Herpetologists' League

Hypotheses, Errors, and Statistical Assumptions
Author(s): Daniel Simberloff
Reviewed work(s):
Source: *Herpetologica*, Vol. 46, No. 3 (Sep., 1990), pp. 351-357
Published by: Herpetologists' League
Stable URL: http://www.jstor.org/stable/3892978
Accessed: 10/02/2012 15:41

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Herpetologists' League is collaborating with JSTOR to digitize, preserve and extend access to *Herpetologica.*

http://www.jstor.org

Herpetologica, 46(3), 1990, 351–357
© 1990 by The Herpetologists' League, Inc.

HYPOTHESES, ERRORS, AND STATISTICAL ASSUMPTIONS

DANIEL SIMBERLOFF

Department of Biological Science, Florida State University, Tallahassee, FL 32306, USA

COSTS OF TYPE I AND TYPE II ERRORS

SEAMAN AND JAEGER (1990) contend that presumptuous use of parametric statistical methods to test hypotheses can lead ecologists astray; I wholeheartedly concur. As they point out, the usual argument against nonparametric alternatives is that they lack power—there is a high probability of failing to reject a false null hypothesis (type II error). To increase the power of a particular test (lower probability of type II error), however, one must generally increase probability of type I error, the probability of rejecting a true null hypothesis (Conover, 1980).

Seaman and Jaeger (1990) suggest that the view that nonparametric tests are less powerful than parametric alternatives is often untrue and rests in many instances on assumptions about the shape of an underlying distribution that are either incorrect or untestable. In addition, even in situations in which nonparametric tests are somewhat less powerful than parametric ones, there may still be a good reason for ecologists to favor the former—relative costs of the two kinds of errors (Connor and Simberloff, 1986; Toft and Shea, 1983).

It is important to distinguish between statistical hypotheses and scientific hypotheses (Boen, 1989; Connor and Simberloff, 1986). Scientific hypotheses are about phenomena in nature. Statistical hypotheses are about properties of populations based on samples. Thus, a statistical hypothesis can be a quantified application of a scientific hypothesis to a specific set of data. Rejection of one or more statistical hypotheses would constitute one piece of evidence to be weighed in deciding whether to reject a scientific hypothesis. A related distinction is between global and local hypotheses (Dolby, 1982). A global hypothesis applies to all of nature, while a local one applies to particular systems. A specific statistical hypothesis is local unless global populations have been sampled.

Error classification was developed specifically for testing statistical hypotheses, but the terms are used metaphorically, and appropriately, for global, scientific hypotheses. In either case, if the null hypothesis is that some process or phenomenon has no effect, then a type I error consists in concluding that the process *does* have an effect when, in fact, it does not. A type II error for the same case would be concluding that the process has no effect when it actually does.

Finally, a distinction must be made between classical statistical hypothesis-testing and decision theory (Kyburg, 1974). Testing statistical hypotheses is often an aid to inferring whether a scientific hypothesis is true or false, as noted above. It does not explicitly take account of costs of errors. By contrast, in some settings a particular statistical hypothesis may be tested over and over again, and the test results acted on each time. For example, one may test samples from batches of computer chips, then save or discard entire batches based on test results. This problem has motivated the development of decision theory, which is a theory of acting rationally (with respect to expected losses and gains) in the face of uncertainty, rather than a theory of inference about nature. If statistics is to be used for this purpose, costs attending various kinds of errors must be explicitly assessed. Below I argue that costs of errors must also be considered when statistical hypotheses are tested as part of testing scientific hypotheses.

Perusing the medical literature nowadays, one can easily conclude that type II error is the greater scourge and entails the larger costs (e.g., Freiman et al., 1978; Marks et al., 1988). Bourne (1987) went so far as to title a paper, "No Statistically Significant Difference?: So What?" His particular hypothetical example was a test of whether a particular treatment of a disease is efficacious, and he betrayed his underlying reasoning at the outset (p. 40):

"In reality, the two groups referred to almost certainly are at least minutely different." He seems to have felt that the observed difference between samples *might* at least partly reflect a treatment effect, and that, in terms of short-term decisions about clinical practice, failing to recognize even a minute effect costs more than attributing the difference to the treatment when it is in fact due to chance. He went on to argue that negative findings (failure to find a significant difference) also entail a long-term cost—they discourage innovation—so editors should reject them unless they are accompanied by a power analysis.

This rationale seems to equate to, "If it can't be shown to be worse, it might be better, so we ought to use it"—effective unless proven otherwise. I imagine if I were working on major medical problems, or suffered from one, I might be inclined to agree. There must be a tremendous temptation in the face of serious illness to be willing to try almost anything not known to be harmful. Certainly this is true in cases of last resort, in which no effective treatment is currently known. An example is the Federal Drug Administration's allowing use of DDI and compound Q to treat A.I.D.S., admitting that they are not yet known to be safe and effective (the usual criteria for certifying a pharmaceutical). This approach is fundamentally in the realm of decision theory, and the chief concern is not a scientific hypothesis about nature. The emphasis on type II errors in medicine seems also to reflect fear that physicians will confuse rejection of a statistical hypothesis with rejection of a scientific hypothesis (e.g., Angell, 1989; Boen, 1989; Fleiss, 1986*a*; Poole, 1987). Often these authors cite studies showing appalling statistical ignorance of medical researchers (e.g., Matthews and McPherson, 1987; Wulff et al., 1987).

Although humanitarian instincts as outlined above would probably prevail, there are at least two potential costs to the decision to use a medical treatment not yet shown to be effective. The first is a decision-theoretic humanitarian matter—could not the financial cost and emotional and physical anguish of tracking down and using a dubious pharmaceutical outweigh a very small chance of benefit? The spectre of distraught cancer patients flocking to Mexico for laetrile must haunt any decision to approve a treatment not shown to be effective. The second cost is more in the realm of inference about scientific theory. If a treatment is widely used in desperation, continuing efforts to assess its value are compromised for want of an adequately controlled experiment. Further, expenditure of effort and funds to refine and elaborate the technique may turn out to be wasted. As I noted, Bourne (1987) feared that accepting an incorrect finding of "no significance" (type II error) will discourage potentially useful innovation. But he seemed unconcerned about the corresponding consequence of a type I error—an expensive wild goose chase to perfect or expand a worthless procedure. Similarly, Angell (1989) and Boen (1989) were at pains to aver that lack of statistical evidence that a treatment *is* effective is not tantamount to evidence that the treatment is *not* effective, and emphasized such problems as small sample size and lack of power. However, they did not state clearly exactly what observations *would* impel them to conclude clinical nonsignificance of a treatment.

In ecology, especially academic ecology, humanitarian considerations are usually not as obvious, and certainly not as immediate. However, the second cost of a type I error—expenditure on a worthless program of scientific research—may be substantial (Connor and Simberloff, 1986). For one thing, many people can waste years on both theoretical development and field measurements. Consider increasingly complex models of limiting similarity and species-packing based on the Lotka-Volterra competition equations. Brown (1981) eloquently suggested a real cost of a decade's worth of this metastasis:

> "Theoreticians were not alone in their enthusiasm for these models. Many more empirical ecologists spent much time studying interspecific competition and trying to measure the α_{ij}'s required to test the theories. Among both theoretical and field ecologists there was widespread belief that interspecific competition was the primary factor which limits diversity, and that working out the mechanisms of competitive interaction was the key to understanding the organization of communities.

> In the last few years enthusiasm has given way to disappointment as this approach has proven unproductive. It is worth inquiring into the reasons for this failure so that we may avoid making the same mistakes in the future." (Page 881)

Brown (1981) was referring, of course, not to a single instance of type I error in a specific statistical test, but to persistent credulity with respect to a global scientific hypothesis. However, a persistent tendency to see an effect is composed of, and colors, numerous individual cases and statistical tests of local hypotheses.

In areas of applied ecology where decision theory may be appropriate, the cost of type I errors can be more than just intellectual churning by a set of professors and graduate students—it can be very concrete. For example, by ANCOVA on data on birds of small islands around Britain, Pimm et al. (1988) found large birds to be at greater risk of extinction than small birds for population sizes above seven pairs. Slopes of a "risk of extinction" variable regressed against population size for different species differed at $P = 0.002$. Pimm et al. (1988) pointed to the utility of this theory for reintroduction programs. For example, they noted that conservation biologists may have to decide whether 50 individuals available for reintroduction should be released as one large propagule or five propagules of 10 individuals each.

Conservationists *are* making such decisions; for example, Craig (1990) discussed introduction of kiwis on small islands around New Zealand. However, it is doubtful that the hypothesis by Pimm et al. is correct even for the birds of Britain. Among myriad problems with the original analysis, Tracy and George (personal communication) found no significant difference between slopes of the regression lines for small and large birds, and no convincing support for the conclusion that large birds are more prone to extinction at high population densities. If reintroduction decisions on rare birds like kiwis were made on the basis of a false hypothesis, a failure owing to this cause could be tragically costly.

What about the cost of a type II error in ecology? Toft and Shea (1983) argued that, in specific decisions that must routinely be made in applied ecological management, costs of type II error can greatly exceed those of type I error. Their hypothetical example concerned two dosages of a pesticide that will be applied to thousands of hectares at substantial cost. If a type II error is made in finding that the lower concentration is as effective as the higher one, several costs may be incurred: the pesticide application may have to be repeated, the crop may fail, the pest may unnecessarily develop resistance, the public might lessen its support of applied research. These costs are all substantial, but the hypothetical example did not rationalize the claim that costs of type I error would be lower. Suppose the researcher finds that the higher concentration *is* more effective, when, in fact, it is not. The cost of the sprayed chemical would be greater than it had to be, the pest would develop resistance more quickly than would have been necessary, direct and indirect effects of the pesticide on non-target organisms would be enhanced, and the public might come to be predisposed against all pesticide use by virtue of the preceding three effects.

In statistical hypothesis-testing to advance scientific understanding, might type II error discourage innovation and potentially productive continuing research, as Bourne (1987) suggested for medicine? I doubt it. First, a major thrust of ecological debate over the past few years has been argument over whether the proper statistical test has been used, particularly in instances where a result seems borderline (say, $0.05 < P < 0.10$). Second, a current trend in ecology is enthusiastic advocacy of "pluralism", the recognition that different ecological systems may be differently structured and governed by different forces (e.g., Schoener, 1986). In the face of this ecumenism, I doubt that a researcher would even claim a finding of "no significant difference" for his/her particular system is applicable to many other systems, and it is doubly doubtful that someone else working on another system would abandon his/her pet hypothesis because of the negative result. There seems to be some appreciation in ecology of the difference between statistical and scientific hypotheses, and concern with how global the latter are. Even had the analysis of Pimm et al. (1988)

not been statistically faulty, it would have been surprising if no one had questioned whether a comparison of large British bird species like the raven and jackdaw with small ones like the robin and chaffinch is really very informative on kiwi ecology.

In sum, it is not a clearcut matter even in medicine that type II errors are inherently more costly. First, the cost of a type II error cannot be considered in a vacuum; it depends on what the alternative is. Further, it may be that statistical naiveté is what renders it costly. In ecology, I am unconvinced that type II errors are intrinsically more costly than type I errors in assessing either statistical or scientific hypotheses. In both fields, the danger of misinterpreting a statistical test in a way that might increase the cost of a type II error could be lessened by routine publication of power estimates if these are available (Boen, 1989; Bourne, 1987; Marks et al., 1988; Toft and Shea, 1983). Even better would be more rigorous statistical training for people who are going to read papers with statistical tests.

Estimation vs. Testing Hypotheses

In medicine, the current emphasis on errors of type II rather than those of type I is reflected in a feeling that parameter estimation is a much more appropriate statistical technique than hypothesis-testing (e.g., Freiman et al., 1978; Marks et al., 1988; Moses et al., 1984; Poole, 1987; Poole et al., 1984; Rothman, 1978). Poole et al. (1984) went so far as to argue that, "Scientists ought to view their general research objective as one of measurement" (p. 1382), and they regretted the fact that choosing the right test has become an area of concern!

There is no formal classification of errors in estimating parameters. One assumes that the estimate will err, and the width of the confidence band gives information on the likely degree of error.

Of course, confidence limits around an estimated parameter can to some extent allow the same sort of inference that one makes in a hypothesis-test. For example, the choice of a confidence coefficient can ordinarily be interpreted as setting a desired probability of type I error; that is, one can envision a 95% confidence interval for a parameter as the set of null-hypothesis values that the data will not reject at level 0.05. Thus, one can simply construe the confidence band as partitioning possible parameter values into two sets—those consistent with the observed data and those not consistent (Thompson, 1987). If the first set contains the null value of the parameter, one would have to be wary. Poole (1987) lamented that much of the current medical literature goes even further, and takes the latter observation as tantamount to accepting the null hypothesis. He pointed out that "confidence intervals focus attention on the magnitude of an estimate of a *meaningful* parameter . . . and, as a separate matter, on the precision of that estimate" (p. 195, my italics). Significance tests, in his view, blend together the magnitude of the estimate and its precision.

Though a significance test does not give an estimate, stating the power and significance level for rejection in a hypothesis-test can to some extent allow the same sort of inference that one normally makes in estimation (how much confidence can be placed in a finding of no significance?). Also, the actual *P*-value of a significance test *does* convey some information that a confidence interval does not provide directly (Thompson, 1987), while Fleiss (1986*a*) suggested some medical contexts in which hypothesis-testing is highly appropriate.

Further, I would argue that categorically favoring estimation over hypothesis-testing reflects a tacit devaluation of type I error. If a number is to be estimated, there is a predisposition to believe that it matters. Perhaps this predisposition is betrayed by Poole's use of the term "meaningful" in the above quote. The important thing seems to be to come up with a number, and, as Bourne (1987) pointed out, with real data the number will never be zero. Whether it differs significantly from zero is of secondary concern. One wonders under exactly what circumstance a factor would be judged irrelevant. Fleiss (1986*b,c*) viewed the explicit criteria for a decision of "no effect" as a general advantage of hypothesis-tests over confidence intervals.

In ecology, estimating parameters has also been recommended as an alternative to hypothesis-testing, particularly because

multiple factors affect virtually every phenomenon. For example, Quinn and Dunham (1983) emphasized multiple causality in suggesting that the real object of ecological investigation is to determine the fraction of variation that can be explained by each of several candidate predictors. Gilpin and Diamond (1984) doubted that hypothesis-testing will ever be useful in community ecology, at least in the post facto examination of pattern data, because one can never prove that the data to test any null hypothesis were free of uncontrolled variation.

Certainly many field ecological systems are so complex that it will be difficult to achieve sufficient replication and control to parse the relative contributions of different factors. Nonlinearities, for example, can render it unlikely that all but the most local of hypotheses would not be rejected. Quinn and Dunham (1983) exemplified the situation with succession, which may be complicated by such features as differing species interactions at different densities or intransitive competitive relationships. They suggested that simply communicating results of observations and experiments designed to measure processes, rather than using the data to test hypotheses, may be a more useful approach (cf. Poole, 1988). I concede that ecological systems are complex, and often very difficult to manipulate experimentally. But it is depressing to think that there will be no criterion for abandoning a view that some factor is influential, and it may be selling ecologists short to concede at the outset that they will be unable to design unambiguous experiments. With respect to pattern data in community ecology, I would argue that the entire null hypothesis approach and the debate surrounding it have been extremely salutary in subjecting unjustifiable claims of causality to intense scrutiny, and raising typical standards of evidence.

Seaman and Jaeger (1990) frame their essay in terms of testing hypotheses, but the same potential advantage (freedom from questionable assumptions) accrues to nonparametric methods in estimating parameters. For example, an estimator of the locational difference between two distributions, and a confidence interval for that estimator, are available for the Wilcoxon rank-sum test (Hollander and Wolfe, 1973). Similarly, one can estimate Kendall's *tau* for association between two variables, and derive a confidence interval (Hollander and Wolfe, 1973). However, by their nature some nonparametric tests seem not to lend themselves to estimation of a parameter because they do not specify an easily interpreted parameter. Thus, the Smirnov two-sample test simply asks whether two parent distributions differ in either location or shape, and the test statistic itself (least upper bound of the difference between the cumulative frequency distributions of the samples) does not correspond to any easily grasped parameter.

Bayesian inference is much more readily applied to estimating parameters than to testing hypotheses. It is thus not surprising that Browner and Newman (1987) have recently argued for a Bayesian approach in medicine, on the grounds that a factor such as disease prevalence can serve as a prior distribution. Poole (1988) foresaw the development of a variety of Bayesian approaches as a healthy alternative to hypothesis-testing. However, to my knowledge there has not yet been a full consideration, in the context of decision theory, of whether the Bayesian maximax approach, or a minimax approach, or some other course of action, is the appropriate way to deal with the cost of errors.

Robustness

Use of a parametric test implies that the researcher either has some intuitive expectation of the underlying structure of the data, some sense that it conforms to a particular underlying distribution, or better, that he/she has examined the data to test for conformation. As Seaman and Jaeger (1990) point out, many ecologists do not perform such tests. Others do, but, if they find that the data violate parametric assumptions, they appeal to the notion, perhaps incorrect, that parametric tests are robust. In addition to the potential pitfalls of using parametric tests pointed out by Seaman and Jaeger (1990), it is worth noting that such tests are especially likely to be inappropriate for questions relating to distributional tails—e.g., variances and outlying percentages—in data from nonnormal distributions (Bloch et al., 1989).

One may get a mistaken impression from Seaman and Jaeger (1990) that nonparametric tests are a panacea and can be routinely applied once they become available in statistics packages. Nonparametric tests are distribution-free—not tied to a specific underlying distribution. This fact does not mean they entail no assumptions whatever. Even a test that does not assume a specific distribution can assume quite a bit about distributions. For example, the Kruskal-Wallace test assumes that the shapes of the distributions of the parent populations are the same, though the locations may differ (Conover, 1980). The Smirnov two-sample test assumes continuous distributions and tests whether shapes are identical (Conover, 1980). The Mann-Whitney-Wilcoxon test tests the hypothesis that two distributions are identical, but its power depends on the parameter $P(Y > X)$, where X and Y are drawn from the two distributions (Moses, personal communication). Locally weighted (LOESS) nonparametric regression analyses do not assume linearity, as does ordinary least-squares regression; they do not even assume monotonicity. But LOESS regression does assume normality of errors and constant variance (Cleveland and Devlin, 1988). Absence of an underlying distribution certainly removes an onerous assumption, but often there has been little study of how robust a particular nonparametric technique is to violation of its assumptions. It is thus important not to use nonparametric tests blindly. Further, many of the same problems that plague parametric analyses—small samples, nonrandom samples, poor experimental design—will also plague nonparametric ones.

Conclusion

In fact, I suspect that a large part of the debates in both medicine and ecology about statistical methods reflects concern about incorrect use and interpretation, rather than inherent properties of the methods. For example, Poole (1987) saw hypothesis-testing as a pernicious way to avoid thinking, in that it constitutes mechanical application of a decision rule: either reject or accept a hypothesis, and think no more about it. If this procedure were part of a decision rule to be applied to scientific hypotheses as opposed to repetitive, routine, and local statistical hypotheses, such a course of action would indeed be an abdication of intellectual responsibility. However, in a context more closely akin to that envisioned by decision theory, a mechanical procedure may not be inappropriate. Further, to the extent that scientific hypotheses may partly be weighed by the combined results of several statistical hypothesis tests, hypothesis-testing is not mindless. Similarly, as I noted above, the relative costs of errors depend greatly on the situation, and the researcher must assess them in this light, as there is no cut-and-dried algorithm for all situations.

In the final analysis, the decision on whether to use parametric or nonparametric methods, and how to interpret results with either type of method, should rest on the judgment of the investigator, and this judgment must be informed by statistical insight and training. Only with adequate statistical sophistication can one make a proper decision on what test to use, based on a particular set of data, a particular set of questions, and a reason for doing a statistical test. Similarly, as Bayesian methods (Poole, 1988) and likelihood ratios (Goodman and Royall, 1988) are more widely deployed, adequate statistical training will be required to prevent misuse and incorrect interpretation.

Acknowledgments.—C. Bodian, F. James, D. Meeter, and L. Moses provided insightful discussion on several aspects of this manuscript.

Literature Cited

Angell, M. 1989. Negative studies (editorial). New Engl. J. Med. 321:464–466.

Bloch, D. A., K. Lorig, B. W. Brown, Jr., and L. E. Moses. 1989. Letter to the editor. Soc. Sci. Med. 29:259–260.

Boen, J. R. 1989. Understanding *P*-value misuse (letter). Statistics in Medicine 8:1413–1414.

Bourne, W. M. 1987. No statistically significant difference?: So what? Arch. Ophthalm. 105:40–41.

Brown, J. H. 1981. Two decades of homage to Santa Rosalia: Towards a general theory of diversity. Am. Zool. 21:877–888.

Browner, W. S., and T. B. Newman. 1987. Are all significant P-values created equal? J. Am. Med. Assoc. 257:2459–2463.

Cleveland, W. S., and S. J. Devlin. 1988. Locally weighted regression: An approach to regression analysis by local fitting. J. Am. Statist. Assoc. 83: 596–610.

Connor, E. F., and D. Simberloff. 1986. Com-

petition, scientific method, and null models in ecology. Am. Sci. 74:155–162.

CONOVER, W. J. 1980. Practical Nonparametric Statistics, 2nd ed. Wiley, New York.

CRAIG, J. L. 1990. Potential for ecological restoration of islands for indigenous fauna and flora. *In* D. Towns, C. Daugherty, and I. Atkinson (Eds.), Ecological Restoration of New Zealand Islands. New Zealand Department of Conservation, Wellington. In press.

DOLBY, G. R. 1982. The role of statistics in the methodology of the life sciences. Biometrics 38: 1069–1083.

FLEISS, J. L. 1986*a*. Significance tests have a role in epidemiologic research: Reactions to A. M. Walker. Am. J. Public Health 76:559–560.

———. 1986*b*. Confidence intervals vs. significance tests: Quantitative interpretation (letter). Am. J. Public Health 76:587.

———. 1986*c*. Dr. Fleiss responds (letter). Am. J. Public Health 76:1033–1034.

FREIMAN, J. A., T. C. CHALMERS, H. SMITH, JR., AND R. R. KUEBLER. 1978. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. New Engl. J. Med. 299:690–694.

GILPIN, M. E., AND J. M. DIAMOND. 1984. Are species co-occurrences on islands non-random, and are null hypotheses useful in community ecology? Pp. 297–315. *In* D. R. Strong, D. Simberloff, L. G. Abele, and A. Thistle (Eds.), Ecological Communities: Conceptual Issues and the Evidence. Princeton University Press, Princeton, New Jersey.

GOODMAN, S. N., AND R. ROYALL. 1986. Evidence and scientific research. Am. J. Public Health 78: 1568–1574.

HOLLANDER, M., AND D. A. WOLFE. 1973. Nonparametric Statistical Methods. Wiley, New York.

KYBURG, H. E., JR. 1974. The Logical Foundations of Statistical Inference. D. Reidel, Dordrecht, Holland.

MARKS, R. G., E. K. DAWSON-SAUNDERS, J. C. BAILAR, B. B. DAN, AND J. A. VERRAN. 1988. Interactions between statisticians and biomedical journal editors. Statistics in Medicine 7:1003–1011.

MATTHEWS, D. R., AND K. MCPHERSON. 1987. Doctors' ignorance of statistics. Brit. Med. J. 294:856–857.

MOSES, L. E., J. D. EMERSON, AND H. HOSSEINI. 1984. Analyzing data from ordered categories (letter). New Engl. J. Med. 311:1383.

PIMM, S. L., H. L. JONES, AND J. DIAMOND. 1988. On the risk of extinction. Am. Nat. 132:757–785.

POOLE, C. 1987. Beyond the confidence interval. Am. J. Public Health 77:195–199.

———. 1988. Feelings and frequencies: Two kinds of probability in public health research. Am. J. Public Health 78:1531–1533.

POOLE, C., S. LANES, AND K. J. ROTHMAN. 1984. Analyzing data from ordered categories (letter). New Engl. J. Med. 311:1382.

QUINN, J. F., AND A. E. DUNHAM. 1983. On hypothesis testing in ecology and evolution. Am. Nat. 122:602–617.

ROTHMAN, K. J. 1978. A show of confidence. New Engl. J. Med. 199:1362–1363.

SCHOENER, T. W. 1986. Overview: Kinds of ecological communities—Ecology becomes pluralistic. Pp. 467–479. *In* J. Diamond and T. J. Case (Eds.), Community Ecology. Harper and Row, New York.

SEAMAN, J. W., JR., AND R. G. JAEGER. 1990. Statisticae dogmaticae: A critical essay on statistical practice in ecology. Herpetologica 46:337–346.

THOMPSON, W. D. 1987. Statistical criteria in the interpretation of epidemiological data. Am. J. Public Health 77:191–194.

TOFT, C. A., AND P. J. SHEA. 1983. Detecting community-wide patterns: Estimating power strengthens statistical inference. Am. Nat. 122:618–625.

WULFF, H. R., B. ANDERSEN, P. BRANDENHOFF, AND F. GUTTLER. 1987. What do doctors know about statistics? Statistics in Medicine 6:3–10.

Accepted: 13 February 1990
Associate Editor: David Cundall

Herpetologica, 46(3), 1990, 357–361
© 1990 by The Herpetologists' League, Inc.

REPLY TO SEAMAN AND JAEGER: AN APPEAL TO COMMON SENSE

CATHERINE A. TOFT

Department of Zoology, University of California, Davis, CA 95616, USA

SOME ecologists, usually beginning practitioners, behave as though statistics were magic. It's as if statistical procedures were a black box into which one stuffs raw numbers and out of which emerges the Truth. To criticize common practice as Seaman and Jaeger (1990) have done is healthy, and their article raises many worthy points. I worry, however, that they have left the reader, especially the beginning practi-