Behav Ecol Sociobiol (2011) 65:57–68
DOI 10.1007/s00265-010-1040-y

REVIEW

Issues in information theory-based statistical inference—a commentary from a frequentist's perspective

Roger Mundry

Received: 14 June 2010 / Revised: 22 July 2010 / Accepted: 1 August 2010 / Published online: 18 August 2010

Abstract After several decades during which applied statistical inference in research on animal behaviour and behavioural ecology has been heavily dominated by null hypothesis significance testing (NHST), a new approach based on information theoretic (IT) criteria has recently become increasingly popular, and occasionally, it has been considered to be generally superior to conventional NHST. In this commentary, I discuss some limitations the IT-based method may have under certain circumstances. In addition, I reviewed some recent articles published in the fields of animal behaviour and behavioural ecology and point to some common failures, misunderstandings and issues frequently appearing in the practical application of IT-based methods. Based on this, I give some hints about how to avoid common pitfalls in the application of IT-based inference, when to choose one or the other approach and discuss under which circumstances a mixing of the two approaches might be appropriate.

Keywords Akaike's information criterion · Null hypothesis significance testing · Data dredging

Introduction

For several decades, the statistical methods applied in biology have been dominated by the concept of null hypothesis significance testing (NHST), in which a 'null hypothesis' is rejected or not, based on a *P* value. Since approximately the 1990s of the last century, a new approach became increasingly popular which does not test a null hypothesis but compares several competing hypotheses (i.e. models) using information theory-based criteria. Among these one or several models are then selected as the most plausible one(s) generating the data observed and as the models with the best expected predictive capability. Whether or not the new approach could or should replace or complement the old one has been hotly debated since then (e.g. Johnson 2002; Guthery et al. 2005; Stephens et al. 2005; Steidl 2006; Lukacs et al. 2007; Sleep et al. 2007).

In this commentary, I discuss some of the issues coming along with the use of information theory (IT)-based statistical inference and point to some frequent misconceptions in its practical application in animal behaviour and behavioural ecology. Throughout, I shall tend to take a frequentist's perspective (i.e. that of someone advocating the classical NHST approach). That is, I shall hint at problems which, from a frequentist's point of view, might arise when applying the new approach.

The classical concept of statistical inference

For several decades now, statistical inference drawn about behavioural data has been strongly dominated by null hypothesis significance testing. Using this approach, one states a null hypothesis of (usually) no difference or no relation and calculates the summed probability of getting data (given the null hypothesis) deviating from that null hypothesis at least as much as the actually observed data. The resulting probability (i.e. the *P* value) is the probability of observing data at least as extreme as those actually

Communicated by L. Garamszegi

This contribution is part of the Special Issue 'Model selection, multimodel inference and information-theoretic approaches in behavioural ecology' (see Garamszegi 2010).

R. Mundry (✉)
Max Planck Institute for Evolutionary Anthropology,
Deutscher Platz 6,
04103 Leipzig, Germany
e-mail: roger_mundry@eva.mpg.de

obtained, given the null hypothesis being true (see, e.g. Siegel and Castellan 1988; Sokal and Rohlf 1995; Zar 1999; Quinn and Keough 2002). The question of how to draw inference based on such a *P* value has been a matter of debate since the advent of this approach. While some have suggested making dichotomous decisions (rejecting or accepting the null hypothesis; but see below) based on the *P* value being below or not below a certain threshold (usually 0.05), others have proposed to treat *P* values as continuous measures of evidence against the null hypothesis (reviewed in, e.g. Stoehr 1999; Quinn and Keough 2002; Hurlbert and Lombardi 2009). Current practice in ethology, behavioural ecology and related disciplines represents a mixture of both approaches: On the one hand, we use terms like 'significant' (usually when a *P* value is below 0.05); on the other hand, many researchers report exact *P* values and consider a 'really small' *P* value (e.g. 0.001) as 'strong' or a 'marginally non-significant' *P* value (e.g. 0.06) as 'some' evidence against the null hypothesis (Quinn and Keough 2002).

Since its inception, the NHST approach has been criticised for several inherent weaknesses and for the way it is used and its results are interpreted (or better: misinterpreted). The main problems of the approach arise from what a *P* value is actually measuring (and not measuring). In fact, a *P* value expresses the probability of getting the data (or more extreme ones) given the null hypothesis being true ($P_{data}|H_0$), but neither the probability of the null hypothesis given the data ($P_{H0}|$data) nor the probability of any particular alternative hypothesis given the data ($P_{HA}|$data; see, e.g. Cohen 1994; Nickerson 2000 or Quinn and Keough 2002 for summaries). This has major implications with regard to the interpretation of a *P* value. In fact, although it is the core of the NHST approach to reject or accept the null hypothesis, such inference is actually based on a rather indirect logic: A large *P* value indicates that the data have a large probability given the null hypothesis but this is not equal to the probability of the null hypothesis given the data, and a small *P* value indicates a small probability of the data given the null hypothesis but again this is not equal to the probability of the null hypothesis (or any other) given the data. Another major criticism of NHST is centred around the fact that whether a null hypothesis is rejected or not depends not only on whether an effect does exist but also on the sample size (the well-known relation between power and sample size; see, e.g. Cohen 1988; Siegel and Castellan 1988; Stoehr 1999). To take this issue to the extreme, consider that given an effect of a certain magnitude, it is the sample size alone that determines the *P* value and hence whether a test reveals significance. From this perspective, a *P* value seems to be a completely pointless measure. However, it has been frequently suggested to complement *P* values with measures of effect size as well as point estimates and confidence intervals or standard errors of the effects considered, which aim to indicate the practical relevance of the phenomenon investigated in addition to its statistical significance (e.g. Stoehr 1999; Anderson et al. 2001; Nakagawa and Cuthill 2007; Garamszegi et al. 2009). Based on this, one could, for instance, neglect the practical relevance of a small but statistically significant effect. Further criticisms of the NHST approach largely refer to its misuse and misinterpretation (both presumably resulting from the aforementioned weaknesses of the approach, Johnson 1999) by researchers in various areas (for summaries of the criticisms of NHST, see, e.g. Cohen 1994, Johnson 1999, Nickerson 2000 or Stephens et al. 2007).

The information theory-based approach to statistical inference

Since approximately a decade, a new approach to statistical inference has become increasingly popular in applied statistics used in research in ecology and, to a lesser extent, also animal behaviour (Garamszegi et al. 2009; Garamszegi 2010; Richards et al. 2010). This approach differs from NHST very fundamentally in how inference about the data is drawn. It is based on comparing relative measures of support for several different (usually competing) models, each representing a hypothesis. More practically, one tries to find, for each of the models to be investigated, those parameters of the model which best explain the data (with the 'parameters' being, e.g. estimated coefficients and residual standard deviation). The probability of the data (i.e. the product of their point probabilities) given the model and its parameter values then reveals the 'likelihood' of the specific parameter values (which are usually searched for by maximising the likelihood; see below). Obviously, the explanatory power (i.e. the likelihood) revealed for a certain model depends on the number of predictor variables it includes (and actually increases with the number of predictor variables). For instance, in the case of a linear regression, the fit of a model with two predictor variables will invariably be better than the fit of a model with only one of them. However, the increased model fit, achieved by adding a parameter, comes along with the cost of increased model complexity. In fact, by adding more and more parameters, one can easily achieve a perfect fit, but at the same time, this fit is likely to be totally trivial. IT-based inference aims in compromising between model fit and model complexity. It compares different models and searches for those which are most parsimonious, i.e. which represent the best compromise between explanatory value and simplicity. This requires being able to compare the fit

of different models with different numbers of predictor variables. This is done by penalizing the likelihood for the number of predictor variables (more precisely, the number of estimated parameters) in the model. The most basic way of practically doing this is using Akaike's information criterion (AIC), which equals twice the negative natural logarithm of the likelihood associated with the model, plus twice the number of estimated parameters. The AIC is then taken as a measure of relative support for a certain model (where a smaller AIC indicates a 'better' model; see Burnham and Anderson 2002 for an introduction). Information criterion-based model selection is supposed to compromise well between underfitting and bias, on the one hand, and overfitting and variance (i.e. uncertainty in parameter estimates), on the other hand. Choosing models based on, e.g. the AIC means to search for a parsimonious model representing a good trade off between bias and variance (Burnham and Anderson 2002), or between model simplicity and model fit.

Among the major advantages of the approach is that it allows to account for uncertainty in the decision about whether to assume a certain model to be effective (or the best in a set of candidate models; e.g. Burnham et al. 2010). In fact, one may investigate a number of competing models and draw inference from the entire set of models or a subset of it ('Multi-Model Inference' (MMI); see Burnham and Anderson 2002 for an introduction): Models with similarly small AIC have similar explanatory value, and one would not (or does not have to) opt for one of them to be the 'best' model (and reject all the others) but can compare the relative support for each of them (usually measured using 'Akaike weights'). Similarly, one can also measure the relative level of support for each of the variables investigated. Hence, one does not (have to) make dichotomous decisions (accept or reject a null hypothesis about a certain model or variable) but collects relative support for a range of models and potentially selects several of them as likely candidates for the 'best' model in the set (for details, see Burnham and Anderson 2002, and for quick introductions, see, e.g. Johnson and Omland 2004, Burnham et al. 2010 or Symonds and Moussalli 2010).

An essential prerequisite of the IT and MMI approach is careful a priori selection of the models investigated. The recommendation most frequently given is that each of the models to be investigated (and hence each of the predictor variables to be included therein) requires good empirical and/or theoretical support (Burnham and Anderson 2002). However, it has also been suggested that under certain circumstances all possible models that could be built out of a set of variables may be investigated (Stephens et al. 2007), though some (e.g. Burnham and Anderson 2002) have termed this approach 'unthoughtful' (p. 39) or 'poor strategy' (p. 147). For strategies of obtaining a candidate set, see, e.g. Burnham et al. (2010) and Dochtermann and Jenkins (2010), and for more general discussions of this topic, see, e.g. Hegyi and Garamszegi (2010) or Symonds and Moussalli (2010).

Differences and communalities of NHST- and IT-based inference

It is important to note that NHST- and IT-based MMI differ only with regard to the questions asked, what conclusions are drawn and how inference is made, but not at all with regard to model fitting (at least regarding the results). Having, for instance, one continuous predictor variable and one continuous response variable and assuming a linear relation between the two (i.e. a simple regression with response $= c_0 + c_1 \times$ predictor), the estimated values for the two coefficients (c_0 and c_1) would be exactly equal, regardless of whether one uses 'ordinary least squares' or maximum likelihood to find them. The same applies to any ordinary least squares based statistic in NHST, i.e. any general linear model (obviously this implies that both procedures also have the same assumptions, e.g. regarding independent and normal errors with constant variance). Having a response variable for which ordinary least squares are not suitable (e.g. a binary or count response), a frequentist would also use maximum likelihood to get the estimated values of the coefficients (i.e. a 'Generalized Linear Model', McCullagh and Nelder 2008). Hence, for any given model, the way that its parameters (i.e. coefficients) are estimated is essentially the same in NHST- and IT-based inference, and the obtained values of the coefficients associated with the predictor variables are virtually identical.

The main difference between NHST- and IT-based inference lies in the kind of questions addressed, how and about what inference is drawn and what it reveals. In NHST, one asks whether a null hypothesis should be rejected or not and (usually) compares one model with the null hypothesis (usually the 'null model' not comprising any predictor variable): If the summed probability of the observed and all other more 'extreme' relationships between the predictor variable(s) and the response variable is sufficiently small (e.g. ≤ 0.05), one rejects the null hypothesis and concludes that the data are unlikely to be generated by chance alone and that the predictor variable(s) do affect the response variable, otherwise not. In IT-based inference, one does not reject or accept hypotheses but measures and compares the relative support that each model (i.e. hypothesis) in a set of candidate models receives. In other words, NHST (sensu Neyman–Pearson; see Stoehr 1999) addresses the question 'what should I do?' (i.e. rejecting the null hypothesis or not) whereas IT addresses

the question 'what does this observation tell me about A *versus* B?' (i.e. which hypothesis receives more or the most relative support; Royall 1997).

A further difference between NHST- and IT-based inference is that IT-based methods are mainly designed and available to compare models fitted using maximum likelihood, whereas NHST can be used in a variety of other contexts and for other questions addressed. For instance, ordination methods such as Principal Components and Principal Coordinates Analysis do, to my knowledge, only exist in the framework of NHST, and also rank-based non-parametric statistics do not have a directly comparable counterpart in the framework of IT-based analysis.

General problems coming along with the use of IT-based inference

Proponents of the IT-based approach have frequently stated that it is superior to NHST when analysing complex and observational ecological data sets with several predictor variables (Burnham and Anderson 2002; Stephens et al. 2007). Occasionally, it has also been stated that IT-based inference is generally superior to NHST (e.g. Lukacs et al. 2007). Hence, the question arises whether we should generally refrain from using NHST and rather use IT-based inference instead. For several reasons, I do not believe that such a procedure is adequate in every case.

First of all, it is a core assumption of IT-based inference that all candidate models are theoretically and/or empirically well-founded (Burnham and Anderson 2002; Burnham et al. 2010; but see Stephens et al. 2007 and Burnham and Anderson 2002, p. 168; note that also in case of an NHST analysis thorough thinking about the models investigated is required). From my understanding, this implies that also the predictor variables included therein are well-founded. This, in turn, requires good knowledge of the system investigated. The systems investigated in behaviour and behavioural ecology, though, are not generally or necessarily well understood. Measuring the 'quality' of an individual or its territory, for instance, is frequently complicated by first being a complex phenomenon comprising a variety of different aspects and second being not very well understood (in the sense that we often do not know what exactly determines a 'high quality' individual or territory). As a consequence, the use of so-called proxies is frequently seen. With a proxy, I here refer to a variable not being the variable of interest in itself (e.g. territory quality) but another variable (e.g. average temperature) assumed (or hoped) to be somewhat correlated with the variable of interest. The point I wish to make is that in research about, for example, animal behaviour or ecology, we frequently are not very confident of whether the (statistical) predictor variables we investigate affect the phenomenon under study at all. In fact, I am convinced that many studies of animal behaviour or behavioural ecology are 'experimental' in the sense that a priori it is completely unknown whether the predictor variables used will show to have any effect on the phenomenon under study. In such a situation, though, I believe that an approach which requires a good foundation of the models investigated and hence the variables out of which they are built to not generally or necessarily be fully appropriate (see also Steidl 2006 or Symonds and Moussalli 2010).

Closely related to this aspect is the question of whether it makes sense at all to test null hypotheses of no relation or no difference (also referred to as 'nil hypothesis', e.g. Cohen 1994). In fact, one of the main criticisms of NHST is that null hypotheses are always wrong (e.g. Johnson 1999; see also review by Nickerson 2000). And indeed, given an effect of a certain magnitude, the *P* value is solely a function of the sample size, and given the latter being large enough, a statistical test will invariably reveal significance (see above). However, this statement only holds as long as there is an effect. While for certain phenomena in certain research areas (e.g. in ecology) it can be assumed 'that everything is connected to everything else' (Johnson 1999), I argue that this is not generally the case. With regard to the null hypotheses tested in research on animal behaviour, for instance, it seems reasonable to assume that a considerable proportion of these are actually true. In principle, this seems plausible because many of the predictor variables we investigate are proxies (as described above) which we hope or believe to be related to what we actually want to measure (but cannot directly, due to its complexity or our lack of knowledge). As a consequence, it is likely to happen (not too rarely, presumably) that such a proxy turns out to not show the slightest indication of an effect. I am also sure that most readers of this article will already have made the disappointing experience of not having found the slightest indication of a null hypothesis investigated being wrong. I believe that such an outcome is usually due to insufficient variables (i.e. 'proxies') and lack of knowledge of the complex systems we investigate. In fact, I am fully convinced that the null hypotheses tested in, for example, ethology and behavioural ecology are not generally as trivial and wrong per se as the great example found by Johnson (1999), which states that density of large trees is unaffected by logging. Hence, establishing the statistical significance of a variable using classical NHST seems to be a reasonable exercise under certain circumstances (see also Forstmeier and Schielzeth 2010). Finally, it seems worth noting that the investigation of proxies, insufficient knowledge about the matter under investigation and the testing of presumably true null hypotheses is not specific to ethology and behavioural ecology but is likely to be an issue in many

research areas (particularly from the life sciences in their widest sense, including, for example, psychology and sociology). For recent contributions to the issue in other research areas see, for example, Ioannidis (2005), Young et al. (2009) or Vul et al. (2010).

A final issue for which NHST has been criticised and which is not a characteristic of IT-based inference is that of making binary decisions (about null hypotheses). It frequently has been argued that the NHST approach of making dichotomous decisions based on an arbitrarily selected threshold (e.g. a P value being or not being ≤ 0.05) will frequently lead to wrong conclusions (e.g. Johnson 1999; Stoehr 1999), and this is certainly true, particularly when the power of an analysis and effect sizes are not considered. On the other hand, however, proponents of NHST have also repeatedly suggested that P values should be considered as continuous variables (see summary by Stoehr 1999) measuring strength of evidence 'against' the null hypothesis, and it frequently has been suggested that the interpretation of P values should generally be accompanied by a consideration of power, measures of effect size as well as point estimates associated with their standard errors or confidence intervals (e.g. American Psychological Association 1994; Stoehr 1999; Anderson et al. 2001; Nakagawa and Cuthill 2007; Garamszegi et al. 2009). So the argument against NHST would be somewhat weakened if these suggestions were followed.

Issues and recommendations regarding the practical application of IT methods

Besides these more fundamental problems outlined above, I see several issues regarding how IT-based statistics are currently practically applied. To evaluate how widespread and/or severe these issues are, I searched the issues of *Animal Behaviour*, *Behavioral Ecology and Sociobiology* and *Behavioral Ecology* that appeared in 2008 (plus some fraction of those that appeared in 2007 and articles in press) for articles including the term 'AIC'. The total number of articles I found was 51 (22 in both *Animal Behaviour* and *Behavioral Ecology* and seven in *Behavioral Ecology and Sociobiology*). However, I did not make an attempt to do an in-depth and quantitative analysis of what I found, largely because I frequently found it hard to follow what actually has been done. Hence, the following account will largely be qualitative and particularly focus on the problems I see (i.e. it will be biased towards articles including problematic statistics). The main problems were data dredging, failure to establish explanatory value of the best model(s), poor documentation of the candidate set of models investigated, mistreatment of interactions, neglecting assumptions and several other issues. Besides treating these issues in the application of the IT approach, I shall give some recommendations regarding its use.

Model selection and significance testing

The problem

The main problem with the practical application of model selection was that it was frequently used in conjunction with significance testing in a way I consider 'data dredging'. With data dredging, I refer to an analysis which searches among a (potentially large) set of variables and models built thereof for those 'significantly' explaining the response variable(s). The problem with data dredging is that the probability of finding a model seemingly explaining the response variable well (e.g. a statistically significant one) increases with the number of variables and models investigated, even in the complete absence of any relation between the predictor variables and the response variable. The results of such an analysis are potentially completely meaningless, and without external reference (e.g. replication or cross-validation), it is impossible to assess whether this is the case. Consequently, both proponents of IT-based inference and NHST have repeatedly warned against data dredging (also referred to as, for example, a 'fishing expedition'; e.g. Lovell 1983; Chatfield 1995; Anderson et al. 2000; Burnham and Anderson 2002; Smith and Ebrahim 2002; Symonds and Moussalli 2010). Since I am adopting a frequentist's perspective, it might be worth repeating Burnham and Anderson (2002, p. 203) as proponents of an IT-based approach in this context: "... do not use AIC_C [or AIC; my addition] to rank models in the set and then test whether the best model is "significantly better" than the second-best model. The classical tests one would use for this purpose are invalid, once the model pairs have been ordered by an information criterion." Obviously, the same applies also for a comparison between the best and any other model, for example, the null model.

Nevertheless, in the 51 papers, I investigated I found at least 15 (29%) presenting at least one analysis I consider being a case of data dredging and approximately five additional papers for which this was potentially the case. Most of these selected one or several best models using an IT-based analysis and subsequently tested their significance (or the significance of the variables included therein) using classical NHST. That such an analysis has a high potential of false positives should be obvious since it, first, represents a classical case of multiple testing and, second, uses the same data to formulate a null hypothesis based on data exploration and then to test its statistical significance (e.g. Chatfield 1995).

To demonstrate that this increased potential of false positives is not somewhat negligible but actually of considerable magnitude, I used a simulation very similar

to that used by Mundry and Nunn (2009). In this simulation, I generated data sets, each comprising two to ten predictor variables and one response variable. Each data set consisted of pure random numbers being normally distributed with a mean of five and a standard deviation of four (both arbitrarily chosen). Since the sample size required for a reliable and stable result is dependent on the number of predictor variables, I set the number of cases (N) as a function of the number of predictor variables (k) with $N = 3 \times (50 + 8 \times k)$ (Field 2005). For each simulation with a given number of predictor variables, I generated 1,000 such data sets. For each data set, I analysed all models that can be built out of its predictor variables (disregarding interactions) using a standard multiple regression (assuming normally distributed and homogeneous errors) and extracted the AIC of each model. Finally, I selected the best model and tested its significance as compared to the null model using a conventional F test. When the best model was the null model, I set the P value to one. The simulation was run using a script written for R (version 2.9.1, R Development Core Team 2009), and the main functions used were 'rnorm' to get the random numbers, 'lm' to calculate the multiple regression, 'extractAIC' to get the AIC and 'anova' to compare the best with the null model. I found that the number of 'significant' best models was invariably above 5% and clearly and greatly increased with the number of predictor variables investigated (Fig. 1). To make this point perfectly clear: If I investigate, say, whether the breeding success of nightingales is related to a set of ten predictor variables, all of them being completely meaningless (e.g. average daily rainfall and temperature at a randomly chosen weather station on a continent on which nightingales do not occur, the value of the Dow Jones index at the date of individual births, the weight of the bar of soap in my bathroom on the day when the individual was included into the study etc.), then I have an almost 50% chance of getting a 'significant' finding when running an all subsets analysis, selecting the best model using AIC and then testing its 'significance' using NHST. This is not to say that MMI or selecting a best model does not make sense. It is only to say that selecting a best model and then testing its statistical significance using classical NHST does not make any sense.

It is worth noting that at least six of the 15 papers mentioned above used AIC (or a derivative of it) in conjunction with stepwise model selection methods for at least one analysis. This was somewhat disturbing since stepwise model selection has been heavily criticised by both proponents of IT-based inference and NHST for a variety of good reasons. Among these are (a) that different methods (i.e. forward selection, backward elimination and the combination of the two) do not necessarily or generally reveal the same solution (reviewed in James and McCulloch 1990), (b) that the method does not allow for a reliable ranking of the predictor variables by their importance, (c) that the final model selected may comprise nuisance variables and does not necessarily include all important predictor variables (Derksen and Keselman 1992), (d) that the solution derived tends to be unstable in the sense that slight changes in the data set can lead to gross changes in the final model (e.g. Austin and Tu 2004) and (e) that stepwise procedures in conjunction with NHST lead to greatly inflated type one error rates (i.e. increased probability of erroneous rejection of a true null hypothesis; e.g. Freedman 1983; Mundry and Nunn 2009; see Whittingham et al. 2006 for a summary of the weaknesses of stepwise regression). None of these shortcomings of stepwise methods is affected at all by whether variables are entered and/or removed using a classical statistical significance criterion or an information criterion like the AIC (Burnham et al. 2010; Hegyi and Garamszegi 2010; Richards et al. 2010). It is worth noting in this context that proponents of IT-based inference have frequently and particularly emphasised the superiority of IT-based MMI compared to stepwise methods (e.g. Burnham and Anderson 2002; Whittingham et al. 2006; Burnham et al. 2010), and I agree with them.

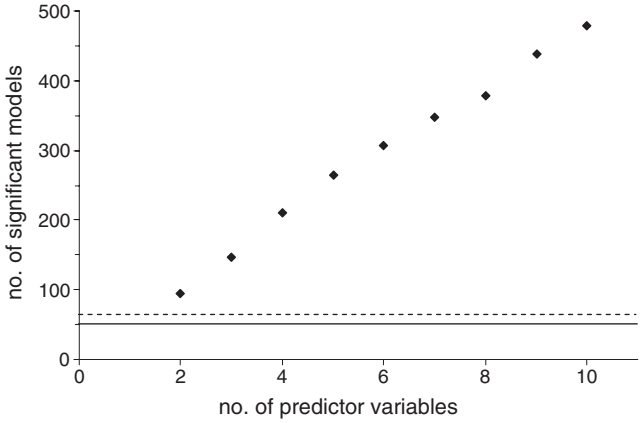

Fig. 1 Numbers of 'significant' best models (out of 1,000) as a function of the number of predictor variables and when the null hypothesis is, by definition, true (tested were pure random data). Best models were selected from all models that can be built out of the respective number of predictor variables and selected using AIC. *Symbols above the dashed horizontal line* represent proportions significantly in excess of chance expectation (50, *solid line*; binomial test). Note that the probability of getting a significant result was always clearly above the desired 5%. Counting the number of data sets for which the difference between the AIC of the null model and that of the best model was at least two revealed almost the same results (i.e. per number of predictor variables the number of data sets for which $AIC_{null} - AIC_{best} \geq 2$ differed by at most 14 from the number of 'significant' best models)

Recommendations

What follows from this is straightforward: Selecting one (or a few) best model(s) based on an IT-based approach and

then testing its (or their) statistical 'significance' using NHST does not make any sense because the probability of such a test to reveal 'significance' is much higher than the nominal error level. In fact, doing so potentially produces nothing else than publication bias, with the researcher actively grabbing one or a few 'significant' P values out of a larger number of non-significant ones being replaced by some automated formalism. Hence, such an exercise is a pointless undertaking which should generally be refrained from. Consequently, Burnham and Anderson (2002) have also repeatedly warned against mixing the two approaches (see also Burnham et al. 2010).

It is worth noting in this context that whether 'significance' is established using a classical significance test or based on effect sizes and confidence intervals does not make a difference, neither practically nor conceptually. This is the case for theoretical reasons as outlined by, for example, Chatfield (1995), who pointed out that in case of no relationships between the predictors and the response and an assessment of the explanatory value of the best model, one would be likely to overestimate this explanatory value. To illustrate this point, I ran the simulation described above (see also Fig. 1) again and this time measured R^2 (adjusted as described in Tabachnick and Fidell 2001, p. 147) of the best model, when it was not the null model. When pooling across all simulations, I found an average R^2 of 0.013 (range 0.0025–0.084). The R^2, averaged separately per number of predictor variables (2–10), ranged from 0.0128 to 0.0140 and decreased slightly but clearly with the number of predictor variables (Spearman correlation: $r_S = -0.73$, $N=9$, $P=0.03$). Although these effect sizes seem very low at a first glance, they are not far below what could be considered 'normal' effect sizes in ecology (Møller and Jennions 2002). Inspecting the estimated coefficients of the predictor variables in the best model together with their standard errors and confidence intervals revealed comparable results. Here I checked only one best model, randomly chosen from simulations with ten predictor variables, of which three where included into the best model. Again, effects are not too strong, but two of the three predictor variables have confidence intervals not including the zero (Table 1; see also Forstmeier and Schielzeth 2010).

Table 1 Estimated parameter coefficients, their standard errors, confidence intervals, t and P values

	Estimate	SE	CI	t value	P
Intercept	5.88	0.46	+4.98 to +6.78	12.85	<0.001
pv_2	0.08	0.05	−0.02 to +0.18	1.61	0.108
pv_8	−0.11	0.05	−0.21 to −0.01	−2.25	0.025
pv_10	−0.13	0.05	−0.23 to −0.03	−2.61	0.009

The model is the best model selected from all models that can be built out of ten predictor variables (not including interactions) and comprises three of them

Failure to establish explanatory value of the best model(s) (data dredging II)

An issue closely related to the previous one is failure to establish the explanatory value of the best model(s) at all. Although a perfect case of this failure did not occur in the sample of papers, I investigated (but almost) this issue seems worth mentioning because I frequently encountered researchers being confused about this point. The failure is to simply take the best model as the 'significant one' without any consideration of its explanatory value (Burnham et al. 2010). The simple reason why this is not a very useful approach is that there will always be a best model (Garamszegi 2010; Symonds and Moussalli 2010), and with an increasing number of predictor variables and models investigated, the probability of the best model being not the null model increases considerably. It seems to me that researchers are not always aware about this issue since I found, for instance, a paper stating that some models 'received substantial support' ignoring the fact that the second best model was the null model with an AIC differing from that of the best model by only 0.8 (though the authors later recognized that all confidence intervals of the variables in these models with 'substantial support' were very wide and comprised the zero).

The recommendation following from this is straightforward: Simply selecting the best model for inference without any consideration of its explanatory value should not be done. Instead, it is essential that its explanatory value is investigated (Burnham et al. 2010; Symonds and Moussalli 2010). Ideally, this would be done using replication (i.e. investigation of an independent data set). In practice, this will hardly be possible, and one will usually have to use methods such as cross-validation (i.e. checking how well the model explains data not used for deriving it) or bootstrapping (for further recommendations regarding inference, see below).

Failure to specify the set of candidate models

Another issue which I encountered in at least nine papers (18%) of the sample was that it remained unclear in at least one analysis, what exactly the set of candidate models investigated was (note that these do not include papers which applied stepwise methods). This is definitely inappropriate since in such a case it is unclear about what inference was actually drawn. A similar problem appeared when stepwise methods were used for which the set of models investigated is not predefined at all and at least with regard to what is eventually published, usually completely unclear.

The recommendation following from this issue is clear: When using IT-based inference, it is essential that the candidate set of models investigated is well-founded, carefully developed and clearly stated (Burnham and Anderson 2002; Burnham et al. 2010; Dochtermann and Jenkins 2010). From an IT-based inference on an unclear set of candidate models, one cannot really learn anything.

Difference between AIC values as a 'significance test'

I also found several papers using a difference in the AIC between two models for some kind of 'significance' test, stating, for instance, 'that a change in AIC of more than 2.00 would be considered biologically meaningful', 'assuming that a difference between models of AIC >2.00 is biologically significant' or 'the model with the highest Akaike weight was considered the best model, but only significantly so if it differed from other candidate models by at least 2 AIC_c units'. I personally do not consider this to be a very serious issue (from a frequentist's perspective, there is nothing to complain about such a decision criterion); however, I assume that most proponents of an IT-based approach would do so. For instance, Burnham and Anderson (2002, p. 203) state explicitly that one should 'avoid terms such as significant' in the context of an IT-based analysis.

It might be worth taking a closer look at what exactly is done with such a 'significance test' based on differences between AIC values. It can be easily seen that the difference in AIC between, e.g. a full and a null model (here $\Delta\mathrm{AIC} = \mathrm{AIC}_{\mathrm{null}} - \mathrm{AIC}_{\mathrm{full}}$) is tightly linked to the test statistic of the likelihood ratio test (change in deviance, ΔD) that a frequentist might use to test the full against the null model (Dobson 2002). In fact, $\Delta D = \Delta\mathrm{AIC} + 2 \times k$, with k being the number of estimated parameters (Sakamoto and Akaike 1978). Hence, the larger the ΔAIC is, the larger the ΔD. Plotting the P value obtained from the likelihood ratio test against the ΔAIC shows that the decision based on the ΔAIC (being >2 or not) and that based on the P value obtained from the likelihood ratio test (being $P \leq 0.05$ or not) are virtually identical in the case of two predictor variables, but that for other numbers of predictor variables the likelihood ratio test is more likely to reveal 'significance' (Fig. 2). As already stated, from a frequentist's perspective, there is nothing to complain about such practice. The question arises, though, as to what the advantage of an IT-based approach is when it is used in such a way. The simple answer is that there is none. In fact, such NHST-like inference based on the difference between two AIC values carries all of the shortcomings of NHST and can be criticised for exactly the same reasons (e.g. making binary and automated decisions according to an arbitrary criterion). Hence, I think one should be honest and replace such NHST-like inference by conventional NHST, if possible, and state a P value.

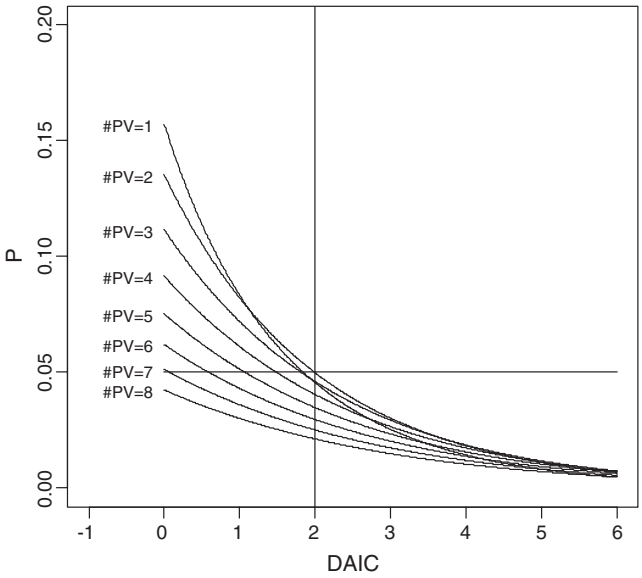

Fig. 2 Relation between P values derived from a likelihood ratio test and the difference between the AIC values derived for the full and the null model for different numbers of predictor variables. P values were derived from testing a χ^2 with k degrees of freedom, derived as $\chi^2 = \Delta\mathrm{AIC} + 2 \times k$, where k is the number of predictor variables

Other issues and recommendations

Besides the above-mentioned issues, which were all more or less specific to the practical application of IT methods, I encountered a number of further issues which are rather generic for statistical modelling approaches in general.

For instance, I encountered at least six papers (12%) presenting models which included interactions but not the respective main effects. Fitting such a model is invalid since the effect of the interaction is confounded with the effects of the two main effects (or more generally, assessment of the effect of an interaction is only possible when all the terms included in it are also in the model). In such a case, a reliable estimation of the effect of the interaction is difficult if not impossible (e.g. Cohen and Cohen 1983; Aiken and West 1991). Another issue related to interactions was that in several cases it was unclear whether continuous variables were centred before being included into an interaction. Centring (to a mean of zero) or standardising (to a mean of zero and a standard deviation of one) might be essential in such a case, though, because otherwise the continuous predictor and the interaction term might strongly correlate, leading to the adverse effects of collinearity (e.g. Tabachnick and Fidell 2001; note that the same applies if powers of a continuous variable are included). Finally, it might be worth mentioning that the interpretation of main effects involved in an interaction is

considered impossible by most statisticians (see, e.g. review by Hector et al. 2010; but see Schielzeth 2010). In addition, one should keep in mind that a main effect in a model including an interaction involving that main effect and the same main effect in a model not including such an interaction have two different meanings and interpretations.

There were also some cases in which there was some imbalance between the number of cases, on the one hand, and the number of variables and models investigated, on the other hand. An extreme case was a data set of a sample size equalling 12 which was used to investigate four predictor variables and ten models built thereof (null model not counted). Although no clear rules are available for what the number of cases should be, given a certain number of predictor variables (and models built thereof), it seems clear that the sample size should be considerably in excess of the number of predictor variables (see, e.g. Tabachnick and Fidell 2001, Harrell 2001 and Quinn and Keough 2002, or see Field 2005 who gives some guidelines about acceptable ratios of the number of cases to the number of predictor variables and Burnham et al. 2010 for an upper limit of the number of models investigated in the relation to the sample size). If this is not the case, models tend to overfit and be unstable (in the sense that even small changes in the data may lead to substantially different findings).

Furthermore, in most of the papers, I found at least some hints that the authors were aware of the assumptions of the analyses conducted, but in a few these were completely missing. However, these matter as much in an IT-based approach as in NHST, and violations of them may severely affect the conclusions drawn. Hence, one should routinely perform the standard checks required for the specific model fitting procedure used (e.g. normality and homogeneity of residuals; influence of certain cases, etc.).

An important assumption in case of having two or more predictor variables in a model is absence of collinearity. However, I found only a total of nine papers showing indications that the authors considered this to be an issue. Collinearity means that the predictor variables are (partly) interrelated (Quinn and Keough 2002; Field 2005) and in the simplest case is indicated by (some) high correlations among them. When collinearity does exist, results tend to be unstable (small differences in the analysed data may lead to large changes in the parameters estimated), and estimates of parameters have large standard errors (implying that the estimated effect of a predictor variable is associated with large uncertainty; Freckleton 2010). In the context of IT-based inference and particularly MMI, collinearity is an issue because it leads to model redundancy, which particularly affects Akaike weights (Burnham and Anderson 2002) and is likely to inflate the number of models in the confidence set. Hence, to me it seems necessary to routinely check models with several predictor variables for collinearity among them. It might be worth mentioning here that collinearity takes place among the predictor variables and has nothing to do with the response. Hence, these checks do not differ at all between, e.g. linear, logistic or Poisson regression (Field 2005).

General recommendations

Following these criticisms and recommendations regarding the practical application of IT-based inference, it might seem worth to give some more general hints regarding the choice between IT- and NHST-based inference and discuss if and how a mixture of the two methods seems appropriate. I want to emphasize, though, that these suggestions are partly preliminary and reflect my personal opinion and, that I am confident that not all these suggestions will be supported by everyone.

When to choose NHST- and IT-based inference?

To me, the main reason why the two approaches are mixed so frequently seems to be some lack of knowledge about the rationales and assumptions of the two approaches and when they are appropriate. So when are the two approaches appropriate?

In any case when one feels the necessity to state a statistical significance (i.e. a *P* value) in order to support some statement, an NHST approach is required (but this must not 'test' a best model selected using an IT-based method or the variables included in the best model). I personally feel some necessity to do such significance tests in case of an 'experimental' study (in the sense outlined above), for which it is not a priori clear whether the predictor variables investigated have any effect on the response variable(s), and I am also convinced that in animal behaviour research, such studies are common (see also Forstmeier and Schielzeth 2010). However, I also strongly suggest to follow the recommendations of others to generally refrain from reporting 'naked' *P* values (Anderson et al. 2001) but regularly complement them with measures of effect sizes, point estimates and their standard errors or confidence intervals (e.g. Stoehr 1999; Anderson et al. 2001; Nakagawa and Cuthill 2007; Garamszegi et al. 2009).

On the other hand, there are definitely situations in which NHST does not make much sense, for instance, in case of 'silly' nulls (e.g. Johnson 1999; Anderson et al. 2001) which are obviously wrong and pointless, as it might frequently be the case in ecology where the relation between two variables might be straightforward. Also when several (potentially non-nested) competing models have to be compared and the strength of evidence in favour of any

of these has to be determined, NHST does not really offer much, whereas IT-based inference offers a lot (Burnham and Anderson 2002). Finally, when prediction is a major purpose of the study, I believe that IT-based methods are clearly superior because 'model averaging' (Burnham and Anderson 2002; Johnson and Omland 2004) does not require to base predictions on a single model, selected based on partly arbitrary thresholds (e.g. *P* values; for a quick introduction, see Symonds and Moussalli 2010). Instead, IT-based model averaging elegantly allows for accounting for uncertainty in model selection (but see Richards et al. 2010). In all these situations, IT-based inference seems clearly superior to me. However, it seems important to emphasise that such analyses do not require classical significance tests, and that classical significance tests in such studies do not offer any additional insights.

Occasionally, it has been suggested that the IT-based approach might be preferable in case of observational studies whereas NHST is the better option in case of an experimental study (Burnham and Anderson 2002; see also Stephens et al. 2007). However, I personally do not find this distinction to be very helpful because, on the one hand, observational studies can be 'quasi-experimental' and, on the other hand, experimental studies can be confounded by a whole set of rather uncontrollable confounding variables turning them, in fact, into 'quasi-experimental' studies. For instance, when testing for age effects in the song of birds, one could do a longitudinal study on wild individuals (i.e. an observational study) investigating each subject in its first and second season. In such a case, potentially confounding variables would be rather well accounted for by incorporating individual identity into the model. An experimental study, on the other hand, may become rather 'observational', for instance, when the number of individuals is limited (and hence one has to take the individuals available), and some potentially confounding variables like age, sex, litter or prior testing experience should be incorporated into the statistical model. Hence, it seems that a decision between one or the other approach is best driven by the specific question addressed rather than by the somewhat arbitrary distinction between experimental and observational studies (see also Stephens et al. 2005; Garamszegi 2010).

Mixing the two approaches

In my opinion, some scenarios do exist in which both approaches can coexist. For instance, IT-based inference usually requires certain assumptions to be fulfilled, and I do not see any reason why, for instance, the normality of residuals should not be tested by using, for example, a Kolmogorov–Smirnov test, in the context of an IT-based analysis (though I frequently have the impression that eyeballing the distribution of the residuals is superior to such a formal test; see also Quinn and Keough 2002). Another situation in which I think mixing of the two approaches might be appropriate is when one wants to use NHST to test the effect of some variables but there are some additional potentially confounding variables to be controlled for. Since including many variables into a model might create overfitting and lead to inflated variance (e.g. Burnham and Anderson 2002), a potential strategy to choose a parsimonious combination of the control variables could be an IT-based approach, also when the final inference about the test variables (but not of the control variables) is based on NHST. A similar situation is one in which different error structures (e.g. Poisson or Gaussian; for an example, see Richards et al. 2010), or different ways of controlling for autocorrelation, are available and one needs to choose the most appropriate one. Here one could potentially choose the error structure or the way autocorrelation is controlled for using an IT-based method and then use NHST to investigate the effects of some other predictor variables. However, more research is needed before such an approach of choosing a parsimonious set of control variables or an appropriate error structure can be trusted with regard to the error rate in an NHST analysis.

There seems to be also an avenue potentially reconciling IT- and NHST-based inference: In the simulation described above, I also included a check of whether the null model (comprising only the intercept but none of the predictors) was in the 95% confidence set (being defined based on summed Akaike weights as described in Burnham and Anderson 2002, p. 169, or Symonds and Moussalli 2010). I found that of 1,000 simulations conducted per number of predictor variables, at most 57 revealed a confidence set which included the null model (average, 46.6). Hence, the probability of the 95% confidence set to comprise the null model was very similar to the error level conventionally applied in NHST (i.e. 5%). Moreover, the number of confidence sets comprising the null model did not obviously correlate with the number of predictor variables (r_S= −0.28, N=9, P=0.47). Hence, it seems that drawing inference based on best models or confidence sets only when the confidence set does not comprise the null model does prevent false positives in the sense of classical NHST (note that the difference in AIC between the best and the null model does not reveal such an option, since in my simulation the probability of $AIC_{null}-AIC_{best}$ to reveal at least two was very similar to the probability of the best model to be 'significant'; see also Fig. 1; see also Burnham et al. 2010 or Dochtermann and Jenkins 2010). I want to stress, though, that this suggestion is very preliminary, and further research is definitely needed before such an approach can be trusted. I also want to emphasize that this is not my idea. In fact, I know several researchers routinely checking whether the null model is in the confidence set,

but I am not aware of a reference suggesting such an approach (but see Burnham et al. 2010). Finally, I want to emphasise that I am confident that several proponents of the IT-based approach would not consider this to be a valid or even necessary procedure. In fact, Burnham and Anderson (2002) warned against routinely including the null model into the set of models investigated.

Finally, I personally would consider it a valid procedure when first the full model is tested using classical NHST, and once this revealed significance, the most parsimonious model (or a set of models) is selected using an IT-based approach (see also Symonds and Moussalli 2010; for more reasons to inspect the full model, see Forstmeier and Schielzeth 2010). But again, I am confident that many proponents of IT-based inference would consider this an unnecessary exercise.

Concluding remarks

To summarise, I do not want to argue in favour or against one of the two approaches. I believe that both have their justifications, are useful under certain circumstances, have specific strengths and limitations, and will probably coexist (presumably supplemented by Bayesian inference, Garamszegi 2010) for a long period.

Given the advances in our understanding of, e.g. ecology and behaviour in the past decades, the NHST approach definitely proved to be useful and allowing for revelations of fruitful insights into natural phenomena. However, these past decades during which statistical inference was heavily dominated by NHST also clearly revealed many of the shortcomings of this approach. Among these are fundamental weaknesses of the approach, promoting scientists to misunderstand what conclusions a significance test of a null hypothesis actually allows to draw. From a more practical perspective, the use of statistical procedures within the NHST framework frequently suffers from neglected assumptions and a lack of understanding of fundamental concepts and problems (e.g. pseudo-replication).

With regard to IT-based inference, the situation is different because the approach still has to prove its usefulness and applicability in animal behaviour research, though it seems likely that it will become an important tool. From a more practical perspective, I see a clear danger that the application of IT-based inference will suffer from the exact same problems as the NHST approach: At best semi-informed researchers will misunderstand the concepts behind it and draw conclusions that are not justified by the data and their analysis. To me, the major sources of such misunderstandings and potential misuses of the approach seem to be the following: (a) Confusion of multi-model inference with data dredging, i.e. investigating a set of models comprising several or all possible models that can be built out of a set of all possible variables that potentially might be somehow related to the phenomenon investigated and believing that the best model (or the models in the confidence set) necessarily represents something meaningful. From my understanding, such an approach can at best be considered as 'hypothesis generating' and is likely to reveal little insight (if any at all) into the phenomenon investigated; (b) neglecting assumptions (although IT-based inference suffers from data analysed using the wrong statistical models (e.g. error structure) as much as NHST-based inference does); (c) misunderstanding the conclusions that can be drawn from the IT-based approach, i.e. making dichotomous decisions based on a probabilistic approach which allows to explicitly incorporate uncertainty in model selection; (d) mixing the two approaches in the sense of selecting models based on an IT analysis and then testing their significance (or the significance of the variables they include) using NHST. However, given that researchers, journal editors and referees are aware of the pitfalls that an IT-based analysis provides and that it is not revealing a 'significance' in the classical sense, I believe that animal behaviour research can benefit a lot from incorporating IT-based analyses where appropriate.

Acknowledgements I wish to thank Peter Walsh and Hjalmar Kühl for introducing some of the concepts of IT-based inference to me and for several invaluable discussions about this approach. Kai F. Abt, Hjalmar Kühl, Kevin Langergraber, Rainer Stollhoff and three anonymous referees provided very helpful comments on an earlier version of this paper. Finally, I wish to thank László Zsolt Garamszegi for inviting me to write this paper and for his patience with me during the process of writing it. This work was supported by the Max Planck Society.

Conflict of interest The author declares to have no conflict of interest.

References

Aiken LS, West SG (1991) Multiple regression: testing and interpreting interactions. Sage, Newbury Park

American Psychological Association (1994) Publication manual of the American Psychological Association, 4th edn. APA, Washington

Anderson DR, Burnham KP, Thompson WL (2000) Null hypothesis testing: problems, prevalence, and an alternative. J Wildl Manage 64:912–923

Anderson DR, Link WA, Johnson DH, Burnham KP (2001) Suggestions for presenting the results of data analyses. J Wildl Manage 65:373–378

Austin PC, Tu JV (2004) Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. J Clin Epidemiol 57:1138–1146

Burnham KP, Anderson DR (2002) Model selection and multimodel inference, 2nd edn. Springer, Berlin

Burnham KP, Anderson DR, Huyvaert KP (2010) AICc model selection in Ecological and behavioral science: some background, observations, and comparisons. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1029-6

Chatfield C (1995) Model uncertainty, data mining and statistical inference. J Roy Stat Soc A Sta 158:419–466
Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates, New York
Cohen J (1994) The earth is round (p<.05). Am Psychol 49:997–1003
Cohen J, Cohen P (1983) Applied multiple regression/correlation analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates, Mahwah
Derksen S, Keselman HJ (1992) Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. Br J Math Stat Psychol 45:265–282
Dobson AJ (2002) An introduction to generalized linear models. Chapman & Hall, Boca Raton
Dochtermann NA, Jenkins SH (2010) Developing multiple hypotheses in behavioral ecology. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1039-4
Field A (2005) Discovering statistics using SPSS. Sage, London
Forstmeier W, Schielzeth H (2010) Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1038-5
Freckleton RP (2010) Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1045-6
Freedman DA (1983) A note on screening regression equations. Am Stat 37:152–155
Garamszegi LZ (2010) Information-theoretic approaches to statistical analysis in behavioural ecology: an introduction. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1028-7
Garamszegi LZ, Calhim S, Dochtermann N, Hegyi G, Hurd PL, Jørgensen C, Kutsukake N, Lajeunesse MJ, Pollard KA, Schielzeth H, Symonds MRE, Nakagawa S (2009) Changing philosophies and tools for statistical inferences in behavioral ecology. Behav Ecol 20:1363–1375
Guthery FS, Brennan LA, Peterson MJ, Lusk JJ (2005) Information theory in wildlife science: critique and viewpoint. J Wildl Manage 69:457–465
Harrell FE Jr (2001) Regression modeling strategies. Springer, New York
Hector A, von Felten S, Schmid B (2010) Analysis of variance with unbalanced data: an update for ecology & evolution. J Anim Ecol 79:308–316
Hegyi G, Garamszegi LZ (2010) Using information theory as a substitute for stepwise regression in ecology and behavior. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1036-7
Hurlbert SH, Lombardi CM (2009) Final collapse of the Neyman–Pearson decision theoretic framework and rise of the neo-Fisherian. Ann Zool Fenn 46:311–349
Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2:696–701
James FC, McCulloch CE (1990) Multivariate analysis in ecology and systematics: panacea or Pandora's box? Annu Rev Ecol Evol Syst 21:129–166
Johnson DH (1999) The insignificance of statistical significance testing. J Wildl Manage 63:763–772
Johnson DH (2002) The role of hypothesis testing in wildlife science. J Wildl Manage 66:272–276
Johnson BJ, Omland KS (2004) Model selection in ecology and evolution. Trends Ecol Evol 19:101–108
Lovell MC (1983) Data mining. Rev Econ Stat 65:1–12
Lukacs PM, Thompson WL, Kendall WL, Gould WR, Doherty PF Jr, Burnham KP, Anderson DR (2007) Concerns regarding a call for pluralism of information theory and hypothesis testing. J Appl Ecol 44:456–460
McCullagh P, Nelder JA (2008) Generalized linear models. Chapman and Hall, London
Møller AP, Jennions MD (2002) How much variance can be explained by ecologists and evolutionary biologists? Oecologia 132:492–500
Mundry R, Nunn CL (2009) Stepwise model fitting and statistical inference: turning noise into signal pollution. Am Nat 173:119–123
Nakagawa S, Cuthill IC (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev 82:591–605
Nickerson RS (2000) Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Methods 5:241–301
Quinn GP, Keough MJ (2002) Experimental designs and data analysis for biologists. Cambridge University Press, Cambridge
R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
Richards SA, Whittingham MJ, Stephens PA (2010) Model selection and model averaging in behavioural ecology: the utility of the IT-AIC framework. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1035-8
Royall R (1997) Statistical evidence, a likelihood paradigm. Chapman & Hall, London
Sakamoto Y, Akaike H (1978) Analysis of cross classified data by AIC. Ann Inst Stat Math 30:185–197
Schielzeth H (2010) Simple means to improve the interpretability of regression coefficients. Methods Ecol Evol 1:103–113
Siegel S, Castellan NJ (1988) Nonparametric statistics for the behavioral sciences, 2nd edn. McGraw-Hill, New York
Sleep DJH, Drever MC, Nudds TD (2007) Statistical versus biological hypothesis testing: response to Steidl. J Wildl Manage 71:2120–2121
Smith GD, Ebrahim S (2002) Data dredging, bias, or confounding. Brit Med J 325:1437–1438
Sokal RR, Rohlf FJ (1995) Biometry—the principles and practice of statistics in biological research, 3rd edn. Freeman, New York
Steidl RJ (2006) Model selection, hypothesis testing, and risks of condemning analytical tools. J Wildl Manage 70:1497–1498
Stephens PA, Buskirk SW, Hayward GD, del Rio CM (2005) Information theory and hypothesis testing: a call for pluralism. J Appl Ecol 42:4–12
Stephens PA, Buskirk SW, del Rio CM (2007) Inference in ecology and evolution. Trends Ecol Evol 22:192–197
Stoehr AM (1999) Are significance thresholds appropriate for the study of animal behaviour? Anim Behav 57:F22–F25
Symonds MRE, Moussalli A (2010) A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's Information Criterion. Behav Ecol Sociobiol. doi:10.1007/s00265-010-1037-6
Tabachnick BG, Fidell LS (2001) Using multivariate statistics, 4th edn. Allyn & Bacon, Boston
Vul E, Harris C, Winkielman P, Pashler H (2010) Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect Psychol Sci 4:274–290
Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP (2006) Why do we still use stepwise modelling in ecology and behaviour? J Anim Ecol 75:1182–1189
Young SS, Bang H, Oktay K (2009) Cereal-induced gender selection? Most likely a multiple testing false positive. Proc R Soc Lond, Ser B 276:1211–1212
Zar JH (1999) Biostatistical analysis, 4th edn. Prentice Hall, New Jersey