Opinion TRENDS in Ecology and Evolution Vol.22 No.4

Inference in ecology and evolution

Philip A. Stephens[1], Steven W. Buskirk[2] and Carlos Martínez del Rio[2]

[1] Department of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, UK
[2] Department of Zoology and Physiology, University of Wyoming, Laramie, WY 82071-3166, USA

Most ecologists and evolutionary biologists continue to rely heavily on null hypothesis significance testing, rather than on recently advocated alternatives, for inference. Here, we briefly review null hypothesis significance testing and its major alternatives. We identify major objectives of statistical analysis and suggest which analytical approaches are appropriate for each. Any well designed study can improve our understanding of biological systems, regardless of the inferential approach used. Nevertheless, an awareness of available techniques and their pitfalls could guide better approaches to data collection and broaden the range of questions that can be addressed. Although we should reduce our reliance on significance testing, it retains an important role in statistical education and is likely to remain fundamental to the falsification of scientific hypotheses.

Introduction

Ecologists and evolutionary biologists study complex systems that are characterized by high natural variability and, not surprisingly, rely heavily on statistics to infer pattern and causation from their data. In that context, the use of null hypothesis significance tests (NHST) predominates (Box 1). Typically, NHST are designed to determine the probability with which an observed effect (e.g. a difference between means or a non-zero regression slope) would be observed if the true effect is zero. Over the past decade, several biologists have emphasized the limitations and problems of NHST, promoting a variety of alternatives (e.g. Refs. [1–4]). In spite of this, NHST remains the main approach to inference in ecology and evolution.

Here, we explore the debate over the value of NHST. We do not provide a user's guide to statistics, as other recent works provide more detailed information [3–7]. Rather, we intend to increase awareness of the debate and of its importance in ecology and evolutionary biology, and to encourage the use of other approaches to inference. To that end, we give a brief review of problems with NHST and highlight alternative approaches. We conclude by identifying some common applications of statistics and by suggesting approaches that are most appropriate in each case. We advocate a pluralistic approach to statistics: no single statistical approach currently available is universally preferable. We believe that NHST remains valuable in the context of hypothesis falsification, but we stress that not all statistical analyses are conducted for that purpose.

Corresponding author: Stephens, P.A. (philip.stephens@bristol.ac.uk).
Available online 13 December 2006.

What's wrong with NHST?

Criticisms of NHST have a long history, reviewed in detail elsewhere (see over 400 citations at http://www.warnercnr.colostate.edu/~anderson/thompson1.html). Briefly, problems with NHST fall into two broad categories: (i) those associated with interpretation; and (ii) those involving deeper philosophical issues. Problems of interpretation are usually attributable to poor understanding or sloppy application. These include an inappropriate focus on statistical significance rather than on biological significance [8], poor understanding of the importance of statistical power (e.g. Ref. [9]), a tendency to encourage arbitrary inferences [10], incomplete reporting [1,10] and publication bias favouring statistical significance [11].

Three particularly common examples of poor interpretation include equating: (i) a failure to reject the null hypothesis with the assertion that the null must be true [10,12]; (ii) the probability with which the data could have been obtained, given the null hypothesis, with the probability that the null hypothesis is true [13–15]; and (iii) poor support for the null hypothesis, with strong support for the alternative hypothesis [13].

As an example of the logical flaws underlying the misinterpretation of NHST, consider the second of these problems. The P value produced by a NHST is the probability of observing the vector of data, $\mathbf{Y}$, or one yielding an even more extreme test statistic, t, assuming that the null hypothesis (H_0) is true [i.e. $P(t \geq t_{\text{obs}}|H_0)$]. Unfortunately, when a suitably low P value is obtained from a test, researchers often interpret it to indicate that H_0 is highly unlikely to be true. Cohen [14] explains why this is not the case. Much deductive reasoning rests on syllogisms, in which two unarguable premises (P1 and P2) lead to a deduced conclusion (C). Problems arise, however, when syllogisms are made probabilistic. For example, a probabilistic syllogism might be: (P1) if I have a lottery ticket, I am unlikely to win the lottery; (P2) I won the lottery; (C) I am unlikely to have had a lottery ticket. Although P1 is unarguable (given the low probability of winning the lottery), C does not follow from P1 and P2. However, in science, such a conclusion is routinely made when we use syllogisms of the form: (P1) if the null hypothesis is true, data $\mathbf{Y}$ are unlikely to be observed; (P2) data $\mathbf{Y}$ were observed; (C) the null hypothesis is unlikely to be true.

The more profound philosophical limitations of NHST arise because, by focusing scientific effort on a null and (typically) a single alternative, NHST can limit advances [3]. The emphasis on falsification by NHST leads to a binary approach to rejection or acceptance that can obscure uncertainty about the best explanation for an observed

  doi:10.1016/j.tree.2006.12.003

Box 1. Current use of statistical approaches in ecology and evolution

We reviewed the last 50 empirical papers published in 2005 in each of four journals in ecology and evolutionary biology (*Behavioral Ecology, Ecology Letters*, *Evolution* and the *Journal of Applied Ecology*). Reviews and purely modelling-based papers were excluded, as we were principally interested in how inferences were drawn from data.

Papers were scored according to whether they used NHST (black bars), information theoretic approaches (IT; red bars), or other approaches (principally Bayesian analyses or likelihood-based phylogenetic tree constructions; green bars). Some papers used more than one type of approach and, thus, totals sum to more than 100%. In all four journals, most papers ($\geq$90% in each case) used NHST, whereas $\leq$10% used IT (Figure I). Other techniques were commonly used only in *Evolution*, which is perhaps unsurprising, given the relative frequency with which authors in that journal deal with phylogenetic inferences.

It has been suggested that NHST approaches can be appropriate in experimental studies but should not be used in observational studies (because variance in the data set has not been generated by experimental manipulation, leaving inference vulnerable to unconsidered confounding factors) [6]. Consequently, we also scored papers according to whether they used observational or experimental data (*Behavioral Ecology*, $n = 21$; *Ecology Letters*, $n = 26$; Evolution, $n = 31$; *Journal of Applied Ecology*, $n = 38$), and assessed the frequency with which NHST approaches were used (blue bars). These frequencies tended to be only marginally lower than the frequencies with which NHST was used overall (Figure I).

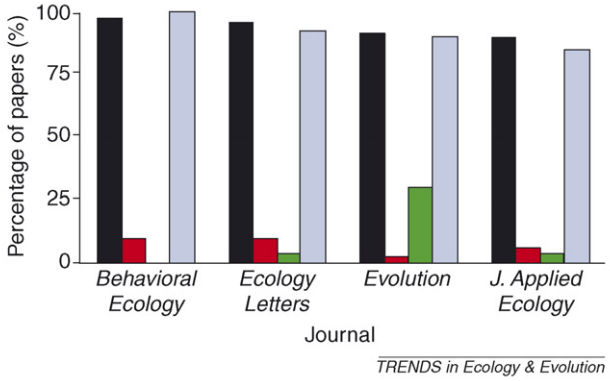

Figure I.

phenomenon [16,17]. NHST provides no option for weighting different hypotheses according to the belief that we have in them, or for multi-model inference where no single model is clearly superior [6].

Why is NHST still so widely used?

The adoption of new methods takes time and many biologists remain confused about appropriate statistical techniques, or unaware of the limitations of the techniques that they use. Even in disciplines such as psychology, in which awareness of the problems with NHST is more widespread, improvements have been slow [18]. Better teaching of statistics is essential to accelerate changes, which requires that those in a position to teach keep abreast of developing techniques. It also suggests that statistics should be taught as rigorously as any other discipline in ecology and evolution, in which background reading, essays and examinations are the norm. Although some statistics courses are already taught this way, we suggest that these are the exception rather than the rule. We suspect that most statistics courses taught to biologists emphasize NHST or concern NHST exclusively. Clearly, the roots and theory of NHST will always be a fundamental component of any introductory course on statistics. Nevertheless, the theory of NHST should not be taught without equal consideration of problems with, and alternatives to, those techniques.

The apparent ease with which classical analyses can be conducted using off-the-shelf statistical software, for which few equivalents exist for more complex analytical approaches, also hinders the uptake of new approaches. However, alternative methods can often be conducted simply, with no more equipment than a calculator or computer spreadsheet (e.g. Ref. [6]). They can also simplify analyses by ensuring that the analyses fit the question, rather than the opposite process. Hobbs and Hilborn note that some model selection algorithms can reduce the necessity to make assumptions at the outset (such as complex assessments of probable error distributions), as it is possible for the inferential approach to distinguish between these [4].

Finally, many experimentalists view the problems of NHST as being relevant primarily to observational science. However, the philosophical limitations of NHST remain, regardless of whether the data are collected experimentally or observationally. Furthermore, experimental and observational work fall on a continuum, ranging from strictly controlled laboratory designs, through partially controlled field manipulations, to purely observational data collection. Even carefully designed experiments can be confounded by the potential for alternative explanation, especially where the subject matter is complex (e.g. Refs. [19,20]). No scientist wishing to retain flexibility within the experimental-to-observational continuum can afford to ignore the wider debate on data analysis and interpretation.

What are the alternatives?

Alternatives to standard NHST include effect size statistics, model selection approaches based on information criteria (sometimes referred to as information theoretic or IT techniques), and Bayesian statistics (Table 1).

Effect size statistics

Effect size statistics overcome many of the problems of interpretation that are inherent in NHST. For instance, consider the simplest effect size statistic, the counternull [21]. This is the non-null magnitude of effect size that is supported by the same amount of evidence as the null. If we estimate the growth rate of a population, for example, our estimate might be $-15\%\ y^{-1}$, with a 95% confidence interval from -32% to $+2\%$. Some would interpret this to mean that we cannot reject the null hypothesis (of a zero growth rate). By contrast, if our estimate of decline is subject to normally distributed error, the counternull indicates that a rate of decline of 30% per annum is just as well supported as an estimate of zero. The counternull thus reminds researchers that a failure to reject the null does not mean that the null effect is more plausible than alternatives. Effect size statistics are underused in ecology and evolutionary research and, although they are seldom

Table 1. An overview of inferential approaches

Approach	Requirements	Outcomes	Advantages	Disadvantages
Null hypothesis significance testing	Data, **Y**, and a statistical null hypothesis, H_0, which designates the test statistic of interest, t	Provides $P(t \geq t_{obs}\|H_0)$, the probability of observing the test statistic (or one more extreme), if the null hypothesis is true. In carefully designed and well replicated experiments, NHST enables H_A, the converse of the null, to be falsified (if experiments repeatedly fail to reject the null with a suitably low P value)	Computational simplicity (with ready availability of user interfaces)	A variety of inherent difficulties with interpretation, as well as deeper philosophical problems that can limit scientific advances
Information theoretic model comparison	Data, **Y**, and a set of two or more competing models, $H_1 \ldots H_n$, which might include the null and its converse, a nested set of arrangements of potential predictor variables, or several disparate, mechanistic descriptions of a system	Provides an information criterion value for each competing model, usually of the form $C = -2\ln[L(H_i\|\mathbf{Y})] + B$, where C is the criterion estimate, $L(H_i\|\mathbf{Y})$ is the model likelihood and B is a penalty imposed by some aspect of the model or data (e.g. Ref. [5])	Enables models to be ranked in order of relative support; evidence ratios to be calculated for any pair of competing models; and model averaging to account for uncertainty in model selection ([6,17,27] but see Ref. [55]). Discourages binary approaches to inference	Unclear which information criterion is most appropriate [26,27] or how well some criteria perform under different conditions [55]
Bayesian statistics	Data, **Y**, a set of competing models (as above) and prior information, which might include previous estimates of the plausibility of each model, as well as their parameter values	Unique in providing $P(H_i\|\mathbf{Y})$ [4]	Including existing knowledge means that knowledge accumulates; sensitivity analyses are intrinsically accommodated (through presenting posteriors for a range of prior assumptions); all uncertainties are integrated out; and Bayesian approaches can deal with complex problems, such as those with both process and observation errors	Computational complexity, as Bayesian approaches require integrating under likelihood functions; this can render even a simple ANOVA complex in a Bayesian framework [56]
Effect size statistics	Data, **Y**, from two or more treatment groups	Measures of the practical significance of an observed effect (e.g. the difference between two means), e.g. the counternull [21], Cohen's d[a] [57], the *CL* statistic of McGraw and Wong[b] [58,59], and several other measures [58,60]	Focuses attention away from statistical significance and resultant biases; improves the potential for meta-analyses of experimental data (but see Ref. [61])	Effect size statistics are largely descriptive and, as such, are often unsatisfactory as sole measures of the outcome of an experiment

[a]A standardized descriptor quantifying, independently of sample size, the degree to which the means of two treatment groups are separated.
[b]An estimate of the probability that a randomly chosen subject in one treatment class will have a higher value of the test statistic than a random subject from another treatment class.

adequate descriptors of the outcome of research, increased use would certainly ameliorate some of the more serious errors of NHST interpretation bemoaned by some observers [14].

Bayesian statistics

The potential for using Bayesian statistics in ecology has been discussed for some time (e.g. Ref. [22]). Bayesian approaches have intuitive appeal because the rejection or acceptance of research hypotheses is fundamentally linked to previous beliefs and assumptions. By contrast, the way in which previous beliefs are incorporated into classical statistical inference is often subjective and vulnerable to bias. If we incline towards our alternative hypothesis but fail to reject the null, we often invest considerable effort in *ad hoc* explanations of why we failed to reject it [23,24]. Similarly, if we were previously sceptical about the hypothesis examined but were able to reject the null, we tend to proffer a range of alternative explanations for this finding (somewhat undermining the purpose of our study) [24]. Bayesian approaches force explicit *a priori* identification of beliefs. Incorporating prior knowledge into statistics is not the sole preserve of Bayesian statistics [4] but, unlike other approaches, prior information is required for Bayesian inference [2]. Early concerns regarding the subjective nature of Bayesian statistics [22] have largely been overcome by methods that emphasize clear, objective analyses of existing data to provide prior information (e.g. Ref. [25]), or use 'uninformative priors' where such data are not available [4].

IT techniques

Of the inferential approaches that we consider here, IT has seen the fastest incorporation into ecology and evolutionary biology. Statisticians remain divided about the best information criterion to use in any given circumstances [26,27], but Akaike's Information Criterion (AIC) [28] and its derivatives are the most common in ecological literature [5,26,29,30] and have been strongly advocated [1,6].

In spite of its increasing use, some authors have been cautious regarding the purported value of IT model selection, and have expressed concern that such approaches might replace NHST as a mechanical approach to inference [31,32]. Others, including us, have noted the difficulties of selecting a candidate set of plausible models that formalize a set of competing hypotheses, warning that

Box 2. IT: more than just sensitivity analysis?

Guthery *et al.* [26] observed that IT analyses are often used as no more than a method of sensitivity analysis or as a means to assess the relative magnitude of effects. To illustrate this point, consider the following example. Over 40 years of climate data, vegetation surveys and population data on ungulate and predator abundances have been collected in the Sikhote-Alin Biosphere Reserve in the Russian Far East. Biologists want to determine the major factors that affect annual changes in population size of red deer *Cervus elaphus* (Figure I). An array of sets of factors might all be expected to have an influence on these population changes, including: density dependence of various orders, availability of different food items, competition with other cervids, climatic conditions, predator abundance and indices of human impact.

Even ignoring interactions and non-linearities, this list contains 15 or more factors that can reasonably be assumed to have a measurable influence on population dynamics. Should we compare thousands of potential models that result from combinations of these factors? It is suggested that such large model sets should be unnecessary as, with a modicum of thought and insight, we should be able to rule out certain combinations [51]. However, with good prior reasons to suppose that all of the posited influences are likely to have an effect, which should we rule out? Arguably, to suggest that any one of the factors listed is unlikely to have an effect would be tantamount to the use of 'silly nulls' (one of the criticisms of the thoughtless use of NHST [1]).

Although an 'all subsets' comparison of models is likely to offend both opponents [26] and proponents [51] of IT–AIC, we contend that it can still be valuable. An IT–AIC comparison of all submodels will reveal which of the factors in the global model appear to have weak effects on the response and, thus, which should be eliminated to reduce bias in the remaining model. The factors with the strongest effects will appear consistently in well supported models, whereas those with less weight will occur in fewer of the better models. Most importantly, the IT–AIC process will enable model averaging [6,27], an approach that is not associated with NHST techniques, such as likelihood ratio tests. Model averaging enables uncertainty in model choice to be explicitly incorporated in the findings and substantially reduces the extent of bias that would arise from basing inference on a single 'best' model (e.g. Ref. [17] but see Ref. [55]).

There is a risk that an all-subsets approach can fit biologically implausible models, but this problem is common to most statistical approaches. For example, in classical approaches to parameter estimation, we do not include prior information. Thus, implausible parameters (or models encoding implausible biological relationships) can emerge from classical statistical analysis. No inferential approach has an innate 'understanding' of biology; this is where the researcher's own judgement is crucially important, regardless of the approach used.

Figure I. Red deer in Sikhote-Alin Biosphere Reserve. Reproduced with permission from D. Miquelle.

IT approaches do not inherently motivate the development of such a set [30,33]. Others counter that this step should be difficult; this is the fundamental stage in the process and, as such, is where skilled scientists should make the greatest contribution [6].

The foregoing are concerns with the application of IT, rather than philosophy. However, a more recent critique went further, suggesting that IT model comparisons were of little value apart from as a method of sensitivity analysis or as a means to assess the relative magnitude of effects [26]. One reason for this criticism is that many studies consider a set of competing models that are nested submodels of a single, global model. Competing models thus represent special cases of the global model, in which one or more parameter coefficients are zero. Previously, such cases were often analysed using likelihood ratio tests (e.g. Ref. [34]) but, owing to several advantages, IT approaches are becoming more common. Those advantages suggest that, even where all factors in a global model have prior support, IT model comparison brings with it substantial benefits through bias reduction, as well as an explicit acknowledgement of uncertainty in model selection (Box 2).

What statistical approaches should be in common use?

Three main objectives of statistical analysis are apparent to us: assessing descriptive findings, generating predictive models and challenging research hypotheses (Table 2).

Descriptive science

In spite of the enormous growth in experimental ecology [35], a large proportion of published ecological work is descriptive. Here, inferential statistics are often used to elevate the observational above the anecdotal. In this context, effect sizes are commonly presented in an exploratory, *ad hoc* fashion. Effects might not have been overtly presented in the context of hypotheses posed at the outset but, nonetheless, they might be worth reporting and can stimulate the development of new hypotheses. These types of conjecture on the existence of a pattern or effect correspond with the 'existential hypotheses' of Guthery *et al.* [36]. Such explorations are usually exercises in parameter estimation and do not require the language of hypothesis falsification. Indeed, this language can be entirely inappropriate. Consider, again, our example of estimating population growth rates. Estimates of population decline are particularly likely to be non-significant in populations that are already small (and thus provide low statistical power when surveyed) [37]. In these cases, we can ill-afford to focus on the need to reject a null hypothesis of 'no effect'.

Clark [2] notes that, in simple cases, traditional approaches to parameter estimation often give results and confidence regions that are similar to those provided by more complex analyses, such as Bayesian approaches with uninformative priors. In these cases, traditional analyses will generally suffice, and it will usually be necessary

Table 2. Summary of statistical applications and suggested analytical approaches

Application	Type of data	Suggested methods	Refs
Descriptive or exploratory analyses	Experimental or exploratory data examined only to gain insight	NHST or IT with effect size statistics	[33,60]
	As above but with prior information on estimable parameters	Likelihood-based or Bayesian	[4,25]
Fitting predictive models	Experimental or observational	IT	[6,17]
	As above but for complex systems or with prior information	Bayesian	[2,7]
Challenging research hypotheses	Rigorous experimental data enabling clear assessment of binary predicted outcomes	NHST	
	Experimental or observational data with a range of possible explanations	IT	[6]

only to present the means with accompanying confidence intervals and sample sizes. Where more complete information on the estimated parameter is desirable, authors can include graphs of the likelihood function. For more complex exercises in parameter estimation, especially where prior information is available, likelihood-based [4] or Bayesian approaches [25] are preferable.

Fitting predictive models

Many applied ecological questions require prediction. For example, researchers are often required to predict the habitat that will be of most value to a species. This is also essentially a question of description, rather than an exercise in hypothesis falsification: data are collected on the presence–absence of the species at some appropriate spatial scale, together with environmental variables collected at the same resolution, and deemed likely to describe patterns of distribution (e.g. Ref. [38]). To formalize that description, measured variables that appear to have the greatest effect on presence are selected from those available, excluding those with small or ambiguous effects. This is an important example of where effect size estimation will be useful and where, through their overt recognition of uncertainty in model selection, IT techniques will be valuable [17].

Although we have emphasized ecological examples, model selection problems are also a common feature of evolutionary studies. Selecting the best-justified model (phylogeny) and its parameters (branch lengths) for a given set of DNA sequences is a complex, multi-dimensional problem. Although IT approaches have been used in this area (e.g. Ref. [39]), the field is dominated by increasingly refined Bayesian methods [40–42]. Ecologists are beginning to follow this example, and Bayesian approaches are also used increasingly for highly complex ecological problems, such as those involving population dynamics [43–45] and species distributions [46].

NHST might also have a role in model selection. In particular, many statisticians view NHST as being useful for model criticism, the assessment of whether a model is 'good enough' without the use of an explicit alternative. With this approach to model assessment, NHST is not an exercise in hypothesis falsification, but can flag a model as being in need of elaboration or modification.

Challenging research hypotheses

The final application of statistics that we distinguish here is that of challenging research hypotheses. This is the area that has led to most dispute, largely because of philosophical differences regarding whether individual hypotheses should be subjected to falsification (the primary aim of NHST [24,47]), or whether multiple competing hypotheses should be compared simultaneously to assess their relative support from the data (the primary aim of IT and Bayesian model comparison approaches) [3]. Key points to note are that the method of multiple working hypotheses [48] does not require that all hypotheses be evaluated simultaneously [31] and that falsification is regarded as a strong use of NHST [14]. Rigorous experimentation can rule out working hypotheses sequentially. Attempting to falsify hypotheses individually can encourage clarity in identifying, differentiating and designing rigorous experiments to test those hypotheses, as exemplified in much behavioural work in which multiple causal hypotheses are possible (e.g. see Ref. [49] for a recent example).

For those biologists who are philosophically inclined towards a less binary approach and who favour an overt acknowledgement of imperfections of knowledge [48], IT approaches are likely to be more attractive. Here, defining models that encapsulate distinctly different hypotheses (rather than specific cases of a single, over-arching hypothesis) is the key, demanding step in the process. Caley and Hone show how this can be achieved [50].

In spite of the philosophical differences between NHST and model selection, we suspect that the inferences they produce regarding research hypotheses will often be similar. For example, Burnham and Anderson stress that weight of evidence, rather than falsification, is the aim of model selection [51]. Nevertheless, for most researchers, models for which information criteria provide 'poor support' relative to their competitors will be considered falsified to the same extent as hypotheses that NHST demonstrates are less plausible than the null.

Conclusions

Most authors continue using NHST, and many have argued for its merits [14,22,34,52–54]. We believe that NHST will continue to have an important role in statistical education, exploratory and descriptive studies, and carefully controlled experiments that explicitly aim to falsify one of a set of competing hypotheses. In spite of this, we are increasingly convinced of the need to move on from an overwhelming reliance on NHST. We must become more conscious of the type of science that we are engaged in and the relevance of alternative methods of inference when the falsification of hypotheses is not our aim. Although we believe that NHST will, and should, remain in use within ecology and evolution, we are increasing our reliance on other approaches. We hope that a proportion of readers will be similarly motivated to reconsider their approaches to inference.

Acknowledgements

We thank R. Anderson-Sprecher, I. Cuthill, R. Freckleton, A. Goldsmith, P. Green, J. Greenwood, M. Hornocker, D. McDonald, S. Mills, S. Richards and A. Russell for advice and interesting discussions, and three anonymous referees for constructive criticisms of earlier drafts.

References

1 Anderson, D.R. *et al.* (2000) Null hypothesis testing: problems, prevalence, and an alternative. *J. Wildl. Manage.* 64, 912–923
2 Clark, J.S. (2005) Why environmental scientists are becoming Bayesians. *Ecol. Lett.* 8, 2–14
3 Hilborn, R. and Mangel, M. (1997) *The Ecological Detective: Confronting Models with Data,* Princeton University Press
4 Hobbs, N.T. and Hilborn, R. (2006) Alternatives to statistical hypothesis testing in ecology: a guide to self teaching. *Ecol. Appl.* 16, 5–19
5 Johnson, J.B. and Omland, K.S. (2004) Model selection in ecology and evolution. *Trends Ecol. Evol.* 19, 101–108
6 Burnham, K.P. and Anderson, D.R. (2002) *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach,* Springer-Verlag
7 Huelsenbeck, J.P. *et al.* (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294, 2310–2314
8 Yoccoz, N.G. (1991) Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bull. Ecol. Soc. Am.* 72, 106–111
9 Peterman, R.M. (1990) The importance of reporting statistical power: the forest decline and acidic deposition example. *Ecology* 71, 2024–2027
10 Johnson, D.H. (1999) The insignificance of statistical significance testing. *J. Wildl. Manage.* 63, 763–772
11 Palmer, A.R. (2000) Quasireplication and the contract of error: lessons from sex ratios, heritabilities and fluctuating asymmetry. *Annu. Rev. Ecol. Syst.* 31, 441–480
12 Johnson, D.H. (2002) The role of hypothesis testing in wildlife science. *J. Wildl. Manage.* 66, 272–276
13 Carver, R.P. (1978) The case against statistical significance testing. *Harv. Educ. Rev.* 48, 378–399
14 Cohen, J. (1994) The earth is round ($P < .05$). *Am. Psychol.* 49, 997–1003
15 Johnson, D.H. (2005) What hypothesis tests are not: a response to Colegrave and Ruxton. *Behav. Ecol.* 16, 323–324
16 Snedecor, G.W. and Cochran, W.C. (1967) *Statistical Methods,* Iowa University Press
17 Whittingham, M.J. *et al.* (2006) Why do we still use stepwise modelling in ecology and behaviour? *J. Anim. Ecol.* 75, 1182–1189
18 Fidler, F. *et al.* (2005) Toward improved statistical reporting in the journal of consulting and clinical psychology. *J. Consult. Clin. Psychol.* 73, 136–143
19 Heyes, C.M. (1993) Anecdotes, training, trapping and triangulating: do animals attribute mental states? *Anim. Behav.* 46, 177–188
20 Huston, M.A. (1997) Hidden treatments in ecological experiments: re-evaluating the ecosystem function of biodiversity. *Oecologia* 110, 449–460
21 Rosenthal, R. and Rubin, D.B. (1994) The counternull value of an effect size: a new statistic. *Psychol. Sci.* 5, 329–334
22 Dennis, B. (1996) Discussion: should ecologists become Bayesians? *Ecol. Appl.* 6, 1095–1103
23 Meehl, P.E. (1967) Theory testing in psychology and physics: methodological paradox. *Philos. Sci.* 34, 103–115
24 Peters, R.H. (1991) *A Critique for Ecology,* Cambridge University Press
25 McCarthy, M.A. and Masters, P. (2005) Profiting from prior information in Bayesian analyses of ecological data. *J. Appl. Ecol.* 42, 1012–1019
26 Guthery, F.S. *et al.* (2005) Information theory in wildlife science: critique and viewpoint. *J. Wildl. Manage.* 69, 457–465
27 Link, W.A. and Barker, R.J. (2006) Model weights and the foundations of multimodel inference. *Ecology* 87, 2626–2635
28 Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (Csaki, F., ed.), pp. 267–281, Akademiai Kiado
29 Rushton, S.P. *et al.* (2004) New paradigms for modelling species distributions? *J. Appl. Ecol.* 41, 193–200
30 Eberhardt, L.L. (2003) What should we do about hypothesis testing? *J. Wildl. Manage.* 67, 241–247
31 Guthery, F.S. *et al.* (2001) The fall of the null hypothesis: liabilities and opportunities. *J. Wildl. Manage.* 65, 379–384
32 Sanderson, M.J. (2005) Where have all the clades gone? A systematist's take on inferring phylogenies. *Evolution* 59, 2056–2058
33 Stephens, P.A. *et al.* (2005) Information theory and hypothesis testing: a call for pluralism. *J. Appl. Ecol.* 42, 4–12
34 Quinn, G.P. and Keough, M.J. (2002) *Experimental Design and Data Analysis for Biologists,* Cambridge University Press
35 Resetarits, W.J. and Bernardo, J. (1998) *Experimental Ecology: Issues and Perspectives,* Oxford University Press
36 Guthery, F.S. *et al.* (2004) In my opinion: hypotheses in wildlife science. *Wildl. Soc. Bull.* 32, 1325–1332
37 Greenwood, J.J.D. and Robinson, R.A. (2006) Principles of sampling. In *Ecological Census Techniques* (Sutherland, W.J., ed.), pp. 11–85, Cambridge University Press
38 Manly, B.F.J. *et al.* (2002) *Resource Selection by Animals: Statistical Design for Field Studies,* Kluwer Academic Publishers
39 Strimmer, K. and Rambaut, A. (2002) Inferring confidence sets of possibly misspecified gene trees. *Proc. R. Soc. B* 269, 137–142
40 Huelsenbeck, J.P. and Ronquist, F. (2001) MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755
41 Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574
42 Opgen-Rhein, R. *et al.* (2005) Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. *BMC Evol. Biol.* 5, 6
43 Newman, K.B. *et al.* (2006) Hidden process models for animal population dynamics. *Ecol. Appl.* 16, 74–86
44 Clark, J.S. and Bjornstad, C.N. (2004) Population time series: process variability, observation errors, missing values, lags, and hidden states. *Ecology* 85, 3140–3150
45 Clark, J.S. *et al.* (2005) Hierarchical Bayes for structured, variable populations: from recapture data to life-history prediction. *Ecology* 86, 2232–2244
46 Latimer, A.M. *et al.* (2006) Building statistical models to analyze species distributions. *Ecol. Appl.* 16, 33–50
47 Medawar, P.B. (1996) Hypothesis and imagination. In *The Strange Case of the Spotted Mice* (Medawar, P.B., ed.), pp. 12–32, Oxford University Press
48 Chamberlin, T.C. (1965) The method of multiple working hypotheses. *Science* 148, 754–759
49 Pompilio, L. *et al.* (2006) State-dependent learned valuation drives choice in an invertebrate. *Science* 311, 1613–1615
50 Caley, P. and Hone, J. (2002) Estimating the force of infection; *Mycobacterium bovis* infection in feral ferrets *Mustela furo* in New Zealand. *J. Anim. Ecol.* 71, 44–54
51 Anderson, D.R. and Burnham, K.R. (2002) Avoiding pitfalls when using information-theoretic methods. *J. Wildl. Manage.* 66, 912–918
52 Hagen, R.L. (1997) In praise of the null hypothesis statistical test. *Am. Psychol.* 52, 15–24
53 Robinson, D.H. and Wainer, H. (2002) On the past and future of null hypothesis significance testing. *J. Wildl. Manage.* 66, 263–271
54 Fidler, F. *et al.* (2006) Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. *Conserv. Biol.* 20, 1539–1544
55 Richards, S.A. (2005) Testing ecological theory using the information-theoretic approach: examples and cautionary results. *Ecology* 86, 2805–2814
56 Box, G.E.P. and Tiao, G.C. (1973) *Bayesian Inference in Statistical Analysis,* John Wiley & Sons
57 Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences,* Lawrence Earlbaum Associates
58 McGraw, K.O. and Wong, S.P. (1992) A common language effect size statistic. *Psychol. Bull.* 111, 361–365
59 Vargha, A. and Delaney, H.D. (2000) A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *J. Educat. Behav. Stat.* 25, 101–132
60 Rosnow, R.L. and Rosenthal, R. (2003) Effect sizes for experimenting psychologists. *Can. J. Exp. Psychol.* 57, 221–237
61 Ray, J.W. and Shadish, W.R. (1996) How interchangeable are different estimators of effect size? *J. Consult. Clin. Psychol.* 64, 1316–1325