Journal of Applied Ecology 2007
44, 456–460

FORUM

Concerns regarding a call for pluralism of information theory and hypothesis testing

PAUL M. LUKACS,* WILLIAM L. THOMPSON,† WILLIAM L. KENDALL,‡ WILLIAM R. GOULD,§ PAUL F. DOHERTY JR,¶ KENNETH P. BURNHAM** and DAVID R. ANDERSON††

**USGS Patuxent Wildlife Research Center and IAP World Services, 1201 Oakridge Dr., Fort Collins, CO 80525, USA; †National Park Service, Southwest Alaska Network, 240 West 5th Ave, Anchorage, AK 99501, USA; ‡USGS Patuxent Wildlife Research Center, 12100 Beech Forest Road, Laurel, MD 20708, USA; §University Statistics Center, New Mexico State University, Las Cruces, NM 88003, USA; ¶Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, CO 80523, USA; **Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, 1484 Campus Delivery, Fort Collins, CO 80523; ††Applied Information Company Inc., 707 Breakwater Dr., Fort Collins, CO 80525, USA*

Summary

1. Stephens *et al*. (2005) argue for 'pluralism' in statistical analysis, combining null hypothesis testing and information-theoretic (I-T) methods. We show that I-T methods are more informative even in single variable problems and we provide an ecological example.

2. I-T methods allow inferences to be made from multiple models simultaneously. We believe multimodel inference is the future of data analysis, which cannot be achieved with null hypothesis-testing approaches.

3. We argue for a stronger emphasis on critical thinking in science in general and less reliance on exploratory data analysis and data dredging. Deriving alternative hypotheses is central to science; deriving a single interesting science hypothesis and then comparing it to a default null hypothesis (e.g. 'no difference') is not an efficient strategy for gaining knowledge. We think this single-hypothesis strategy has been relied upon too often in the past.

4. We clarify misconceptions presented by Stephens *et al*. (2005).

5. We think inference should be made about models, directly linked to scientific hypotheses, and their parameters conditioned on data, Prob(H_j | data). I-T methods provide a basis for this inference. Null hypothesis testing merely provides a probability statement about the data conditioned on a null model, Prob(data | H_0).

6. *Synthesis and applications*. I-T methods provide a more informative approach to inference. I-T methods provide a direct measure of evidence for or against hypotheses and a means to consider simultaneously multiple hypotheses as a basis for rigorous inference. Progress in our science can be accelerated if modern methods can be used intelligently; this includes various I-T and Bayesian methods.

Key-words: Akaike's information criterion, information theory, model selection, multimodel inference, null hypothesis testing, statistical analysis

Journal of Applied Ecology (2007) **44**, 456–460
doi: 10.1111/j.1365-2664.2006.01267.x

Correspondence: Paul M. Lukacs, Colorado Division of Wildlife, 317 W. Prospect Road, Fort Collins, CO 80526, USA (e-mail Paul.Lukacs@state.co.us).

Introduction

A recent paper by Stephens *et al*. (2005) questions the rejection of null hypothesis testing (NHT) and calls for a 'pluralism' of analysis methods. Much of their paper

restates prior criticisms of the common misuses of NHT, with which we are in agreement. The problems with NHT and *P*-values have been discussed extensively in the literature (for a list of > 400 citations see http://www.warnercnr.colostate.edu/~anderson/null.html (accessed 18 December 2006)), so we do not repeat them here. We agree with Stephens *et al*. (2005) that both NHT and information-theoretic (I-T) methods can be used inappropriately and careful attention must be paid to the use of all statistical methods. However, there are several occasions in Stephens *et al*. (2005) in which the potential for misunderstanding is great; thus we feel further clarity regarding I-T approaches is warranted. For example, we interpreted their call for plurality as the notion that combining IT methods and NHT in data analysis can provide stronger inference. The authors describe a study that combined usage of likelihood ratio testing and I-T methods as a powerful analysis approach. Further correspondence has indicated their intent was to suggest that I-T methods should be placed alongside NHT in the biologist's statistical toolbox and not to suggest that both approaches be used in concert with one another.

NHT remains in use in scientific research. NHT is not mathematically wrong; it is just relatively uninformative for scientific questions compared with modern analysis methods. Scientists are changing their views of inference as better methods, such as I-T and Bayesian, are replacing NHT. I-T and Bayesian methods provide statistical frameworks for pursuing the multiple-hypothesis approach to science advanced by Chamberlin (1897) and endorsed as strong inference by Platt (1964). They are also consistent with modern approaches to decision making in the face of competing ecological models, found in adaptive resource management (Walters 1986; Williams, Nichols & Conroy 2002). We do not object to NHT being in a statistical toolbox if used carefully. We would rarely use NHT ourselves but often we, as journal referees, would not argue against its use, for example as a means of assessing goodness-of-fit.

NHT and I-T represent two philosophically different views of data analysis and inference. NHT attempts to present a binary choice between the null hypothesis (H_0) and the alternative (H_A), based on an arbitrary α level and the resulting *P*-value, the probability of the data (X) and unobserved, more extreme, data given the null hypothesis, Prob($X \mid H_0$). In contrast, I-T provides simple ways to quantify directly the evidence for two or more science hypotheses. This evidence usually stems from a simultaneous analysis of multiple hypotheses and includes a ranking of the models based on information loss, the probability of each model given the data [Prob($H_j \mid X$) for $j = 1, 2, \ldots, R$, where R is the number of models] and evidence ratios (Table 1).

Stephens *et al*.'s (2005) abbreviation 'information-theoretic model comparison (ITMC)' poorly denotes the breadth of this approach. Kullback–Leibler information and its asymptotic estimator (Akaike's information criterion) is much more than a 'model comparator'.

Table 1. Summary of the types of evidence provided under null hypothesis testing and information-theoretic approaches

Null hypothesis testing	Information-theoretic
P-values = Prob($X \mid H_0$)	Ranking of H_j $j = 1, \ldots R$
	Probability of hypothesis j = Prob($H_j \mid X$)
	Evidence ratios, hypothesis i vs. j
	Model averaging
	Unconditional estimates of precision

Hence we have adopted the abbreviation I-T in discussing this class of methods, including several procedures to allow rigorous inference from more than a single model (multimodel inference). Stephens *et al*. (2005) further state that NHT is a more appropriate tool for some questions that they later clarify as being single-parameter studies. We do not agree that NHT is more appropriate than I-T methods for the scenarios they presented, and offer an explanation for our disagreement. We wish to make two main points. First, we demonstrate that I-T is directly applicable in single-parameter problems and, moreover, that it is more informative than NHT. Secondly, we stress that developing scientific hypotheses is a difficult process and it should rightly be challenging. Hypothesizing is the cornerstone of science and must be given considerable thought. I-T methods encourage greater a priori thinking than NHT, which focuses on the testing of a null hypothesis against an alternative.

The single-parameter case

Stephens *et al*. (2005) suggested that NHT serves well in single-parameter situations. We argue that I-T methods remain more informative even in a case where only a single additional parameter is in question. Our disagreement with Stephens *et al*.'s (2005) position can be illustrated using the capture–recapture data collected for the European dipper *Cinclus cinclus* (Lebreton *et al*. 1992) to contrast the two approaches. The points we make here are general and are not restricted to this particular example. The frog–atrazine case in Stephens *et al*. (2005) could easily be handled with the more useful I-T methodology.

The basic research question in the European dipper study related to whether the apparent survival probability of birds differed in years when floods occurred during the breeding season (this species nests near streams) vs. a normal year when a flood did not occur. In each case there are two models: (1) $\{\varphi(\cdot),p(\cdot)\}$,implying that apparent survival (φ) and recapture probabilities (p) are approximately constant over years; and (2) $\{\varphi(n),\varphi(f),p(\cdot)\}$, where years are partitioned into normal years (n) and flood years (f) in terms of apparent survival probabilities. These models are clear representations of two science hypotheses, one where a flood has no impact on apparent survival, and one where a

Table 2. Maximum likelihood estimates (MLE) and confidence intervals (CI) for apparent survival (φ) of European dippers in flood (*f*) and normal (*n*) years from two models

Model	MLE	95% CI	
{φ(·),*p*(·)}	0·5602	0·5105	0·6088
{φ(*n*),φ(*f*),*p*(·)}			
(*n*)	0·6071	0·5451	0·6658
(*f*)	0·4688	0·3858	0·5537

flood does impact apparent survival. The first model has $K = 2$ parameters, whereas the second model has $K = 3$ parameters. The research question (a simple one-parameter observational study) relates to the possible change in survival probability during flood years.

The maximum likelihood estimates (MLE) for φ and measures of precision for parameters in the two models are presented in Table 2. The difference in estimates of survival probability (the 'effect size') is 0·1383, SE = 0·0532, with a 95% confidence interval for this difference of (0·0340, 0·2425). The MLE of p is 0·9025 (SE = 0·0286) vs. 0·8997 (SE = 0·0293) for the two models, respectively.

NULL HYPOTHESIS TESTING

The simpler model is nested in the three-parameter model and a simple likelihood ratio test of the two models provides a test statistic of 6·735 and, assuming this is χ^2 distributed on one degree of freedom, we obtain a *P*-value of 0·0095. This would be ruled 'significant'; some would say 'highly significant' and others would include '**' in tabular material to emphasize its high significance. Note that only the null hypothesis (H_0) is the subject of the test.

Formally, the *P*-value of 0·0095 is the probability of a value as large as 6·735 or larger, given the null model {φ(·),*p*(·)} is true. Given that this is such a small probability, one concludes (by default) that the alternative model {φ(*n*),φ(*f*),*p*(·)} is 'significantly' better. The proper interpretation of the *P*-value is strained; this provides some explanation regarding why so many people erroneously believe the *P*-value means something else (e.g. the probability that the null model is true).

I-T APPROACH

Under this approach, one obtains the model probabilities directly:

Model	Probability
{φ(.),*p*(.)}	0·0868
{φ(*n*),φ(*f*),*p*(.)}	0·9132.

In addition, these are mathematically equivalent to Bayesian posterior model probabilities (Burnham & Anderson 2004). These model probabilities provide direct evidence regarding the empirical support for the two models, without having to assume that either model is 'true' (there are no true models). We believe that most scientists and resource managers would view these model probabilities as more meaningful forms of evidence compared with *P*-values.

The quantification of information loss ($\Delta_i = \mathrm{AIC}_i - \mathrm{minAIC}$) allows the computation of the likelihood of model g_i, given the data:

$$L(g_i \mid X) = e^{-\frac{1}{2}\Delta_i}, \; i = 1, \ldots, R.$$

The probability of model i is a normalization of the model likelihoods:

$$w_i = \frac{L(g_i \mid X)}{\sum_{j=1}^{R} L(g_j \mid X)}.$$

The w_i are 'Akaike weights' or model probabilities. These weights are quite unlike *P*-values (probability of the data, given the null model), instead they are the probability of model i, given the data (Table 1). Finally, an evidence ratio (E) is useful in comparing the relative strength of evidence for two hypotheses, i and j:

$$E = \frac{L(g_i \mid X)}{L(g_j \mid X)} = \frac{w_i}{w_j}.$$

Burnham & Anderson (2002) provide a discussion of evidence ratios and model probabilities. Evidence ratios provide a measure of the relative likelihood of one hypothesis vs. another. Here, likelihood has a technical meaning, can be quantified and should not be confused with probability. For example, if person A holds three raffle tickets and person B has one, person A is three times more *likely* to win than person A. We do not know the absolute probability of either person winning without knowing the total number of raffle tickets. In the dipper example, the evidence ratio gauges the relative support for the two alternatives: 0·9132/0·0868 = 10·52. Given the available data, a difference in survival probability having occurred between normal and flood years is 10·52 times more likely than no difference having occurred. This suggests somewhat limited to moderate evidence for a flood effect on apparent survival probability (strong evidence of a flood effect is not warranted, contrary to the result from NHT). Evidence ratios are invariant to other models in the model set and are the statistic used in legal settings, such as criminal trials relying on DNA evidence (Evett & Weir 1998). Evidence ratios are a continuous measure, but some useful guidelines have existed in the statistical literature (Table 3; Jeffreys 1948; Evett & Weir 1998). Inference should be about models and parameters, given data; however, *P*-values are probability statements about data, given null models. Model probabilities and evidence ratios provide a means to make inference directly about models and their parameters, given data.

I-T methods can be used in single-parameter problems such as the pollutant problem posed by Stephens *et al.* (2005), despite their claim that AIC was not applicable

Table 3. General guidelines for the amount of support given by an evidence ratio based on Evett & Weir (1998)

Evidence ratio	Verbal description
1–10	Limited support
10–100	Moderate support
100–1000	Strong support
> 1000	Very strong support

because 'AIC cannot be used to compare models of different data sets' (Stephens *et al*. 2005). Stephens *et al*. (2005) misinterpret what is meant in the statistical sciences as a 'data set.' In particular, a data set does not mean just one vector of numerical values. The example presented by Stephens *et al*. (2005) is a case of a control–treatment design, which assumes a control (sites are similar) and independent samples at each site. In actuality, paired control–treatment samples would be recorded at similar times because pollutant effects are time-dependent as a result of stream flow, but we disregard this to be consistent with Stephens *et al*.'s (2005) original example. Thus, the analysis could be framed as two models:

$$Y = \beta_0$$
$$Y = \beta_0 + (\beta_1 + 5)X$$

where Y is the concentration of the pollutant, β_0 is the overall mean concentration (the intercept), β_1 is the treatment effect, 5 is the constant added representing the minimum treatment effect of interest, and X is an indicator of the upstream (control) or downstream (treatment) site. The intercept-only model treats the control and treatment observations as if they were collected at the same site, whereas the indicator-variable model constrains these observations to be site-specific in the analysis. In this case the response variable is the same, therefore models can be built to represent the hypotheses and I-T methods are applicable. If the response variables were different at each site, neither I-T nor NHT could be used.

Stephens *et al*. (2005) were mistaken when they claimed that NHT can provide 'the probability with which H_A could be supported'. NHT does not provide information about the probability of the alternative hypothesis because only the null hypothesis is the subject of the test. I-T methods provide the probability of the alternative hypothesis that the authors seek and both the model weights and the evidence ratios quantify the empirical support for the hypotheses, whether there are two or more such hypotheses.

I-T methods allow us to go a step further in our analysis and make formal inference from multiple models simultaneously (Burnham & Anderson 2002). The $\{\varphi(\cdot),p(\cdot)\}$ and $\{\varphi(n),\varphi(f),p(\cdot)\}$ apparent survival estimates can be model-averaged to produce an estimate of flood and normal year apparent survival that takes model selection uncertainty into account. The model-averaged estimates are a weighted average of the estimates of the two models, with the weights based on the model probabilities (Burnham & Anderson 2002). The model-averaged variance accounts for sampling variance and variation in parameter estimates across models (Burnham & Anderson 2002). The model averaged effect size for the dipper example is 0·1263 (SE = 0·0639). Note the estimate is a bit smaller than that for the $\{\varphi(n),\varphi(f),p(\cdot)\}$ model and the standard error is larger. The difference represents the uncertainty as to which model is actually best in terms of Kullback–Leibler information loss. This estimate is conditional on the set of models considered rather than a single model. NHT offers no procedure for model averaging or for computing the unconditional estimates of sampling variation or covariation.

It remains true that NHT is not wrong, but it is relatively uninformative in most cases. The scientific method in combination with NHT has increased knowledge since its formalization. However, the theory underlying NHT is weak in that it is based on the probability of the data, given the null model. We believe I-T approaches represent an improved methodology because these methods encourage greater a priori thinking about plausible scientific hypotheses (even if there are only two) and because the outputs are more directly interpretable, regardless of the problem sophistication.

Developing scientific hypotheses

Statistical models should represent a translation of scientific hypotheses to their equivalent mathematical expression. An ecologist may need to collaborate with a statistician to turn a scientific hypothesis into such a mathematical statement. Such collaboration would be likely to be fruitful for both the ecologist and statistician, especially when initiated prior to collecting any data, and this has been our experience. There may be cases where a null hypothesis is plausible, and in those cases its model should be included in the set of models under consideration. The science hypotheses and statistical models should always be very tightly linked.

Stephens *et al*. (2005) suggest that developing scientific hypotheses and the statistical models that represent them is a difficult process, therefore exploratory data analysis (EDA) is needed to suggest new hypotheses. EDA is often promoted as a method allowing researchers to uncover relationships they would not have thought of previously. We agree that developing scientific hypotheses is challenging; however, the ultimate role of the scientist is that of developing and evaluating plausible hypotheses. EDA is a risky method for developing scientific hypotheses. The probability of finding an effect that is spurious is often quite high with EDA (Freedman 1983). Frequent chasing of spurious effects slows the progress of science (Anderson *et al*. 2001).

We encourage some post-hoc examination of data, but only after the a priori investigations have been completed. Then it is important to keep the two types of

inferences separate. EDA based on NHT methods such as stepwise selection simply removes thought from data analysis. NHT provides little to an EDA to either develop unknown relationships or extract new information. All of the relationships are tied to the estimation method, not NHT. Therefore, if one were to use I-T or NHT it would make little difference in the exploratory analysis. I-T does provide the opportunity to model-average and hedge one's bet against spurious effects. We stress that hard thinking about the scientific question combined with theory about a problem presents a far better way to develop new hypotheses.

We urge researchers to place substantial mental effort to derive a set of plausible scientific hypotheses. Hypothesizing is the centre of science. Soule (1987) stated: 'models are tools for thinkers, not crutches for the thoughtless'. Ecologists must actively engage in developing meaningful hypotheses, rather than always defaulting to the standard null hypothesis, 'no effect'.

Are analyses employing both frequentist and I-T approaches more revealing?

We question whether analysing the same data set with both NHT and I-T approaches has any justification. Stephens *et al*. (2005) suggest that 'if our objective is to maximize our understanding of a system and develop a model that best approximates reality … , there is an argument that we should use whatever means are available to do so'. We are not aware of any statistical theory that exists to guide an analyst in a combined analysis. What does one do if the results conflict? Stephens *et al*. (2005) suggest exploring assumptions, but that should be done as part of any analysis regardless of the method. Why would one feel better about an inference if the results agree? The same data are being used for the two analyses and the estimation procedures are in large part identical. The two analyses are not independent lines of evidence. Simply using 'whatever means are available' in the absence of guiding theory provides a weak basis for science.

Conclusions

I-T and NHT represent two very different philosophical views of data analysis and inference. We argue that I-T methods provide a more informative approach to inference. I-T methods provide a direct measure of evidence for or against hypotheses and a means to consider simultaneously multiple hypotheses as a basis for rigorous inference. It is the evidence ratio that allows a researcher to consider the relative support for competing scientific hypotheses, rather than merely selecting an alternative hypothesis by default because the probability of the data is small given the null hypothesis. I-T methods allow a scientist to make inference from models and parameters conditioned on data, rather than probability statements about the data conditional on a null model (Table 1). While it is true that both I-T and NHT methods can be misused, no one is advocating misuse of any statistical methods. When used properly, I-T methods provide more information to the researcher.

Historically, NHT has occupied a large place in the statistical toolbox. We ourselves have used NHT in the recent past in limited applications. Increasingly we see the place for NHT getting smaller. It is hard to understand why one would cling to an inefficient tool when better options exist. Our statistical toolbox has grown rapidly in recent decades. I-T and Bayesian methods should occupy the top shelf of the toolbox, with NHT getting only a small drawer.

Stephens *et al*. (2005) argue that many ecologists are confused about analysis methods. We see the confusion dissipating and the use of I-T methods rapidly increasing. We also see an increase in the use of Bayesian methods, particularly for models with random effects. Here we have attempted to help clarify I-T methodology and demonstrate its clear advantages over NHT.

References

Anderson, D.R., Burnham, K.P., Gould, W.R. & Cherry, S. (2001) Concerns about finding effects that are actually spurious. *Wildlife Society Bulletin*, **29**, 311–316.

Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Multimodel Inference*, 2nd edn. Springer-Verlag, New York, NY.

Burnham, K.P. & Anderson, D.R. (2004) Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods in Research*, **33**, 261–304.

Chamberlin, T.C. (1897) The method of multiple working hypotheses. *Journal of Geology*, **5**, 837–848.

Evett, I.W. & Weir, B.S. (1998) *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sinauer Associates Inc., Sunderland, MA.

Freedman, D.A. (1983) A note on screening regression equations. *The American Statistician*, **37**, 152–155.

Jeffreys, H. (1948) *Theory of Probability*. Oxford University Press, Oxford, UK.

Lebreton, J.-D., Burnham, K.P., Clobert, J. & Anderson, D.R. (1992) Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monographs*, **62**, 67–118.

Platt, J.R. (1964) Strong inference. *Science*, **146**, 347–353.

Soule, M.E. (1987) Where do we go from here? *Viable Populations for Conservation* (ed. M.E. Soule), pp. 175–183. Cambridge University Press, Cambridge, UK.

Stephens, P.A., Buskirk, S.W., Hayward, G.D. & Martinez del Rio, C. (2005) Information theory and hypothesis testing: a call for pluralism. *Journal of Applied Ecology*, **42**, 4–12.

Walters, C.J. (1986) *Adaptive Management of Renewable Resources*. MacMillan, New York, NY.

Williams, B.K., Nichols, J.D. & Conroy, M.J. (2002) *Analysis and Management of Animal Populations*. Academic Press, San Diego, CA.

Received 30 January 2006; final copy received 13 November 2006
Editor: Rob Freckleton