Confidence intervals were used in some manner to make inferences in only 16% (12–29) of the quantitative environmental chemistry and toxicology publications. The 2005 ecology and conservation biology journal survey [13] reported only a slightly higher level of confidence interval use. These percentages are well below those of the *British Medical Journal,* which after editorial policy changes increased from 4% (1977) to 62% (1994) [12]. Similarly, the *American Journal of Epidemiology* had a 70% confidence interval use rate in 1990. More engagement of statistical editors, as done for the *British Medical Journal,* might improve this situation in environmental chemistry and toxicology journals.

Relative to alternate approaches, only 3% (1–9) of the quantitative publications applied information theory–based approaches. None used Bayesian methods.

Quantitative results were analyzed in relative isolation from other experiments in 84% (72–91) of the surveyed studies. The exceptions included those compiling large toxicological data sets. No study estimated PPV or FPRP.

Generally, applications of hypothesis testing in environmental toxicology and chemistry were similar to those in other environmental sciences. The results for ecology and conservation biology discussed above led Fidler et al. [13] to conclude that ''further efforts are clearly required to move the discipline toward improved practices.'' The same conclusion seems relevant to environmental chemistry and toxicology.

Unquestionably, hypothesis testing is a major tool that is misapplied in many fields. Interpretation of p values is confused (e.g., the reject-or-accept nil hypothesis routine based on 0.05). Power is ignored or given short shrift, being applied post hoc incorrectly in most of the few studies that give it attention. None of the surveyed environmental toxicology or chemistry publications set α, β, and ES a priori or attempted to estimate R or π from the literature. Therefore, the PPV or FPRP could not be estimated from results. The value of calculating PPV in environmental health risk was clearly illustrated in a study by Rizak and Hrudey [30], in which water-quality professionals were presented with a hypothetical detection of a pesticide. Not understanding the value of calculating PPV, most reported high certainty (80–100% chance) of the pesticide being present when, in fact, a low chance (5%) existed. To end on a positive note, however, 16% of the surveyed studies (particularly environmental chemistry studies) used confidence intervals effectively to assess results.

CONCLUSIONS ABOUT IMPROVING STATISTICAL INFERENCE

Two general recommendations suggest themselves for immediate implementation based on the materials summarized above. First, any interpretation of hypothesis testing as currently practiced should explicitly address any relevant test shortcomings and not extend inferences beyond those limits. Second, the teaching of statistics to environmental science students should shift away from a traditional emphasis on hypothesis testing to a more flexible approach embracing other valuable vantages, especially the Bayesian and information theory–based vantages.

Nine specific recommendations also are offered. The first is that confidence intervals be used instead of hypothesis tests whenever possible. Other alternate methods include Bayesian and information theory–based techniques. If hypothesis testing is done, the following eight recommendations are made: Define and justify Type I and II error rates a priori; define and justify an ES a priori; do not confuse $p(E \mid H_0)$ and $p(H_0 \mid E)$ during interpretation of results; design tests to allow estimation of PPV; publish negative results; estimate a priori, not post hoc, power; avoid nil hypotheses as much as reasonable; and avoid definitive inferences from isolated tests.

Acknowledgement—Thanks are due to the editor-in-chief for allowing me to carefully explore this issue and understand more clearly my many past errors. The author is very grateful to John Carriger, Mark Crane, Philip Dixon, Margaret Mulvey, and two anonymous reviewers for their invaluable comments on earlier versions of this manuscript.

REFERENCES

1. Woodworth GG. 2004. *Biostatistics. A Bayesian Introduction.* John Wiley, Hoboken, NJ, USA.
2. Goodman SN. 1993. P values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *Am J Epidemiol* 137:485–496.
3. Sterne JAC, Davey Smith G. 2001. Sifting the evidence—What's wrong with significance tests? *BMJ* 322:226–230.
4. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. 2004. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *J Natl Cancer Inst* 96:434–442.
5. Ioannidis JPA. 2005. Why most published research findings are false. *Public Library of Science Medicine* 2:e124–e126.
6. Ioannidis JPA, Trikalinos TA. 2005. Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *J Clin Epidemiol* 58:543–549.
7. Altman M. 2004. Statistical significance, path dependency, and the culture of journal publication. *The Journal of Socio-Economics* 33:651–663.
8. McCloskey DN. 1995. The insignificance of statistical significance. *Am Sci* 272:32–33.
9. Ziliak ST, McCloskey DN. 2004. Significance redux. *The Journal of Socio-Economics* 33:665–675.
10. Cohen J. 1994. The earth is round ($p < .05$). *Am Psychol* 49: 997–1003.
11. Cohen J. 1995. The earth is round ($p < .05$): Rejoinder. *Am Psychol* 50:1104.
12. Fidler F, Cumming G, Burgman M, Thomason N. 2004. Statistical reform in medicine, psychology, and ecology. *The Journal of Socio-Economics* 33:615–630.
13. Fidler F, Burgman MA, 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.
14. Hoenig JM, Heisey DM. 2001. The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician* 55:1–6.
15. Crane M, Newman MC. 2000. What level of effect is a no observed effect? *Environ Toxicol Chem* 19:516–519.
16. Gigerenzer G. 2004. Mindless statistics. *The Journal of Socio-Economics* 33:587–606.
17. Fidler F, Thomason N, Cumming G, Finch S, Leeman J. 2004. Editors can lead researchers to confidence intervals, but can't make them think. *Psychological Science* 15:119–126.
18. Howson C, Urbach P. 1989. *Scientific Reasoning. The Bayesian Approach.* Open Court, La Salle, IL, USA.
19. Hacking I. 2001. *An Introduction to Probability and Inductive Logic.* Cambridge University Press, Cambridge, UK.
20. Pedhazur EJ, Pedhazur Schmelkin L. 1991. *Measurement, Design, and Analysis. An Integrated Approach.* Lawrence Erlbaum Associates, Hillsdale, NJ, USA.
21. Gelman A, Carlin JB, Stern HS, Rubin DB. 1997. *Bayesian Data Analysis.* Chapman & Hall/CRC, Boca Raton, FL, USA.
22. Nakagawa S. 2004. A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behav Ecol* 15:1044–1045.
23. Cohen J. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, Mahwah, NJ, USA.
24. Van der Hoeven N. 1998. Power analysis for the NOEC: What is the probability of detecting small toxic effects on three different species using the appropriate standardized test protocols? *Ecotoxicology* 7:355–361.
25. Cumming G, Finch S. 2005. Inference by eye. *Am Psychol* 60: 170–180.
26. Di Stefano J. 2004. A confidence interval approach to data analysis. *For Ecol Manag* 187:173–183.
27. Altman D, Machin D, Bryant TN, Gardner MJ. 2000. *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*, 2nd ed. British Medical Journal Books, London, UK.
28. SAS Institute. 2003. *SAS Statistical Package,* Ver 9.1. Cary, NC, USA.
29. Kraufvelin P. 1998. Model ecosystem replicability challenged by the ''soft'' reality of a hard bottom mesocosm. *J Exp Mar Biol Ecol* 222:247–267.
30. Rizak SN, Hrudey SE. 2006. Misinterpretation of drinking water-quality monitoring data with implications for risk management. *Environ Sci Technol* 40:5244–5250.