Bulletin of the Ecological Society of America
72 (1991)

commentary

# USE, OVERUSE, AND MISUSE OF SIGNIFICANCE TESTS IN EVOLUTIONARY BIOLOGY AND ECOLOGY

In marked contrast to what is advocated by most statisticians, most evolutionary biologists and ecologists overemphasize the potential role of significance testing in their scientific practice. Biological significance should be emphasized rather than statistical significance. Furthermore, a survey of papers shows that the literature is infiltrated by an array of misconceptions about the use and interpretation of significance tests.

Significance tests and their associated *P* values are de rigueur in research publications in evolutionary biology and ecology. Few would dare write "the litter size observed in population A is different from the one observed in population B" or "the observed sex ratio was different from 1:1" without adding the magical word "significant" or some formula like "*t* test, $P < .05$."

However, in the statistical literature, one finds statements like:

> *Overemphasis on tests of significance at the expense especially of interval estimation has long been condemned* (Cox 1977).

> *The continued very extensive use of significance tests is alarming* (Cox 1986).

> *The author believes that tests provide a poor model of most real problems, usually so poor that their objectivity is tangential and often too poor to be useful* (Pratt 1976).

> *We do not perform an experiment to find out if two varieties of wheat or two drugs are equal. We know in advance, without spending a dollar on an experiment, that they are not equal* (Deming 1975).

In this paper, I first evaluate the current state of affairs with respect to significance testing. I then explain what significance tests can be used for, or rather what they cannot be used for. Finally, I propose some rules that should improve the use of statistical methodology.

Distressingly, nothing I say in this paper is new for statisticians. However, most of the discussion has been among statisticians and biomedical researchers (see, however, Jones and Matloff 1986, Perry 1986, Krebs 1989, Wiens 1989 for a very few related comments in the ecological literature). As the literature survey shows, most biologists and other users of statistical methods seem still to be unaware that significance testing by itself sheds little light on the questions they are posing.

## *A Survey of the Literature*

> *Most readers of* The American Statistician *will recognize the limited value of hypothesis testing in the science of statistics. I am not sure that they all realize the extent to which it has become the primary tool in the religion of Statistics* (Salsburg 1985).

I have surveyed recent issues of primary evolution and ecology journals for papers with respect to statistical treatment of data (see Table 1). This sample possesses none of the characteristics that a "good" sample should have: it is not random (what is a "population" of papers?) and the papers read were not independent (each journal having its own publication policy). If errors of evaluation have been made, however, the patterns shown below are so clear that re-analysis will not modify my conclusions.

I comment only on the most common errors in the application of significance testing found in these papers, without giving any reference to a particular paper. If the problem was restricted to only a few scientists, there would be no reason for writing this paper.

By far the most common error is to confound statistical significance with biological (scientific) significance (Berkson 1942 and many others). Statements like "the two populations are significantly different relative to parameter *X* ($P = .004$)" are found with no mention of the estimated difference. The difference is perhaps *statistically* significant at the level .004, but the reader has no idea if it is *biologically* significant. Biological significance may be intimated in such remarks as



Bulletin of the Ecological Society of America 1991 72

Table 1. Literature survey. 120 papers in five journals of ecology and evolutionary biology have been read in order to evaluate the use of significance testing. The table below gives the numbers of papers found concerning the interpretation of $P$ values between .05 and .10 (i.e., nonsignificant or near significance), statements about finding significant difference without presenting estimates of the actual difference, ANOVA tables reduced to sum of squares, $F$, and $P$ values, and finally how $P$ value is believed to measure the magnitude of an effect. In addition, three papers have been found to say explicitly that the $P$ value is the probability that the null hypothesis is true, but statements like "the population means were the same ($P > .2$)," which are far more common, implicitly have the same erroneous meaning.

| | Journal* | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | Total |
| .05 < $P$ value < .10 | | | | | | |
| "Not significant" | 4 | 6 | 4 | 3 | 3 | 20 |
| "Near significance" | 5 | 2 | 4 | 1 | 4 | 17 |
| No estimate of difference | 9 | 4 | 4 | 1 | 1 | 19 |
| ANOVA: only $F$ values | 5 | 2 | 2 | 1 | 0 | 10 |
| $P$ value = effect size | 4 | 0 | 2 | 2 | 0 | 8 |
| Papers with statistical analysis | 37 | 22 | 30 | 12 | 19 | 120 |

* 1: Ecology 68(4) and Ecological Monographs 59(3); 2: Evolution 42(5) and 43(3); 3: Oikos 55(1 and 2), 56(3); 4: American Naturalist 132(4, 5, and 6); 5: Journal of Animal Ecology 58(3).

"the differences, although small, were significant." "Small" is a relative concept but may have some biological meaning. If the difference is not statistically significant, even less information is given (sometimes just "NS").

Many ANOVA tables give only the sum of squares and the $F$ values. We are not told what the mean and standard error for each treatment are. For regressions just the correlation is given (without any graph or residual analysis), and we do not know what the regression line is, or we are not given the standard error for the slope. This may be discomfiting when the Spearman rank correlation is given; we then have no idea of the shape of the relationship, as it can be linear or not. Also, a "significant" linear correlation coefficient does not imply that the underlying relationship is linear, and a nonsignificant one does not mean that the variables are independent.

The $P$ value is *not* the probability that the null hypothesis is true. This error is rarely explicit, though I have found, e.g., "the probability that the observed coefficient equals 0 appears . . ."; the error is, nonetheless, pervasive.

Significance tests may be performed before further statistical analysis (e.g., Kolmogorov-Smirnov for the normality assumption, ANOVA before pooling different samples), and the $P$ value is used to make a decision concerning further analysis, usually with the .05 limit. This is of no interest if we do not know what the costs are of making one or the other decision. For example, even a small difference between different groups can be important in subsequent analysis. In the same way, a statistical test can be robust against some departure from normality, and not against others, but the "test of normality" may give the same result in both cases.

Finally, the value of .05 has become the absolute limit between two worlds: difference on one side, equality on the other (see Table 1). "A significant difference was found between [A; the names of the variables have been changed to protect the guilty] and [B] ($P = .045$), but not between [B] and [C] ($P = .055$)," is not so rare, and I have even found "no significant difference, $P = .05$"! Expressions like: "marginally significant," "near significance," "barely significant," "bordered on statistical significance," "approached significance," are frequent when $P$ values are between .05 and .10. A $P$ value greater than .10 is nearly always "not significant."

Another problem is the tendency to give only "$P > .05$" or "NS." Recalculating the $P$ value, I have found that $P$ values described as "NS" may be equal to .06 or .07, as well as .9. Others wrongly use the test statistic to rank different effects, with no regard to the sample size or the degrees of freedom: in one paper, a $P$ value of .10 indicated an effect, .147 a little effect, and .187 no effect.

I have not found in any paper of the survey a consideration of the power of the tests used (see Toft and Shea 1983, Rotenberry and Wiens 1985 for discussion of statistical power in an ecological context). At most, some au-

thors say "the small sample size reduced the power of the test." For example, researchers looking at sex ratios seem not to be aware that the chi-square test is sensitive to sample size and that they will not detect the same differences with samples of different sizes.

To conclude this survey, statistical significance testing dominates the practice of statistics in evolutionary biology or ecology. The *P* value is often the only criterion used in decision-making, without any reference to the sample size or the experimental/survey design, or to the potential costs of such a decision. Significant difference is not said to be only significant *statistically* (vs. *biologically*). No considerations of the power of the test are made. Finally, the *P* value is often implicitly believed to be the likelihood that the null hypothesis is true.

*Tests of Significance*

> *Tests appear to many users to be a simple way to discharge the obligation to provide some statistical treatment of the data* (Roberts 1976).
>
> *In practice, of course, tests of significance are not taken seriously* (Guttman 1985).

There are many statistical papers on the meaning of *P* values and significance tests. The points of view expressed vary between rejection of them by Bayesian statisticians (e.g., Berger and Sellke 1987, Berger and Berry 1988, and references therein) and quite specific use (e.g., Cox 1977). It is thus difficult to give a presentation of significance tests that every statistician would agree with (see the discussion in Johnstone 1986; Fisher did not give a clear explanation of how to use significance tests [Kempthorne 1976, 1984], except to criticize the Neyman-Pearson approach [Fisher 1956]).

A *significance test* is "a procedure for measuring the consistency of data with a null hypothesis" (Cox 1977; see also Kempthorne [1976, 1984]). A *test statistic, t,* is a function of the data observed $X_{obs}$. The observed value of $t$, $t_{obs}$, is equal to $t(X_{obs})$. The larger the value of $t$, the stronger is the inconsistency with the null hypothesis $H_0$. If $T$ is the random variable denoting the distribution of $t$ under the null hypothesis $H_0$, then the *observed level of significance,* or $p_{obs}$, is (Cox 1977):

$$P_{obs} = \text{Prob}(T \geq t_{obs} = t(X_{obs}); H_0). \quad (1)$$

We should then specify that a *statistical* level of significance of $p_{obs}$ has been reached (and not, for example, $P < .05$).

The probability that $H_0$ is true, given the data observed $X_{obs}$, is equal to $\text{Prob}(H_0; X_{obs})$. It is not equal to the statistical level of significance, and they can be quite different (Bayesian analysis links the probability or belief that $H_0$ is true to the probability of observing the sample when $H_0$ is true; see Berger and Sellke 1987 for a review).

Traditional practice in evolutionary biology and ecology has been to calculate the *P* value for a given set of data and conclude that if it is less than .05, we have observed some "significant difference." There is nothing sacred about the value of .05 (even if, for practical reasons, statistical tables give values only for $\alpha = .05$ or .01): "No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses for a given set of data" (Fisher 1956). In decision theory (used in, e.g., industry), the level of significance is chosen according to some cost function measuring the cost of type I error (i.e., the error committed rejecting the null hypothesis when it is true). There is no such cost function in scientific practice, and accordingly no a priori reason to choose a given significance level in ecology or evolutionary biology.

Also, one cannot compare the *P* values obtained from different studies, as they are functions of sample size, of the design of the study, etc. . . (Gibbon and Pratt 1975). In addition, one cannot compare results where only the significance levels are given; such "summaries" of the data waste information. Editors may save space allowing such summaries, but with undesirable consequences.

Theoretically, one should decide on all the statistical analyses to be done *before* the data collection (Cox 1977). If the statistical analyses are chosen after "looking" at the data, then the *P* values are biased. In a similar way, if one calculates the confidence intervals of a population mean or a regression slope only when a "significant" value has been found, the confidence intervals are biased (Olshen 1973, Scheffé 1977; see Meeks and D'Agostino 1983 for an example). This practice is frequent in allometric studies.

Finally, when rejecting the null hypothesis, one must be aware that every component of the null hypothesis can be "wrong." For ex-

ample, when finding a statistically significant *t* value for a comparison of two population means, this does not necessarily imply that the population means are different. It can be rather that the sampling was deficient (i.e., nonrandom, or the samples are not independent), that the distributions are not normal or the variances not equal. Even though the latter aspects might be considered, the former one is often overlooked.

*Biological and Statistical Significance*

Standard practice is to compare two samples, and decide according to a test statistic "*M*" if the difference is significant or not, usually with .05 as a threshold. But why do we do that? If it is to say that, for example, two litter sizes are different, then we do not need statistics to say that (Deming 1975). They cannot be *exactly* equal, and to show that there is a statistically significant difference is just a matter of sample size. Also, a statement like "the litter sizes differ statistically by 0.001 young per litter" is probably biologically uninteresting.

The objective of many studies may be unclear: if we are comparing litter sizes, it is often to say more than, e.g., they are equal to 3.2 and 3.7. We do this study in relation to, for example, life history theory. According to theoretical work in this field, a difference of, say, 0.01 does not matter, but a difference of 0.5 does. The problem can then be restated in a biologically meaningful way: do the litter sizes differ by more than 0.5 or not? Note that it is quite possible that a "decision" will be impossible to make (see Tukey 1960 for an illuminating discussion of conclusions vs. decisions). We might find that the estimated difference is $0.5 \pm 0.6$, this wide confidence interval being due to, for instance, sampling size. Then, the litter sizes are not statistically significantly different, but it is quite possible that a biologically significant difference exists: further work is needed.

Only biological considerations can be used to find the amount of difference we are (or should be) looking for. To put it in another way, we have to look at the robustness of our theoretical models in order to decide how big a difference must be to be *biologically* significant. These questions should be asked before, and are dependent on the objectives of the study.

There will be no simple answer, but, to quote Tukey (1980), "finding the question is often more important than finding the answer."

*What Can Be Done: Advice to Authors, Referees and Editors*

I do not think that significance testing should be completely abandoned (as Carver [1978] or Guttman [1985] argue), and I don't expect that it will be. But I urge researchers to provide estimates, with confidence intervals: scientific advance requires parameters with known reliability estimates. Classical confidence intervals are formally equivalent to a significance test, but they convey more information. It is redundant, for example, to add "$\alpha \neq 0$; $P < .05$" to a statement like "$\alpha = 2.60 \pm 0.10$ (C.I. 95%)".

An ANOVA table that contains only *F* values is almost useless, and cannot by itself be considered as a summary of the data. The means and standard errors are needed, as well as an estimate of the differences and their confidence intervals. Editors and referees have clearly a role to play here, and they must not tolerate nearly meaningless sentences standing alone like "a significant correlation was found (Spearman rank correlation = 0.8, $P = .01$)." A graph conveys far more information, and is often all that is necessary; do not practice "statistical overkill," e.g., test for a difference between two distributions that do not overlap (this has been found in one paper sampled; see also Chatfield 1985).

Another important aspect, although indirectly related to significance testing, concerns the planning stage (e.g., sample size) of an experimental/observational study. This stage depends on the biological objectives that must be specified before data are collected. Differences that are biologically significant should be decided on before the study, and not after.

Finally, pay more importance to the sampling processes that are part of the "null hypothesis": random sampling, and independence (Felsenstein 1985, Kruskal 1988). Do not force a biological problem into an inappropriate mathematical framework (e.g., classical tests assuming independence) because no appropriate one exists. Following the golden rule of applied mathematics, it is always better to give an approximate answer to the right question than a precise answer to the wrong question.

*Acknowledgments*

I have expressed personal views; this paper nevertheless has its origins in numerous discussions with Wilhelm Falck, Rolf A. Ims, Larry Kirkendall, Harald Steen, Nils Chr. Stenseth, Jogeir Stokland, and John Wiens. I thank them warmly for having patiently endured my enthusiasm for biometry. William Z. Lidicker and two anonymous referees have also provided useful comments.

*Literature Cited*

Berger, J. O., and D. A. Berry. 1988. Statistical analysis and the illusion of objectivity. American Scientist **76**:159–165.

Berger, J. O., and T. Sellke. 1987. Testing a point null hypothesis: the irreconcilability of P values and evidence. Journal of the American Statistical Association **82**: 112–122.

Berkson, J. 1942. Tests of significance considered as evidence. Journal of the American Statistical Association **37**:325–335.

Carver, R. P. 1978. The case against statistical testing. Harvard Educational Review **48**:378–399.

Chatfield, C. 1985. The initial examination of data. Journal of the Royal Statistical Society, Series A, **148**:214–253.

Cox, D. R. 1977. The role of significance tests. Scandinavian Journal of Statistics **4**:49–70.

———. 1986. Some general aspects of the theory of statistics. International Statistical Review **54**:117–126.

Deming, W. E. 1975. On probability as a basis for action. American Statistician **29**: 146–152.

Felsenstein, J. 1985. Phylogenies and the comparative method. American Naturalist **125**:1–15.

Fisher, R. A. 1956. Statistical methods and scientific inference. Oliver and Boyd, London, England.

Gibbon, J. D., and J. W. Pratt. 1975. P-values: interpretation and methodology. American Statistician **29**:20–25.

Guttman, L. 1985. The illogic of statistical inference for cumulative science. Applied Stochastic Models and Data Analysis **1**: 3–9.

Johnstone, D. J. 1986. Tests of significance in theory and practice. Statistician **35**: 491–504.

Jones, D., and N. Matloff. 1986. Statistical hypothesis testing in biology: a contradiction in terms. Journal of Economic Entomology **79**:1156–1160.

Kempthorne, O. 1976. Of what use are tests of significance and tests of hypotheses. Communication in Statistics, Series A, **5**: 763–777.

———. 1984. Statistical methods and science. Pages 287–308 *in* P. S. R. S. Rao and J. Sedransk, editors. W. G. Cochran's impact on statistics. John Wiley and Sons, New York, New York, USA.

Krebs, C. J. 1989. Ecological methodology. Harper and Row, New York, New York, USA.

Kruskal, W. 1988. Miracles and statistics: the casual assumption of independence. Journal of the American Statistical Association **83**:1045–1054.

Meeks, S. L., and R. B. D'Agostino. 1983. A note on the use of confidence limits following rejection of a null hypothesis. American Statistician **37**:134–136.

Olshen, R. A. 1973. The conditional level of the F-test. Journal of the American Statistical Association **68**:692–698.

Perry, J. N. 1986. Multiple-comparison procedures: a dissenting view. Journal of Economic Entomology **79**:1149–1155.

Pratt, J. W. 1976. A discussion of the question: for what use are tests of hypotheses and tests of significance. Communication in Statistics, Series A, **5**:779–787.

Roberts, H. V. 1976. For what use are tests of hypotheses and tests of significance. Communication in Statistics, Series A, **5**: 753–761.

Rotenberry, J. T., and J. A. Wiens. 1985. Statistical power analysis and community-wide patterns. American Naturalist **125**:164–168.

Salsburg, D. S. 1985. The religion of statistics as practiced in medical journals. American Statistician **39**:220–223.

Scheffé, H. 1977. Note on a reformulation of the S-method of multiple comparison. Journal of the American Statistical Association **72**:143–146.

Toft, C. A., and P. J. Shea. 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. American Naturalist **122**:618–625.

Tukey, J. W. 1960. Conclusions vs. decisions. Technometrics 2:423–433.
———. 1980. We need both exploratory and confirmatory. American Statistician 34: 23–25.
Wiens, J. A. 1989. The ecology of bird communities. Cambridge University Press, Cambridge, England.

Nigel G. Yoccoz
Laboratoire de Biométrie
Université Claude Bernard Lyon 1
69622 Villeurbanne Cédex
France
Tel. (33) 72 44 81 42
FAX: (33) 72 44 84 66
and
Division of Zoology
Department of Biology
University of Oslo
P.O. Box 1050
Blindern, N-0316 Oslo 3
Norway

## QUESTIONS FREQUENTLY ASKED OF ECOLOGY PROGRAM OFFICERS AT NSF

Program Officers at the National Science Foundation spend a large part of their time answering questions about operations associated with receiving, reviewing, and evaluating proposals submitted by investigators. Many of the questions are specific to a particular proposal, but others are more general and reveal a lack of information and a variety of misconceptions about the processes that take place within the Programs. These problems arise from the size, diversity, and changing nature of the investigator community and NSF, and the limited resources NSF personnel have to provide information to investigators.

In an attempt to promote understanding of the evaluation and decision-making process in the Ecology Program, I will address several frequently asked questions. The questions chosen are those that are commonly asked, or that indicate the greatest misunderstanding about the Program. The information presented pertains specifically to typical research proposals in the Ecology Program, although it is generally accurate for other programs within the Division of Biotic Systems and Resources (BSR). In addition to typical research proposals, Programs within the Division are involved with almost two dozen other types of proposals across the Foundation (e.g., international, minority, or undergraduate programs), and many of these engender their own sets of questions. I have chosen to discuss only the most common type of research proposal, and I will address the questions in the order that an investigator, preparing a new proposal, might encounter them.

A brief word is in order about the organization of NSF and about the review process. The most common contact a scientist has with NSF is through Program Officers because the Program is the operational unit in NSF. The Ecology Program usually has a temporary Program Director (1–3 years) and a permanent Associate Program Director. The Program is in the Division of Biotic Systems and Resources, which also houses the Programs of Systematic Biology, Population Biology and Physiological Ecology, Ecosystem Studies, and Biological Research Resources. The Division is within a Directorate that includes other biological sciences as well as social and behavioral sciences. This Directorate (Biological, Behavioral, and Social Sciences; BBS) is one of eight within the Foundation.

Standard research proposals currently have two target dates a year (15 June and 15 December; these are discussed below). Proposals are received in a central processing section and forwarded to the Programs within 2–5 weeks of receipt. Program Officers then send the proposals to ad hoc reviewers and panelists, and the support staff collate information and return reviews into the official "jacket" (file). Approximately 4 months after a target date an Advisory Panel, assembled by the Program Officers, meets to discuss each proposal (except those few submitted by current or recent panelists and Program Officers). A panel is composed of approximately 15 researchers (depending on proposal load and the distribution of proposals by subdiscipline) and panelists usually serve for 3 years or six