NULL HYPOTHESIS TESTING: PROBLEMS, PREVALENCE, AND AN ALTERNATIVE

DAVID R. ANDERSON,[1,2] Colorado Cooperative Fish and Wildlife Research Unit, Room 201 Wagar Building, Colorado State University, Fort Collins, CO 80523, USA
KENNETH P. BURNHAM,[1] Colorado Cooperative Fish and Wildlife Research Unit, Room 201 Wagar Building, Colorado State University, Fort Collins, CO 80523, USA
WILLIAM L. THOMPSON, U.S. Forest Service, Rocky Mountain Research Station, 316 E. Myrtle St., Boise, Idaho 83702, USA

Abstract: This paper presents a review and critique of statistical null hypothesis testing in ecological studies in general, and wildlife studies in particular, and describes an alternative. Our review of *Ecology* and the *Journal of Wildlife Management* found the use of null hypothesis testing to be pervasive. The estimated number of *P*-values appearing within articles of *Ecology* exceeded 8,000 in 1991 and has exceeded 3,000 in each year since 1984, whereas the estimated number of *P*-values in the *Journal of Wildlife Management* exceeded 8,000 in 1997 and has exceeded 3,000 in each year since 1994. We estimated that 47% (SE = 3.9%) of the *P*-values in the *Journal of Wildlife Management* lacked estimates of means or effect sizes or even the sign of the difference in means or other parameters. We find that null hypothesis testing is uninformative when no estimates of means or effect size and their precision are given. Contrary to common dogma, tests of statistical null hypotheses have relatively little utility in science and are not a fundamental aspect of the scientific method. We recommend their use be reduced in favor of more informative approaches. Towards this objective, we describe a relatively new paradigm of data analysis based on Kullback-Leibler information. This paradigm is an extension of likelihood theory and, when used correctly, avoids many of the fundamental limitations and common misuses of null hypothesis testing. Information-theoretic methods focus on providing a strength of evidence for an a priori set of alternative hypotheses, rather than a statistical test of a null hypothesis. This paradigm allows the following types of evidence for the alternative hypotheses: the rank of each hypothesis, expressed as a model; an estimate of the formal likelihood of each model, given the data; a measure of precision that incorporates model selection uncertainty; and simple methods to allow the use of the set of alternative models in making formal inference. We provide an example of the information-theoretic approach using data on the effect of lead on survival in spectacled eider ducks (*Somateria fischeri*). Regardless of the analysis paradigm used, we strongly recommend inferences based on a priori considerations be clearly separated from those resulting from some form of data dredging.

JOURNAL OF WILDLIFE MANAGEMENT 64(4):912–923

Key words: AIC, Akaike weights, *Ecology*, information theory, *Journal of Wildlife Management*, Kullback-Leibler information, model selection, null hypothesis, *P*-values, significance tests.

Theoretical and applied ecologists continually strive for rigorous, objective approaches for making valid inference concerning science questions. The dominant, traditional approach has been to frame the question in terms of 2 contrasting statistical hypotheses: 1 representing no difference between population parameters of interest (i.e., the null hypothesis, H_o) and the other representing either a unidirectional or bidirectional alternative (i.e., the alternative hypothesis, H_a). These hypotheses basically correspond to different models. For example, when comparing 2 groups of interest, the assumption is that they are from the same population so that the difference between their true means is 0 (i.e., H_o is $\mu_1 - \mu_2 = 0$, or $\mu_1 = \mu_2$). A test statistic is computed from sample data and compared to its hypothesized null distribution to assess the consistency of the data with the null hypothesis. More extreme values of the test statistic suggest that the sample data are not consistent with the null hypothesis. A substantially arbitrary level (α) is often preset to serve as a cutoff (i.e., the basis for a decision) for statistically significant versus statistically nonsignificant results. This procedure has various names, including null hypothesis testing, significance testing, and null hypothesis significance testing. In fact, this procedure is a hybridization of Fisher's (1928) significance testing and Neyman and Pearson's (1928, 1933) hypothesis testing (Gigerenzer et al. 1989, Goodman 1993, Royall 1997).

There are a number of problems with the application of the null hypothesis testing ap-

[1] Employed by U.S. Geological Survey, Division of Biological Resources.

[2] E-mail: anderson@picea.cnr.colostate.edu

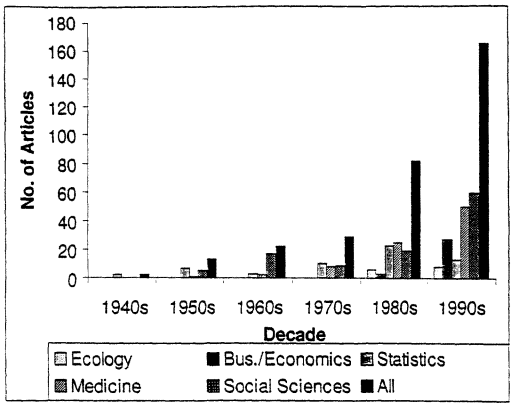

Fig. 1. Sample of articles, based on an extensive sampling of the literature by decade, in various disciplines that questioned the utility of null hypothesis testing in scientific research. Numbers shown for the 1990s were extrapolated based on sample results from volume years 1990–96.

proach, some of which we present herein (Carver 1978, Cohen 1994, Nester 1996). Although doubts among statisticians concerning the utility of null hypothesis testing are hardly new (Berkson 1938, 1942; Yates 1951; Cox 1958), criticisms have increased in the scientific literature in recent years (Fig. 1). Over 300 references now exist in the scientific literature that warn of the limitations of statistical null hypothesis testing. A list of citations is located at *http://www.cnr.colostate.edu/~anderson/thompson1.html* and *http://www.cnr.colostate.edu/~anderson/nester.html*. The former website also includes a link to a list of papers supporting the use of tests. We believe that few wildlife biologists and ecologists are aware of the debate regarding null hypothesis testing among statisticians. Discussion and debate have been particularly evident in the social sciences, where at least 3 special features (*Journal of Experimental Education* 61(4); *Psychological Science* 8(1); *Research in the Schools* 5(2)) and 2 edited books (Morrison and Henkel 1970, Harlow et al. 1997) have debated the utility of null hypothesis tests in scientific research. The ecological sciences have lagged behind other disciplines with respect to awareness and discussion of problems associated with null hypothesis testing (Fig. 1; Yoccoz 1991; Cherry 1998; Johnson 1999).

We present information concerning prevalence of null hypothesis testing by reviewing papers in *Ecology* and the *Journal of Wildlife Management*. We chose *Ecology* because it is widely considered to be the premier journal in the field and hence, should be indicative of statistical usage in the ecological field as a whole. We chose the *Journal of Wildlife Management* as an applied journal for comparison. We review theoretical or philosophical problems with the null hypothesis testing approach as well as its common misuses. We offer a practical, theoretically sound alternative to null hypothesis testing and provide an example of its use. We conclude with our views concerning data analysis and the presentation of scientific results, as well as our recommendations for changes in editorial and review policies of biological and ecological journals.

PROBLEMS WITH NULL HYPOTHESIS OR SIGNIFICANCE TESTING

The fundamental problem with the null hypothesis testing paradigm is not that it is wrong (it is not), but that it is uninformative in most cases, and of relatively little use in model or variable selection. Statistical tests of null hypotheses are logically poor (e.g., the arbitrary declaration of significance). Berkson (1938) was one of the first statisticians to object to the practice.

The most curious problem with null hypothesis testing, as the primary basis for data analysis and inference, is that nearly all null hypotheses are false on a priori grounds (Johnson 1995). Consider the null H_o: $\theta_o = \theta_1 = \theta_2 = \ldots = \theta_5$, where θ_o is an expected control response and the others are ordered treatment responses (e.g., different nitrogen levels applied to agricultural fields). This null hypothesis is almost surely false as stated. Even the application of sawdust would surely make some difference in response. The rejection of this strawman hardly advances science (Savage 1957), nor does it give meaningful insights for conservation, planning, management, or further research. These issues should properly focus on the estimation of effects or differences and their precision and not on testing a trivial (uninformative) null. Other general examples of a priori false null hypotheses include (1) H_o: $\mu_c = \mu_t$ (mean growth rate is equal in control vs. aluminum-treated bullfrog, *Rana catesbeiana*); (2) H_o: $S_{jC} = S_{jD}$ (survival probability in week j is the same for control vs. lead-dosed gull chicks, *Larus spp.*), and (3) H_o: $\rho_{yx} = 0$ (zero correlation between variables Y and X). Johnson (1999) provided additional examples of null hypotheses that are clearly false before any testing was conducted; the focus of such investigations should properly

estimate the size of effects. Statistical tests of such null hypotheses, whether rejected or not, provide little information of scientific interest and, in this respect, are of little practical use in the advancement of knowledge (Morrison and Henkel 1969).

A much more well known, but ignored, issue is that a particular α-level is without theoretical basis and is therefore arbitrary except for the adoption of conventional values (commonly 0.1, 0.05 or 0.01, but often 0.15 in stepwise variable selection procedures). Use of a fixed α-level arbitrarily classifies results into biologically meaningless categories significant and nonsignificant and is relatively uninformative. This Neyman-Pearson approach is an arbitrary reject or not reject decision when the substantive issue is one of strength of evidence concerning a scientific issue (Royall 1997) or estimation of size of an effect.

Consider an example from a recent issue of the *Wildlife Society Bulletin,* "Response rates did not vary among areas (χ^2 = 16.2, 9 df, P = 0.06)." Thus, the null must have been $R_1 = R_2 = R_3 = \ldots = R_{10}$; however, no estimates of the response rates (the $\hat{R}_j$) or their associated precision or even sample size were provided. Had the P-value been 0.01 lower, the conclusion would have been that significant differences were found and the estimates $\hat{R}_i$ and their precision given. Alternatively, had the arbitrary α level been 0.10 initially, the result would have been quite different (i.e., response rates varied among areas, χ^2 = 16.2, 9 df, P = 0.06). Here, as in most cases, the null hypothesis was false on a priori grounds. Many examples can be found where contradictory or nonsensical results have been reported (Johnson 1999). Legal hearings concerning scientific issues are unproductive and lead to confusion when 1 party claims significance (based on α = 0.1), whereas the opposing party argues nonsignificance (based on α = 0.05).

The cornerstone of null hypothesis testing, the P-value, has problems as an inferential tool that stem from its very definition, its application in observational studies, and its interpretation (Cherry 1998, Johnson 1999). The P-value is defined as the probability of obtaining a test statistic at least as extreme as the observed one, conditional on the null hypothesis being true. There are 2 important points to consider about this definition. First, a P-value is based not only on the observed result (the data collected), but also on less likely, unobserved results (data sets never collected) and therefore overstates the evidence against the null hypothesis (Berger and Sellke 1987, Berger and Berry 1988). A P-value is more of a statement about the events that never occurred than it is a concise statement of the evidence from an actual observed event (i.e., the data). Bayesians (people making statistical inferences using Bayes' theorem; Gellman et al. 1995) find this property of P-values objectionable; they tend to avoid null hypothesis testing in their paradigm.

A second consequence of its definition is that a P-value is explicitly conditional on the null hypothesis (i.e., it is computed based on the distribution of the test statistic assuming the null hypothesis is true). The null distribution of the test statistic (e.g., often assumed to be F, t, z, or χ^2) may closely match the actual sampling distribution of that statistic in strict experiments, but this property does not hold in observational studies. In these latter studies, the distribution of the test statistic is unknown because randomization was not done, and hence there are problems with confounding factors (both known and unknown). In observational studies, the distribution of the test statistic under the null hypothesis is not deducible from the study design. Consequently, the form of the distribution is not known, only naively assumed, which makes interpretation of test results problematic.

It has long been known and criticized that the P-value is dependent on sample size (Berkson 1938). One can always reject a null hypothesis with a large enough sample, even if the true difference is trivially small. This points to the difference between statistical significance and biological importance raised by Yoccoz (1991) and many others before and since. Another problem is that using a fixed α-level (e.g., 0.1) to decide to reject or not reject the null hypothesis makes little sense as sample size increases. Here, even when the null hypothesis is true and sample size is infinite, a Type I error (rejecting a null that is true) still occurs with probability α (e.g., 0.1), and therefore this approach is not consistent (theoretically, α should go to zero as n goes to infinity). Still another issue is that the P-value does not provide information about either the size or the precision of the estimated effect. The solution here is to merely present the estimate of effect size and a measure of its precision.

A pervasive problem in the use of *P*-values is in their misinterpretation as evidence for either the null or alternative hypothesis (see Ellison 1996 for recent examples of such misuse). The proper interpretation of the *P*-value is based on the probability of the data given the null hypothesis, not the converse. We cannot accept or prove the null hypothesis, only fail to reject it. The *P*-value cannot validly be taken as the probability that the null hypothesis is true, although this is often the interpretation given. Similarly, the magnitude of the *P*-value does not indicate a proper strength of evidence for the alternative hypothesis (i.e., the probability of H_a, given the data), but rather the degree of consistency (or inconsistency) of the data with H_o (Ellison 1996). Phrases such as highly significant (often denoted as ** or even ***) only reinforce this error in interpretation of *P*-values (Royall 1997).

Presentation of only *P*-values also limits the effectiveness of (future) meta-analyses. There is a strong publication bias whereby only significant *P*-values tend to get reported (accepted) in the literature (Hedges and Olkin 1985:285–290, Iyengar and Greenhouse 1988). Thus, the published literature is itself biased in favor of results arbitrarily deemed significant. It is important to present parameter estimates (effect size) and their precision from any well designed study, regardless of the outcome; these become the relevant data for a meta-analysis.

A host of other problems exist in the null hypothesis testing paradigm, but we will mention only a few. We generally lack a rigorous theory for testing null hypotheses when a model contains nuisance parameters (e.g., sampling probabilities in capture-recapture studies). The distribution of the likelihood ratio test statistic between models that are not nested is unknown and this makes comprehensive analysis problematic. Given the prevalence of null hypothesis testing, we warn against the invalid notion of post-hoc or retrospective power analysis (Goodman and Berlin 1994, Gerard et al. 1998) and note that this practice has become more common in recent years.

The central issues here are twofold. First, scientists are fundamentally interested in estimates of the magnitude of the differences and their precision, the so-called effect size. Is the difference trivial, small, medium, or large? Is this difference biologically meaningful? This is an estimation problem. Second, one often wants to know if the differences are large enough to justify inclusion in a model to be used for inference in more complex science settings. This is a model selection problem. These central issues that further our understanding and knowledge are not properly addressed with statistical hypothesis testing. Statistical science is much more than merely significance testing, even though many statistics courses are still offered with an unfounded emphasis on null hypothesis testing (Schmidt 1996). Many statisticians question the practical utility of hypothesis testing (i.e., the arbitrary α-levels, the false null hypotheses being tested, and the notion of significance) and stress the value of estimation of effect size and associated precision (Goodman and Royall 1988, Graybill and Iyer 1994:35).

PREVALENCE OF FALSE NULL HYPOTHESES AND *P*-VALUES

We randomly sampled 20 papers in the *Articles* section from each volume of *Ecology* for years 1978–97 to assess the prevalence of trivial null hypotheses and associated *P*-values in published ecological studies. We then randomly sampled 20 papers from each volume of the *Journal of Wildlife Management* (JWM) for years 1994–98 for comparison. In each sampled article, we noted whether the null hypotheses tested seemed at all plausible. In addition, we counted the number of *P*-values and equivalent symbols, such as statistics with superscripted asterisks or comparisons specifically marked nonsignificant. We tallied the number of cases where only a *P*-value was given (some papers also provided the test statistic, degrees of freedom, or sample size), without an estimate of effect size, its sign or its precision, even in an associated table, for papers appearing in the JWM during the 1994–98 period. However, our counts did not include comparisons that were both nonsignificant and unlabeled or unspecified, nor did they include all possible statistical comparisons or tests. Consequently, ours is an underestimate of the total number of statistical tests and associated *P*-values contained within each article.

In the 347 sampled articles in *Ecology* containing null hypothesis tests, we found few examples of null hypotheses that seemed biologically plausible. Perhaps 5 of 95 articles in JWM contained ≥ 1 null hypothesis that could be considered a plausible alternative. Only 2 of 95 articles in JWM incorporated biological importance into the interpretations of results, the re-

Table 1. Median, mean (SE), and range of the number of *P*-values per article, and estimated total (SE) number of *P*-values per year, based on a random sample of 20 papers each year from the *Articles* section of *Ecology* for 1978-97.

Volume year	Total articles	Estimated no. of *P*-values per article			Estimated yearly total (SE) *P*-values
		Median	$\bar{x}$ (SE)	Range	
1978	131	1.5	10 (3)	0–68	1,310 (393)
1979	122	8	18 (5)	0–114	2,196 (610)
1980	154	3.5	12 (4)	0–76	1,848 (616)
1981	153	11.5	16 (4)	0–84	2,448 (612)
1982	179	15	23 (8)	0–183	4,117 (1,432)
1983	157	5	14 (3)	0–48	2,198 (471)
1984	178	11	32 (14)	0–317	5,696 (2,492)
1985	189	25.5	25 (4)	0–94	4,725 (756)
1986	158	16	32 (7)	0–109	5,056 (1,106)
1987	190	25	32 (4)	0–79	6,080 (760)
1988	182	24	36 (7)	1–155	6,552 (1,274)
1989	159	35	48 (9)	3–190	7,632 (1,431)
1990	198	21.5	30 (6)	0–91	5,940 (1,188)
1991	182	40	44 (8)	0–204	8,008 (1,456)
1992	181	16.5	22 (3)	0–53	3,982 (543)
1993	203	8.5	28 (7)	0–125	5,684 (1,421)
1994	189	17	26 (5)	0–96	4,914 (945)
1995	203	30.5	31 (5)	0–85	6,293 (1,015)
1996	172	23	37 (10)	1–208	6,364 (1,720)
1997	183	26	42 (11)	0–208	7,686 (2,013)

mainder merely used statistical significance. In the vast majority of cases, the null hypotheses we found in both journals seemed to be obviously false on biological grounds even before these studies were undertaken. A major research failing seems to be the exploration of uninteresting or even trivial questions. Common examples included null hypotheses assuming survival probabilities were the same between juveniles and adults of a species, assuming no correlation or relationship existed between variables of interest, assuming density of a species remained the same across time, assuming net primary production rates were constant across sites and years, and assuming growth rates did not differ among individuals or species.

We estimate that there have been a minimum of several thousand *P*-values appearing in every volume of *Ecology* (Table 1) and JWM (Table 2) in recent years. Given the conservatism of our counting procedure, the number of null hypothesis tests that were actually performed in each study was probably much larger. Approximately 47% (SE = 3.9%) of the *P*-values that we counted in JWM appeared alone, without estimated means, differences, effect sizes, or associated measures of precision. Such results, we maintain, are particularly uninformative (e.g., not even the sign of the difference being indicated). The key problem here is the general failure to explore more relevant questions and to report informative summary statistics (e.g., estimates of effect size and their precision), even when significance was found. The secondary problem is not recognizing the arbitrariness of α, hence perpetuating an arbitrary classification of results as significant or not significant.

A PRACTICAL ALTERNATIVE TO NULL HYPOTHESIS TESTING

We advocate Chamberlin's (1890, 1965) concept of multiple working hypotheses rather than a single statistical null vs. an alternative—this seems like superior science. However, this approach leads to the multiple testing problem in statistical hypothesis testing, and arbitrariness in the choice of α-level and of which hypothesis to serve as the null. Although commonly used in practice, significance testing is a poor approach to model selection and variable selection in regression analysis, discriminant function analysis, and similar procedures (Akaike 1974, McQuarrie and Tsai 1998:427–428).

Akaike (1973, 1974) developed data analysis procedures that are now called *information theoretic* because they are based on Kullback-Leibler (1951) information. Kullback-Leibler information is a fundamental quantity in the sciences and has earlier roots back to Boltzmann's concept of entropy. The Kullback-Leibler information between conceptual truth, *f*, and ap-

Table 2. Median, mean (SE), and range of the number of *P*-values per article, and estimated total (SE) number of *P*-values per year, based on a random sample of 20 papers (excluding *Invited Papers* and *Comment/Reply* articles) each year from the *Journal of Wildlife Management* for years 1994–98.

Year	Articles	Estimated number of *P*-values per article			Estimated yearly total (SE) *P*-values
		Median	$\bar{x}$ (SE)	Range	
1994	101	21	32 (8)	0–139	3,232 (808)
1995	106	24	37 (10)	0–171	3,922 (1,060)
1996	104	21	54 (24)	0–486	5,616 (2,496)
1997	150	24	56 (16)	0–263	8,400 (2,400)
1998	166	28	31 (6)	1–122	5,146 (996)

proximating model g is defined for continuous functions as the integral

$$I(f, g) = \int f(x)\log_e\left(\frac{f(x)}{g(x\,|\,\theta)}\right) dx,$$

where f and g are n-dimensional probability distributions. Kullback-Leibler information, denoted $I(f, g)$, is the information lost when model g is used to approximate truth, f. The right hand side looks difficult to understand, however it can be viewed as a statistical expectation of the natural logarithm of the ratio of f (full reality) to g (approximating model). That is, Kullback-Leibler information could be written as

$$E_f\left[\log_e\left(\frac{f(x)}{g(x\,|\,\theta)}\right)\right],$$

where the expectation is taken with respect to full reality, f. Using the property of logarithms, this expression can be further simplified as the difference between 2 expectations,

$$I(f, g) = E_f[\log_e(f(x))] - E_f[\log_e(g(x\,|\,\theta))].$$

Clearly, full reality is unknown, but it is fixed across models, thus a further simplification can be written as

$$I(f, g) = C - E_f[\log_e(g(x|\theta))],$$

where the expectation of the logarithm of full reality drops out into a simple scaling constant, C. Thus, the focus in model selection is on the term $E_f[\log_e(g(x|\theta))]$.

One seeks an approximating model (hypothesis) that loses as little information as possible about truth; this is equivalent to minimizing $I(f, g)$, over the set of models of interest (we assume there are R a priori models, each representing an hypothesis, in the candidate set). Obviously, Kullback-Leibler information, by itself, will not aid in data analysis as both truth (f) and the parameters (θ) are unknown to us.

Model Selection Criteria

Akaike (1973) found a formal relationship between Kullback-Leibler information (a dominant paradigm in information and coding theory) and maximum likelihood (the dominant paradigm in statistics; deLeeuw 1992). This finding makes it possible to combine estimation and model selection under a single theoretical framework—optimization. Akaike's breakthrough was deriving an estimator of the expected, relative Kullback-Leibler information, based on the maximized log-likelihood function. This led to Akaike's information criterion (AIC),

$$\text{AIC} = -2\log_e(\ell(\hat{\theta}|data)) + 2K,$$

where $\log_e\ell(\hat{\theta}|data)$ is the value of the maximized log-likelihood over the unknown parameters (θ), given the data and the model, and K is the number of parameters estimated in that approximating model. There is a simple transformation of the estimated residual sum of squares (RSS) to obtain the value of $\log_e(\ell(\hat{\theta}|data))$ when using least squares, rather than likelihood methods. The value of AIC for least squares models is merely,

$$\text{AIC} = n\cdot\log_e(\hat{\sigma}^2) + 2K,$$

where n is sample size and $\hat{\sigma}^2 = RSS/n$. Such quantities are easy to compute once the RSS values for each model are available using standard computer software.

Assuming a set of a priori candidate models (hypotheses) has been defined and well supported, AIC is computed for each of the approximating models in the set (i.e., g_i, $i = 1, 2, \ldots, R$). The model where AIC is minimized is selected as best for the empirical data at hand. This concept is simple, compelling, and is based on deep theoretical foundations (i.e., Kullback-Leibler information). The AIC is not a test in any sense: no single hypothesis (i.e., model) is

made to be the null, no arbitrary α level is set, and no notion of significance is needed. Instead, there is the concept of a best inference, given the data and the set of a priori models, and further developments provide a strength of evidence for each of the models in the set.

It is important to use a modified criterion (called AIC_c) when K is large relative to sample size n,

$$AIC_c = -2\log(\mathcal{L}(\hat{\theta}\,|\,\text{data})) + 2K + \frac{2K(K+1)}{(n-K-1)},$$

and this should be used unless $n/K >$ about 40 (Burnham and Anderson 1998). As sample size increases, $AIC = AIC_c$, thus, if in doubt, always use AIC_c as the final term is also trivial to compute. Both AIC and AIC_c are estimates of expected (relative) Kullback-Leibler information and are useful in the analysis of real data in the "noisy" sciences.

Ranking Models

The evidence for each of the alternative models can best be done by rescaling AIC values such that the model with the minimum AIC (or AIC_c) has a value of 0, i.e.,

$$\Delta_i = AIC_i - \text{minAIC}.$$

The Δ_i values are easy to interpret and allow a quick strength of evidence comparison and scaled ranking of candidate models. The larger the Δ_i, the less plausible is the fitted model i as being the best approximating model in the candidate set. It is generally important to know which model (biological hypothesis) is ranked second best as well as some measure of its standing with respect to the best model. Such ranking and scaling can be done easily with the Δ_i values.

Likelihood of a Model, Given the Data

The simple transformation $\exp(-\frac{1}{2}\Delta_i)$, for $i = 1, 2, \ldots, R$, provides the likelihood of the model, given the data: $\mathcal{L}(g_i|data)$. These are functions in the same sense that $\mathcal{L}(\theta|data, g_i)$ is the likelihood of the parameters θ, given the data (x) and the model (g_i). It is convenient to normalize these values such that they sum to 1, as

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{r=1}^{R}\exp\left(-\frac{1}{2}\Delta_r\right)}.$$

The w_i, called Akaike weights, can be interpreted as approximate probabilities that model i is, in fact, the Kullback-Leibler best model in the set of models considered. Akaike weights are a measure of the weight of evidence that model i is the actual Kullback-Leibler best model in the set. The relative likelihood of model i versus model j is just w_i/w_j. Inference here is conditional on both the data and the set of a priori models considered.

Unconditional Sampling Variance

Typically, estimates of sampling variance are conditional on a given model. When model selection has been done, a variance component due to uncertainty in model selection should be incorporated into estimates of precision such that these estimates are unconditional on the selected model, but still conditional on the models in the set. An estimator of the unconditional variance for the parameter θ from the selected (best) model is,

$$\widehat{\text{var}}(\hat{\theta}) = \left[\sum_{i=1}^{R} w_i\sqrt{\widehat{\text{var}}(\hat{\theta}_i\,|\,g_i) + (\hat{\theta}_i - \bar{\hat{\theta}})^2}\right]^2,$$

where

$$\bar{\hat{\theta}} = \sum_{i=1}^{R} w_i\hat{\theta}_i.$$

This estimator, from Buckland et al. (1997), includes a term for the conditional sampling variance, given model g_i (denoted as $\widehat{\text{var}}(\hat{\theta}_i|g_i)$ here) and a variance component for model selection uncertainty, $(\hat{\theta}_i - \bar{\hat{\theta}})^2$. These variance components are multiplied by the Akaike weights, which reflect the degree of model importance. Precision of the estimated parameters can be assessed using this unconditional variance with the usual 95% confidence interval, $\hat{\theta} \pm 2\widehat{se}(\hat{\theta})$, or intervals based on log- or logit-transformation (Burnham et al. 1987: 211–214), profile likelihood (Royall 1997:158–159), or bootstrap (Efron and Tibshirani 1993) methods.

Multi-model Inference

Rather than base inferences on a single selected best model from an a priori set of models, inference can be based on the entire set of models (multi-model inference, MMI). Such inferences can be made if a parameter, θ, is in common over all models (as θ_i in model g_i), or our goal is prediction. Then by using the

Table 3. Example of multi-model inference based on models and results presented in Tables 3 and 4 in Grand et al. (1998) on effect of lead poisoning on spectacled eider annual survival probability; ϕ denotes annual survival probability; subscripts *e* and *u* denoted exposed or unexposed to lead, respectively. From each model we get an estimate of lead-effect on survival ($\hat{effect} = \hat{\phi}_u - \hat{\phi}_e$), and estimated conditional standard error, $\hat{se}(\hat{effect}|g_i)$, given the model; see text for further explanation.

Model g_i[a]	K_i	Δ_i	w_i	Lead effect $\hat{\phi}_u - \hat{\phi}_e$	$\hat{se}(\hat{effect}\|g_i)$
$\{\phi_l\ p.\}$	3	0.00	0.673	0.337	0.105
$\{\phi_{s+l}\ p.\}$	4	2.07	0.239	0.335	0.148
$\{\phi_{s*l}\ p.\}$	5	4.11	0.086	0.330	0.216
$\{\phi.\ p.\}$	2	12.71	0.001	0.000	0.000
$\{\phi_s\ p.\}$	3	14.25	0.001	0.000	0.000
model averaged				0.335	0.125

[a] Notation follows that of Lebreton et al. (1992): s = site; l = lead exposure; . = constant across years, s and l; p = recapture probability.

weighted average, $\hat{\bar{\theta}} = \Sigma w_i \hat{\theta}_i$, we are basing point inference on the entire set of models. This approach has both practical and philosophical advantages (Gardner and Altman 1986, Henderson 1993, Goodman and Berlin 1994). Where a model-averaged estimator can be used, it often has better precision and reduced bias compared to the estimator of that parameter from only the selected best model.

An Example: Lead-Effect on Spectacled Eider Survival

Grand et al. (1998) evaluated the effect of lead exposure on annual survival probability (ϕ) of female spectacled eiders. Data were from 3 years of a larger capture-recapture study at 2 sites on the Yukon-Kuskokwim Delta in Alaska. Nesting female eiders were captured in May–June of 1994–96. At capture in 1994 and 1995, blood was drawn to use in determining lead exposure (assumed to be from ingested lead pellets). Grand et al. (1998) classified each female either as exposed or unexposed to lead. For analysis of lead-effect on annual survival they used 5 models determined a priori (but partly based on analysis of the larger data set). They used program MARK (White and Burnham 1999, White et al. 2000) to model the capture-recapture data and estimate model parameters.

The parameterization of all 5 models was structurally similar in that each model was based on an annual probability of survival (ϕ) and a recapture probability (p), conditional on a bird being alive at the beginning of year j. Grand et al. (1998) let the recapture probabilities be constant across years, denoted as $p.$, and let the survival probabilities vary by lead exposure (l) and site (s). The notation is standard in the capture-recapture literature (Lebreton et al. 1992). Thus, model $\{\phi_l, p.\}$ represents the hypothesis that annual survival probability varied by lead exposure, but not year or site, while recapture probability was constant over years, sites, and exposure. Model $\{\phi_{s+l}, p.\}$ represented the hypothesis that annual survival was constant across years, but varied by site and exposure but with no interactions. Model $\{\phi_{s*l}, p.\}$ represented the hypothesis that survival was constant across years, varied by site and exposure, but with an interaction term, $s \times l$. Model $\{\phi., p.\}$ assumed that both survival and recapture probabilities were constant across years, site, and exposure, while model $\{\phi_s, p.\}$ assumed that survival varied by site, but not year or exposure. Thus, empirical support for a hypothesized lead effect must stem from models $\{\phi_l, p.\}$, $\{\phi_{s+l}, p.\}$, and $\{\phi_{s*l}, p.\}$.

Model selection results presented by Grand et al. (1998) in their Table 3 are basically just the AIC differences, Δ_i; they base inference about the lead-effect (their Table 4) only on the selected best model. Here we extend their results to incorporate multi-model inference (MMI; Burnham and Anderson 1998). First, we have the Akaike weights, w_i, shown in our Table 3. The best model has ϕ varying by lead exposure, but only has $w_1 = 0.673$ as a strength of evidence for this best model. This weight suggests that model $\{\phi_l, p.\}$ is not convincingly the best model if other replicate data sets were available. The next two models add little support for a site-effect, either with or without interaction terms. This can be seen by considering AIC,

$$\text{AIC} = -2\log_e(\mathscr{L}(\phi, p)) + 2K.$$

AIC is an estimator of Kullback-Leibler information loss and embodies the principle of parsimony as a byproduct, not in its derivation. The first term in AIC is a lack of fit component and gets smaller as more parameters are fitted in the model, however, the second component gets

larger as a penalty for adding additional parameters. Thus, it can be seen that AIC enforces a trade-off between bias and variance as the number of parameters is increased. From Table 3, one can see that Δ_i for model $\{\phi_{s+l}, p.\}$ with $K = 4$ parameters increased by 2 units over the best model, while model $\{\phi_{s*l}, p.\}$ with $K = 5$ parameters increased by 4 units over the best model. The fit of the first 3 models in Table 3 is nearly identical; the additional hypothesized effect of site, with or without an interaction term, is not supported by the data. In each case, the Δ_i values increase by about 2 as the number of parameters increases by one. In total, the evidence for a lead-effect is very strong in that the sum of the Akaike weights for these 3 models is 0.998. Empirical support for the 2 models without a lead-effect is lacking (Table 3) as both models have $w_i = 0.001$. The evidence strongly suggests the presence of an effect on annual survival caused by ingestion of lead.

The evidence that model $\{\phi_l, p.\}$ is the best over replicated data sets can be easily judged by the ratio of the Akaike weights of the best model and the second ranked model. This evidence (e.g., $w_1/w_2 = 0.673/0.239 = 2.8$) is insufficient to justify ignoring issues of model selection uncertainty. Hence, from the Akaike weights, it is clear that a lead-effect on survival is required for a model (hypothesis) to be plausible here.

Finally, rather than ignore model selection uncertainty, we can use the model-averaged estimate of lead-effect on annual survival and its unconditional standard error (from Table 3). As is often the case, the model averaged estimate of effect size is very similar to the estimate from just the best model (0.335 vs. 0.337). However, the unconditional standard error, 0.125, is about 20% larger than the conditional standard error, 0.105, from the best model. This increase reflects model selection uncertainty and is an honest measure of uncertainty in the estimated effect of lead on eider survival probabilities.

Summary of the Information-Theoretic Approach

The principle of parsimony, or Occam's razor, provides a philosophical basis for model selection; Kullback-Leibler information provides an objective target based on deep, fundamental theory; information criteria (AIC and AIC_c), along with likelihood-based inference, provide a practical, general methodology for use in data analysis. Objective data analysis can be rigorously based on these principles without having to assume that the true model is contained in the set of candidate models. There are surely no true models in the biological sciences. Papers using information-theoretic approaches are beginning to appear in theoretical and applied journals in the biological sciences.

At a conceptual level, reasonable data and a good model allow a separation of information and noise. Here, information relates to the structure of relationships, estimates of model parameters, and components of variance. Noise then refers to the residuals: variation left unexplained. We can use the information extracted from the data to make proper inferences. The goal here is an approximating model that minimizes information loss, $I(f, g)$, and properly separates noise (non-information or entropy) from structural information. In an important sense, we are not trying to model the data, but rather we want to model the information in the data.

Information-theoretic methods are relatively simple to understand and practical to employ across a large class of empirical situations and scientific disciplines. The methods can be computed by hand if necessary (assuming one has the parameter estimates, maximized log-likelihood values, and $\widehat{\text{var}}(\hat{\theta}_i|g_i)$ for each of the R a priori models). The information-theoretic methods are easy to understand and we believe it is important that people understand the methods they employ. Further material on information-theoretic methods can be found in recent books by Burnham and Anderson (1998) and McQuarrie and Tsai (1998). Akaike's collected works have been recently published by Parzen et al. (1998).

CONCLUSIONS

The overwhelming occurrence of false null hypotheses in our sample of articles from *Ecology* and JWM seems sobering. Why are such strawmen being continually tested and the results accepted as science? We believe researchers in the applied sciences have been indoctrinated into thinking that statistical null hypothesis testing is a fundamental component of the scientific method. Researchers commonly treat scientific hypotheses and statistical null hypotheses as one in the same, which they are not (Romesburg 1981, Ellison 1996). As a result, ecologists live or die by the arbitrarily assigned

significant *P*-value (see Nester 1996 for a colorful description of the different types of emotional response to significance testing results).

In the worst, but common, case, only a *P*-value is presented, without even the sign of the supposed difference. Null hypothesis testing does not represent a fundamental aspect of the scientific method, but rather a pseudoscientific approach that provides a false sense of objectivity and rigor to analysis and interpretation of research data. Carver (1978:394) offers the extreme statement, "... statistical significance testing usually involves a corrupt form of the scientific method and, at best, is of trivial importance ...." Much of the statistical software currently available aggravates this situation by computing and displaying quantities related to various tests.

Results from null hypothesis testing lead to relatively little increase in understanding and divert attention from the important issues—estimation of effect size, its sign and its precision, and meaningful mechanistic modeling of predictive and causal relationships. We urge researchers to avoid using the words significant and nonsignificant as if these terms meant something of biological importance. Do not rely on statistical hypothesis tests in the analysis of data from observational studies, do not report only *P*-values, and avoid reliance on arbitrary α-levels to judge significance. Editors and referees should be wary of trivial null hypotheses being tested, the related *P*-values, and the implication of supposed significance.

There are alternatives to the traditional null hypothesis testing approach in data analysis. For example, the standard likelihood ratio provides a more realistic basis for strength of evidence (Edwards 1972, 1992; Royall 1997). There is a great deal of current research on Bayesian methods and practical approaches are forthcoming for use in the sciences. However, the Bayesian approaches seem computationally difficult and there may continue to be objections of a fundamental nature (Foster 1995, Dennis 1996, Royall 1997) to the use of Bayesian methods in strength-of-evidence assessments and conclusion-oriented, empirical science.

Information-theoretic methods offer a more useful, general approach in the analysis of empirical data than the mere testing of null hypotheses. The information-theoretic paradigm avoids statistical hypothesis testing concepts and focuses on relationships of variables (via model selection) and on the estimation of effect size and measures of its precision. These relatively new approaches are conceptually simpler and easily computable, once the model statistics are available. This paradigm is useful in providing evidence and making inferences from either a single (best) model or from many models (e.g., using MMI based on Akaike weights). Information-theoretic approaches cannot be used unthinkingly; a good set of a priori models is essential and this involves professional judgment and integration of the science of the issue into the model set.

Increased attention is needed to separate those inferences that rest on a priori considerations from those resulting from some degree of data dredging. Essentially no justifiable theory exists to estimate precision (or test hypotheses, for those still so inclined) when data dredging has taken place (the theory (mis)used is for a priori analyses, assuming the model was the only one fit to the data). This glaring fact is either not understood by practitioners and journal editors or is simply ignored. Two types of data dredging include (1) an iterative approach where patterns and differences observed after initial analysis are chased by repeatedly building new models with these effects included, and (2) analysis of all possible models (unless, perhaps, if model averaging is used). Data dredging is a poor approach to making reliable inferences about the sampled population. Both types of data dredging are best reserved for more exploratory investigations that probably should often remain unpublished. The incorporation of a priori considerations is of paramount importance and, as such, editors, referees, and authors should pay much closer attention to these issues and be wary of inferences obtained from post hoc data dredging.

LITERATURE CITED

AKAIKE, H. 1973. Information theory as an extension of the maximum likelihood principle. Pages 267–281 *in* B. N. Petrov and F. Csaki, editors. Second International Symposium on Information Theory. Akademiai Kiado, Budapest.

———. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control *AC* 19:716–723.

BERGER, J. O., AND D. A. BERRY. 1988. Statistical analysis and the illusion of objectivity. American Scientist 76:159–165.

———, 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. 1938. Some difficulties of interpretation encountered in the application of the chi-squared test. Journal of the American Statistical Association 33:526–536.

———. 1942. Tests of significance considered as evidence. Journal of the American Statistical Association 37:325–335.

BUCKLAND, S. T., K. P. BURNHAM, AND N. H. AUGUSTIN. 1997. Model selection: an integral part of inference. Biometrics 53:603–618.

BURNHAM, K. P., AND D. R. ANDERSON. 1998. Model selection and inference: a practical information-theoretic approach. Springer-Verlag, New York, New York, USA.

———, ———, G. C. WHITE, C. BROWNIE, AND K. H. POLLOCK. 1987. Design and analysis methods for fish survival experiments based on release recapture. American Fisheries Society Monograph 5.

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

CHAMBERLIN, T. 1965 (1890). The method of multiple working hypotheses. Science 148:754–759. (reprint of 1890 paper in Science).

CHERRY, S. 1998. Statistical tests in publications of The Wildlife Society. Wildlife Society Bulletin 26: 947–953.

COHEN, J. 1994. The earth is round ($p<.05$). American Psychologist 49:997–1003.

COX, D. R. 1958. Some problems connected with statistical inference. Annals of Mathematical Statistics 29:357–372.

DELEEUW, J. 1992. Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. Pages 599–609 *in* S. Kotz and N. L. Johnson, editors. Breakthroughs in statistics. Volume 1. Springer-Verlag, London, United Kingdom.

DENNIS, B. 1996. Discussion: should ecologists become Bayesians? Ecological Applications 6:1095–1103.

EDWARDS, A. W. F. 1972. Likelihood: an account of the statistical concept of likelihood and its application to scientific inference. Cambridge University Press, London, United Kingdom.

———. 1992. Likelihood: expanded edition. The Johns Hopkins University Press, Baltimore, Maryland, USA.

EFRON, B., AND R. J. TIBSHIRANI. 1993. An introduction to the bootstrap. Chapman & Hall, New York, New York, USA.

ELLISON, A. M. 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. Ecological Applications 6:1036–1046.

FISHER, R. A. 1928. Statistical methods for research workers. Second edition. Oliver and Boyd, London, United Kingdom.

FOSTER, M. R. 1995. Bayes and bust: the problem of simplicity for a probabilist's approach to confirmation. British Journal for the Philosophy of Science 46:399–424.

GARDNER, M. J., AND D. G. ALTMAN. 1986. Confidence intervals rather that P values: estimation rather than hypothesis testing. British Medical Journal 292:746–750.

GELMAN, A., J. C. CARLIN, H. STERN, AND D. B. RUBIN. 1995. Bayesian data analysis. Chapman & Hall, New York.

GERARD, P. D., D. R. SMITH, AND G. WEERAKKODY. 1998. Limits of retrospective power analysis. Journal of Wildlife Management 62:801–807.

GIGERENZER, G., Z. SWIJTINK, T. PORTER, L. DASTON, J. BEATTY, AND L. KRUGER. 1989. The empire of chance: how probability changed science and everyday life. Cambridge University Press, Cambridge, United Kingdom.

GOODMAN, S. N. 1993. P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. American Journal of Epidemiology 137:485–496.

———, AND J. A. BERLIN. 1994. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Annals of Internal Medicine 121:200–206.

———, AND R. ROYALL. 1988. Evidence and scientific research. American Journal of Public Health 78:1568–1574.

GRAND, J. B., P. L. FLINT, M. R. PETERSEN, AND C. L. MORAN. 1998. Effect of lead poisoning on spectacled eider survival rates. Journal of Wildlife Management 62:1103–1109.

GRAYBILL, F. A., AND H. K. IYER. 1994. Regression analysis: concepts and applications. Duxbury Press, Belmont, California, USA.

HARLOW, L. L., S. A. MULAIL, AND J. H. STEIGER, EDITORS. 1997. What if there were no significance tests. Lawrence Erlbaum Associates, Mahwah, New Jersey, USA.

HEDGES, L. V., AND I. OLKIN. 1985. Statistical methods for meta-analysis. Academic Press, London.

HENDERSON, A. R. 1993. Chemistry with confidence: should *Clinical Chemistry* require confidence intervals for analytical and other data? Clinical Chemistry 39:929–935.

IYENGAR, S., AND J. B. GREENHOUSE. 1988. Selection models and the file drawer problem. Statistical Science 3:109–135.

JOHNSON, D. H. 1995. Statistical sirens: the allure of nonparametrics. Ecology 76:1998–2000.

———. 1999. The insignificance of statistical significance testing. Journal of Wildlife Management 63:763–772.

KULLBACK, S., AND R. A. LEIBLER. 1951. On information and sufficiency. Annals of Mathematical Statistics 22:79–86.

LEBRETON, J. D., K. P. BURNHAM, J. CLOBERT, AND D. R. ANDERSON. 1992. Modeling survival and testing biological hypotheses using marked animals: case studies and recent advances. Ecological Monographs 62:67–118.

MCQUARRIE, A. D. R., AND C-L. TSAI. 1998. Regression and time series model selection. World Scientific Press, Singapore, Malaysia.

MORRISON, D. E., AND R. E. HENKEL. 1969. Statistical tests reconsidered. The American Sociologist 4:131–140.

———, AND ———, EDITORS. 1970. The significance test controversy—a reader. Aldine Publishing, Chicago, Illinois, USA.

NESTER, M. R. 1996. An applied statistician's creed. Applied Statistics 45:401–410.

NEYMAN, J., AND E. S. PEARSON. 1928. On the use and interpretation of certain test criteria for purposes of statistical inference, I & II. Biometrika 20A:175–200, 263–294.

———, AND ———. 1933. On the problem of the most efficient test of statistical hypotheses. Philosophical Transactions of the Royal Statistical Society A 231:289–337.

PARZEN, E., K. TANABE, AND G. KITAGAWA, EDITORS. 1998. Selected papers of Hirotugu Akaike. Springer-Verlag Inc., New York, New York, USA.

ROMESBURG, H. C. 1981. Wildlife science: gaining reliable knowledge. Journal of Wildlife Management 45:293–313.

ROYALL, R. M. 1997. Statistical evidence: a likelihood paradigm. Chapman & Hall, London, United Kingdom.

SAVAGE, I. R. 1957. Nonparametric statistics. Journal of the American Statistical Association 52:331–344.

SCHMIDT, F. L. 1996. Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. Psychological Methods 1:115–129.

WHITE, G. C., AND K. P. BURNHAM. 1999. Program MARK—survival estimation from populations of marked animals. Bird Study 46 (supplement): S120–S139.

———, ———, AND D. R. ANDERSON. 2000. Advanced features of program MARK. Pages xxx–xxx *in* Integrating People and Wildlife for a Sustainable Future, R. Fields, editor. Proceedings of the Second International Wildlife Management Congress. The Wildlife Society, Bethesda, Maryland.

YATES, F. 1951. The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. Journal of the American Statistical Association 46:19–34.

YOCCOZ, N. G. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. Bulletin of the Ecological Society of America 72:106–111.

Received 5 May 1999.
Accepted 20 June 2000.
Associate Editor: Bunck.