Behavioral Ecology
doi:10.1093/beheco/arp137
Advance Access publication 2 November 2009

Review

Changing philosophies and tools for statistical inferences in behavioral ecology

László Zsolt Garamszegi,[a] Sara Calhim,[b] Ned Dochtermann,[c] Gergely Hegyi,[d] Peter L. Hurd,[e] Christian Jørgensen,[f] Nobuyuki Kutsukake,[g] Marc J. Lajeunesse,[h] Kimberly A. Pollard,[i] Holger Schielzeth,[j] Matthew R.E. Symonds,[k] and Shinichi Nakagawa[l]

[a]Department of Evolutionary Ecology, Estación Biológica de Doñana-CSIC, c/Americo Vespucio, s/n, 41092, Seville, Spain, [b]Department of Biology, Queen's University, Kingston, ON, K7L 3N6, Canada, [c]Program in Ecology, Evolution and Conservation Biology, Department of Biology, University of Nevada, Reno, NV 89557, USA, [d]Department of Systematic Zoology and Ecology, Eötvös Loránd University, Pázmány Péter sétány 1/C, H-1117, Budapest, Hungary, [e]Department of Psychology, University of Alberta, Edmonton, Alberta T6G 2E9, Canada, [f]Department of Biology, University of Bergen, PO Box 7803, N-5020, Bergen, Norway, [g]Department of Evolutionary Studies of Biosystems, The Graduate University for Advanced Studies, Hayama, Miura-gun, Zushi, Kanagawa 240-0193, Japan, [h]National Evolutionary Synthesis Center, 2024 W. Main Street, Suite A200, Durham, NC 27705-4667, USA, [i]Department of Ecology and Evolutionary Biology, University of California, Life Sciences Building, 621 Charles E. Young Drive South, Los Angeles, CA 90095-1606, USA, [j]Max Planck Institute for Ornithology, Eberhard-Gwinner-Str. 5, 82319 Seewiesen, Germany, [k]Department of Zoology, University of Melbourne, Victoria 3010, Australia, and [l]Department of Zoology, University of Otago, 340 Great King Street, PO Box 56, Dunedin, New Zealand

Recent developments in ecological statistics have reached behavioral ecology, and an increasing number of studies now apply analytical tools that incorporate alternatives to the conventional null hypothesis testing based on significance levels. However, these approaches continue to receive mixed support in our field. Because our statistical choices can influence research design and the interpretation of data, there is a compelling case for reaching consensus on statistical philosophy and practice. Here, we provide a brief overview of the recently proposed approaches and open an online forum for future discussion (https://bestat.ecoinformatics.org/). From the perspective of practicing behavioral ecologists relying on either correlative or experimental data, we review the most relevant features of information theoretic approaches, Bayesian inference, and effect size statistics. We also discuss concerns about data quality, missing data, and repeatability. We emphasize the necessity of moving away from a heavy reliance on statistical significance while focusing attention on biological relevance and effect sizes, with the recognition that uncertainty is an inherent feature of biological data. Furthermore, we point to the importance of integrating previous knowledge in the current analysis, for which novel approaches offer a variety of tools. We note, however, that the drawbacks and benefits of these approaches have yet to be carefully examined in association with behavioral data. Therefore, we encourage a philosophical change in the interpretation of statistical outcomes, whereas we still retain a pluralistic perspective for making objective statistical choices given the uncertainties around different approaches in behavioral ecology. We provide recommendations on how these concepts could be made apparent in the presentation of statistical outputs in scientific papers. *Key words:* BeStat, Bonferroni correction, frequentist approach, information theoretic approach, measurement error, model selection, *P* value, prior, statistical power. *[Behav Ecol 20:1363–1375 (2009)]*

Behavioral ecologists rely on statistical analyses to make inferences from observational or experimental data. However, statistics is a dynamically developing discipline, and there is little consensus on how to choose from the available statistical methods or how to present results. For the last several decades, null hypothesis testing (NHT) based on statistical significance levels (*P* values) has dominated data analysis in the study of behavior (or other variables in the focus of behavioral ecologists). Recently, analytical tools that incorporate alternative statistical philosophies have been highlighted (e.g., Burnham and Anderson 2002; Gill 2002; Rushton et al. 2004; Clark and Gelfand 2006; Nakagawa and Cuthill 2007), but their use is still limited in behavioral ecology, mainly because of their unfamiliarity and of the prevalence of NHT-biased statistical training (Stephens, Buskirk, and del Rio 2007). Hereafter, we use the term "new" or "recent" statistical methods to mean new or recent to behavioral ecology.

The strength of behavioral studies in the field or laboratory is that they allow trait manipulation, which is a powerful way to reveal causal relationships (Quinn and Keough 2002). Nearly

Address correspondence to L.Z. Garamszegi. E-mail: laszlo.garamszegi@ebd.csic.es.

Received 5 December 2008; revised 29 July 2009; accepted 29 August 2009.

© The Author 2009. Published by Oxford University Press on behalf of the International Society for Behavioral Ecology. All rights reserved.
For permissions, please e-mail: journals.permissions@oxfordjournals.org

half of the papers published in "Behavioral Ecology" employ experimental approaches, and the vast majority of them use NHT-based statistics to make inferences from the data (Stephens, Buskirk, Hayward, and del Rio 2007). Controlled experiments are designed to reveal the causal relationship between 2 variables and to reduce the number of confounding factors experimentally by reducing the variance in confounding factors or by balanced group assignment. In carefully designed experiments, balanced or randomized group assignment allows the data to be analyzed with straightforward statistical tests like analysis of variance (ANOVA), *t*-test, regression or mixed-models (Ruxton 2006; Stephens, Buskirk, and del Rio 2007). However, although NHT is appropriate for many, carefully designed experiments, "new" statistical tools are not inappropriate for analyzing experimental data, and, in fact, it is misleading to believe that experimental approach necessarily calls for NHT.

The outcomes of ANOVAs and *t*-tests can be interpreted by using "novel" methods, which provide flexibly interpretable results irrespective to whether the data are collected experimentally or observationally (Lukacs et al. 2007). For example, Information Theoretic (IT) and Bayesian inferences allow complex modeling of multiple hypotheses and to incorporate prior knowledge into the analysis of experimental data (see below). Moreover, experimental data like any results in behavioral ecology could be quantified by effect sizes (see below). It follows that the advantages of the outlined methods, over NHT, seem more relevant to interpretation, rather than the experimental design per se. Accordingly, subsequent discussions and examples equally correspond to the analysis of both observational and experimental data.

In general, the statistical tools one adopts have consequences for the biological inferences that can be made from statistical analyses. NHT involves "binary" thinking (i.e., an effect was either demonstrated or not) and frames research hypotheses in the context of falsification. Although NHT does not preclude the estimation of parameters that reflect the strength of a biological effect, the strong attention paid to *P* values and arbitrary interpretations from NHT results shift the focus from biological significance to statistical significance (Stephens, Buskirk, and del Rio 2007). In contrast, recently proposed alternatives inherently treat biological effects or evidence on a continuous scale and focus on modeling of data instead of hypothesis testing. They allow the simultaneous assessment of multiple competing biological hypotheses that are translated to statistical models (for simplicity, we assume that one statistical model or hypothesis corresponds to one biological hypothesis, and hereafter we refer to "hypotheses" in a general sense covering both levels; however, see some relevant discussion on http://bestat.ecoinformatics.org). These alternative methods typically deal with the magnitude of effects and their biological relevance.

The statistical framework used for analysis also influences the manner in which researchers report results. Traditionally, the use of NHT has put high emphasis on the presentation of significance thresholds, whereas effect sizes and their precision were of secondary concern. Furthermore, the historical reporting of *P* values has implications for future syntheses of information using meta-analyses because they provide little information on the magnitude and direction of a research outcome. These problems are compounded because NHT use can also lead to publication bias (i.e., significant results are published preferentially) because many nonsignificant results are likely to remain in "file-drawers" (Rosenthal 1979; Møller and Jennions 2001). However, recently NHT has been more often combined with the presentation of effect sizes. We welcome this trend because it partly solves some of the issues by introducing a continuous perspective to biological evidence.

Consequently, the shift that we are experiencing in ecological statistics is expected to have a strong influence on how we conduct studies of behavior, as the statistical choice applied can influence research design and the interpretation of data. Hence, there is a compelling case for behavioral ecologists to become familiar with developments in the field of statistics. To this end, Garamszegi and Nakagawa organized a post-conference symposium for the 12th International Behavioral Ecology Congress, held at Cornell University in 2008, to cover a topical review of various methods. Our goal was to initiate a discussion about the most recent advancements in ecological statistics that were emerging in behavioral ecological studies but were not widely appreciated. These topics included 1) IT approaches that allow the comparative evaluation of multiple biological hypotheses; 2) Bayesian inference, in which new empirical evidence is combined with past knowledge to update or newly infer the probability of the hypotheses being tested; 3) issues about effect sizes and confidence intervals (CIs), which differentiate between the strength of biological effects and the precision by which these effects could be estimated; and 4) concerns about data quality and repeatability that undermine the biological relevance of the statistical results.

This paper synthesizes much of the discussions from this symposium by providing illustrative examples to demonstrate what these approaches offer in association to behavioral data (both experimental and correlational). Throughout this review, we avoid suggesting general support for one method over another. Rather, in accordance with recommended practices in ecology (Stephens et al. 2005), we emphasize the importance of a pluralistic approach, by which the researcher must carefully choose the most appropriate statistical method based on the questions at hand. Along this line, we consider the NHT approach as a mathematically correct tool that, if interpreted correctly, can still be used for testing certain questions in behavioral ecology. In the following sections, we discuss the advantages and disadvantages of recent statistical approaches.

MULTIMODEL INFERENCE, MODEL SELECTION, AND INFORMATION THEORY

Behavioral ecologists often deal with complex systems and seek to understand how interacting genetic and environmental factors have shaped behavioral phenotypes. Such complex systems are composed of a large number of potential relationships between different factors, and a scientific study aims at identifying predictors in the most appropriate combination that are responsible for a certain biological phenomenon. This task requires statistical approaches that can handle data with multiple predictors and can be used to evaluate different biological hypotheses in the form of statistical models, which summarize the predicted relationship between the response and predictor variables.

Multi-predictor problems in ecological and behavioral research have typically been treated by model simplification approaches using threshold-based removal-reintroduction algorithms (thresholds usually being $P < 0.05$–0.10), that is, stepwise selection (Miller 1992). The purpose of such a model simplification procedure is to follow an iteration process based on significance level to reach a single, parsimonious model, which contains few variables only but has a strong descriptive value (Ginzburg and Jensen 2004). The stepwise method, however, has been criticized for multiple reasons, including the use of arbitrary significance thresholds, biased distribution of the resulting parameter estimates, incongruence of different selection algorithms, redundancy of repeated parameter testing, and the poor coverage of possible model space and potentially incorrect reliance on a single final model (Anderson et al.

Figure 1
Analyses of data with multiple predictors using NHT and IT approaches. In a comparative analysis of Australian birds in the superfamily Meliphagoidea, Symonds and Johnson (2006) sought to identify what factors best-predicted patterns of mean abundance (the density at which individuals of each species exist). They tested the prediction that species that exploited a greater number of niches would tend to exist at higher abundances than species that were highly specialized. They split niche breadth into 2 components, diet breadth (the number of different food types eaten) and habitat breadth (a diversity index of how many habitats the species was found in). They also considered the effect of body mass and latitudinal position based on the findings of previously published studies of bird abundance.

They initially employed bivariate Pearson correlations in a phylogenetic generalized least squares framework and showed that body mass, latitude, diet breadth, and abundance, all appeared to be interlinked and correlated in some way. Habitat breadth did not show any significant associations with any other variable and was disregarded. Therefore they chose body mass, latitude, and diet breadth to be predictors of abundance in a general linear model:

Variable	Slope	SE	t	P
Body mass	−3.771	1.124	−3.24	0.001
Latitude	0.013	0.004	3.41	0.001
Diet breadth	0.023	0.027	0.84	0.404

The model $R^2 = 0.111$, $F_{3,114} = 5.894$, $P = 0.001$.

These results showed that body mass and latitude were the only significant predictors of mean abundance in these birds. Stepwise regression also produces a model that contains only these 2 parameters. Symonds and Johnson (2006) therefore concluded that neither aspect of niche breadth had any role in predicting patterns of local abundance. However, if they had considered an IT approach using AIC, their analysis would have led to a more cautious conclusion.

Starting with the a priori knowledge that both body mass and latitude are important predictors of abundance in these birds (based on previous studies, not the above analysis), and should be included, we eliminate comparisons of all possible combinations of variables (an all-subset approach) and focus on the question of how models including aspects of niche breadth compare to the model with body mass and latitude alone. There is no a priori reason to consider that interaction or polynomial terms might be important so they are not included.

Figure 1 continued

Model	AIC	ΔAIC[a]	w[b]	ER[c]
Body mass + latitude	−178.44	0	0.33	
Body mass + latitude + habitat breadth	−178.23	0.21	0.30	1.1
Body mass + latitude + diet breadth	−177.65	0.79	0.22	1.5
Body mass + latitude + habitat breadth + diet breadth	−176.88	1.56	0.15	2.1

[a] The difference in AIC between the first-ranked model and the given model.

[b] Akaike weight, that is, the weight of evidence that a given model is the best approximating model.

[c] Evidence ratio, model weight of the first-ranked model relative to that of the given model.

Although the model that includes only body mass and latitude is still the most likely, we cannot say with certainty that it is the only model that could apply to the data. With an Akaike weight of only 0.33 (i.e., it could be considered to be 33% probable that it is the best model), it is only just a better model than the ones that also contain habitat breadth and diet breadth (it is only 1.1 times more likely than model including habitat breadth). We cannot confidently rule out an effect of either aspect of niche breadth. Individual variable weights obtained by model averaging also support a more nuanced interpretation:

Variable	Averaged slope	95% CI	summed w[a]
Body mass	−2.483	−0.760 to −4.206	1.00
Latitude	0.010	0.002 to 0.018	1.00
Habitat breadth	0.113	−0.052 to 0.278	0.45
Diet breadth	0.030	−0.027 to 0.087	0.37

[a] The summed Akaike weight for the variable.

With parameter weights of 0.45 and 0.37, we can interpret aspects of niche breadth as having around 40% probability that they may indeed play a role in determining patterns on abundance in these birds. Notice also that the averaged slope estimates are different than those obtained in the single model that Symonds and Johnson (2006) produced. Although model details were not crucial to that particular analysis, in other cases (e.g., estimation of allometric slopes), AIC analyses and model averaging can provide different estimates of scaling exponents than consideration of single models.

2000; Whittingham et al. 2006). Accordingly, many authors warn against the use of stepwise selection and P value approaches for model selection and subsequent parameter estimation (Whittingham et al. 2006; Lukacs et al. 2007; Mundry and Nunn 2009). In spite of this, stepwise methods have been and are still being commonly used in our field, and we suspect that, because of their easiness, they will continue serving for some comparisons and exploratory analyses in the future. We expect this to occur when the interest is to determine which set of variables provides the best explanation for the variation in the data or when the aim is to contrast results with previous studies that used the same stepwise approach. Note that a recent comparison of the predictive ability of seven model selection approaches revealed that stepwise-based variable selection performed similarly to other algorithms when applied to 12 ecological datasets (Murtaugh 2009). However, we advocate that the shortcomings of stepwise methods should be carefully examined when interpreting results.

The recently introduced IT model comparison method allows the concurrent assessment of several, competing biological hypotheses that are defined a priori (Burnham and Anderson 2002; Johnson and Omland 2004). Information criteria (such as Akaike's information criterion [AIC]) quantify the relative fit of each candidate hypothesis (represented as a statistical model) based on the balance between the likelihood of the data given the model and parsimony in the number of parameters (Burnham and Anderson 2002). An entire suite of models reflecting different biologically relevant hypotheses can be ranked based on their relative criterion values without the need for a threshold of significance. Although probably the most widely used criterion in ecology is AIC (a reason why we also focus on it for our demonstrative purposes), it is just one criterion in the IT framework. Examples of other criteria include Bayesian information criterion (BIC) and deviance information criterion (DIC) (Congdon 2006; Claeskens and Hjort 2008; Ward 2008). The result of an IT process is the list of considered models that are ranked according to the information criterion used. A formal strength of evidence for each model can be acquired by calculating model likelihoods or AIC weights. Such information can be used for effective biological interpretations because the comparison of these metrics in the form of evidence ratio permits evidentiary statements about the plausibility of different hypotheses, given the data (see examples in Figures 1 and 2). Therefore, IT approaches offer more reliable

Figure 2
Interpreting experimental results by using the IT approach. A field study on the collared flycatcher (*Ficedula albicollis*) tested the predictions of parasite-mediated sexual selection theory by following the effect of an experimental immune challenge on song production (Garamszegi et al. 2004). An NHT-based approach demonstrated that sheep red blood cell injected males significantly decreased their song rate after the manipulation of health status, whereas control birds receiving physiological water (placebo) did not change their song output. By contrast, other song traits, such as song duration, were not affected significantly by the treatment. A reanalysis of the same data using an IT approach provided similar conclusions, but it gave explicit statements about the plausibility of different hypotheses. AIC scores for the 2 hypotheses (here H_0: no treatment effect and H_1: treatment effect) in association with song rate showed that more support could be obtained for the alternative hypothesis than to the null hypothesis. This corresponds to an evidence ratio of 9.3, which might be judged as "given the available data, a difference in the change in song rate between the 2 experimental groups is approximately 9.3 times more likely than no difference having occurred." This suggests limited (perhaps moderate) evidence for a treatment effect on song rate (see Lukacs et al. 2007), contrary to the result from NHT, which emphasizes that the evidence is significant. There is more uncertainty about the relative support of different hypotheses in relation to song duration. The NHT approach fails to reject the H_0 hypothesis, and the researcher is tempted to conclude that there is no treatment effect. However, the IT-AIC approach, given the available data, more or less equally supports both hypotheses ($\Delta AIC < 2$, evidence ratio $<< 10$). Such uncertainties are also evident when applying interpretations based on the effect sizes theorem (see Cohen's d).

Song rate					**Song duration**			
NHST-approach	t	df	P	R-squared	t	df	P	R-squared
	2.601	25	0.015	0.213	-0.507	25	0.617	0.010
Effect size approach	Cohen's d 1.007		95% CI (upper/lower) 0.298/2.562		Cohen's d 0.196		95% CI (upper/lower) -0.453/0.748	
IT-AIC approach	*AIC*	*ΔAIC*	*w*[c]	ER[d]	*AIC*	*ΔAIC*	*w*[c]	ER[d]
$H_0 (Y = \beta_0)$[a]	25.31	4.47	0.097	9.346	-15.15	0	0.704	0.421
$H_1 (Y = \beta_0, + \beta_1 * X)$[a,b]	20.84	0	0.903	(H_1 is ~9 times more likely than H_0)	-13.42	1.73	0.296	(H_0 is ~2.5 times more likely than H_1)

[a] Y: expressed song trait; β_0: overall mean of the trait (intercept).
[b] X: treatment assignment (SRBC vs. placebo); β_1: treatment effect (slope).
[c] Akaike weight.
[b] Evidence ratio.

model selection (but not necessarily model simplification) based on the simultaneous evaluation of multiple hypotheses.

The IT methods can also be used for parameter estimation. In fact, the strength of the IT approaches is that they allow model averaging, a technique that provides parameter estimates that incorporate model uncertainty and are based on multiple statistical models (Johnson and Omland 2004; Richards 2005; Claeskens and Hjort 2008). Model averaging, therefore, shifts the focus from the probability of models to the independent effect of each explanatory variable summed across supported models (see example in Figure 1). Frequently, many alternative models are all approximately equally likely (i.e., have similar AIC values). In every single model (as in a single "best" model), estimates and standard errors (SEs) are conditional on the model being correct. However, if we are unsure about the model structure, point estimates and SEs should incorporate this source of uncertainty. To do so, parameter estimates from alternative candidate models are weighted by the evidence for the respective models (e.g., measured as AIC weights) and are averaged across all candidate models (or a subset of best models, Burnham and Anderson 2002). It is also possible to identify the parameters that are more strongly represented across all well-supported models and are thus more likely to have important predictive value. This is done by calculating the cumulative evidence for the models containing a particular parameter (Burnham and Anderson 2002). Parameters can then be ranked according to their representation in models with good fit to the data (Burnham and Anderson 2002).

In addition to model parameters and model fits, another way of dealing with the relative importance of different predictors is to use an estimate of the explained variation (Burnham and Anderson 2002). The overall variation explained by any model relative to a random/null expectation can be calculated based on the log likelihoods of these models containing different combinations of predictors. By the careful consideration of the models being compared, the explained variance approach provides a powerful tool for assessing the contribution of predictors of interest. Such a statistic is a useful accessory to model fit statistics because although a model may be ranked as best in a given set it may explain only a small proportion of variance in the focal variable (Eberhardt 2003).

The initial model set should closely reflect the biological and theoretical background as well as research design (see example in Figure 1). Although the subsequent ranking of all models based on criterion values may lend support to more than one nonexclusive hypothesis, model selection results will be conditional on the initial model set considered. However, the decision which 2models to include in the initial model set is controversial and subject to philosophical issues (see Anderson 2008). Selecting from a large number of possible parameter combinations is sometimes cognitively intractable. The construction of a plausible, multiparameter candidate model set is left to the judgment of the model builder. If no prior information is available, decisions about which initial models to include may be challenging and require exploratory data analysis or the simplification of models containing a large set of potentially important terms.

In addition to the difficulties associated with the definition of the initial model set, there are other issues concerning the IT-based method that warrant attention and future test. For

example, it may be questionable to assume that the application of the statistical concept of parsimony based on model complexity and fit, which is at the heart of IT-AIC methods, has any biological relevance (Guthery et al. 2005). Moreover, there is no generally accepted benchmark for ranking competing models, as there are different information criteria (other than AIC) available for model comparison, which all have different consequences for model selection and subsequent parameter estimation (Claeskens and Hjort 2008). Finally, AIC has been suggested to be prone to overfitting, which results in that the most supported models are too complex, and often include variables and interactions, with very small effects (Pan 1999; Forster 2000; Seghouane 2006).

The scope of this paper only allows coverage of the most important philosophical aspects of the IT approaches based on AIC and to provide some examples (Figures 1 and 2). For those who intend to implement AIC-based and other selection methods into their research practice, we suggest Anderson (2008) as an introduction to the topic and Burnham and Anderson (2002) and Claeskens and Hjort (2008) as more advanced readings. In addition, an upcoming special issue in "Behavioral Ecology and Sociobiology" will deal with particular problems that researchers in our field may meet when analyzing behavioral data in an IT framework (Garamszegi 2010).

BAYESIAN APPROACH

Biologists including behavioral ecologists have traditionally had strong loyalty to the falsificationist approach as proposed by Karl Popper (1963), in which evidence is used to challenge scientific theory until it can be rejected. In other words, data are used to examine a null hypothesis against a single alternative hypothesis. This tradition relies on the binary nature of questions that researchers can ask in simple experiments by testing hypotheses about single parameter causation (i.e., means are different between the control and experimental groups). Although the NHT approach offers an appropriate, and mathematically correct, statistical framework for controlled experiments (Stephens et al. 2005; Whittingham et al. 2006), the resulting *P* values can only be used to make inferences about the validity of the null hypothesis, whereas the degree to which the data support alternative hypotheses remains unexplored (Lukacs et al. 2007). The philosophy of the falsification approach (i.e., the rejection or acceptance of a single hypothesis) thus ignores the uncertainty about the best explanation that can be given for an observed phenomenon (Stephens, Buskirk, and del Rio 2007). Evidence against the null hypothesis may be mistaken as evidence for a specific alternative hypothesis. Bayesian thinking, on the other hand, recognizes that data rarely provide full support for a single hypothesis, but they should only affect the extent to which we interpret which hypothesis is more likely. Note that this philosophical approach also applies to the IT methods. However, only Bayesian methods can make "true" probabilistic statements on any hypothesis (note that Akaike weights used in IT approaches are often treated as representing model probabilities, i.e., the probabilities of hypotheses, but Akaike weights are only the approximations of such probabilities under a large-sample size condition, by assuming each model has equal probabilities prior to data collection; Burnham and Anderson 2002; McCarthy 2007).

Bayesian statistics has recently gained popularity in many areas including phylogeny construction and complex ecological modeling (Ronquist 2004; Clark 2005; Clark and Gelfand 2006; McCarthy 2007). Behavioral ecologists seem to be among the last to employ this flexible framework in their routine analysis (Stephens, Buskirk, and del Rio 2007). This may be because researchers are unfamiliar with this approach or they think that it is inappropriate to analyze experimental data. Bayesian statistics does have concepts and properties which some researchers in the field may not be familiar with (e.g., prior and posterior probabilities; see below), but this does not necessarily mean that the statistical background of behavioral ecologists is useless in the face of Bayesian statistics. Moreover, it is misleading to assume that Bayesian methods preclude experimentation, as the underlying philosophy is concerned with how data are analyzed and interpreted but not with how they are collected. Here, we outline the key properties of Bayesian statistics that differentiate it from the well-known NHT tools. Readers are recommended to consult with accessible introductory books (Gelman and Hill 2007; McCarthy 2007) on Bayesian statistics or more thorough reviews (Gelman et al. 1995; Gill 2002; Congdon 2003, 2005, 2006).

The philosophy of Bayesian statistics is fundamentally different from the one we follow in traditional statistics (e.g., based on NHT). In traditional statistics, we generally rely on the frequentist viewpoint of probability, which is the expected frequency of occurrence of an event in a large number of trials given a particular statistical null hypothesis. According to the Bayesian definition of probability, it is the plausibility of an event given the evidence of the event. To assess the probability of a hypothesis, Bayesians specify a certain prior probability, which reflects our current belief in it and is then updated in the light of the new data. More precisely, Bayesian statistics allows us to obtain the probability of hypotheses (or parameters of interest) given observed data, as described by Bayes' theorem:

$$\Pr(\theta|D) = \frac{\Pr(\theta) \times \Pr(D|\theta)}{\Pr(D)},$$

where θ is a parameter to be estimated (e.g., a intercept or slope; or θ can be read as a hypothesis) and D represents data; $\Pr(\theta)$ (probability of parameter or hypothesis) is what is often referred to as the prior probability or prior, $\Pr(\theta|D)$ (probability of parameter or hypothesis given the data) is the posterior probability or posterior, $\Pr(D|\theta)$ (probability of the data given a parameter or hypothesis) is referred to as the likelihood function (in the NHT context, the *P* value is the probability of data given the null hypothesis being true), and $\Pr(D)$ is the probability of observing data (it acts as a normalizing constant). Therefore, the posterior distribution is proportional to the combination of the prior distribution and the likelihood function. Accordingly, the goal of Bayesian statistics is the estimation of posterior distribution with a given prior and likelihood function. Therefore, the Bayesian approach focuses on the probability of hypotheses, given the data, whereas NHT is concerned with the probability of data, given a null hypothesis.

With Bayesian inference, evidence or observations are used to update or to newly infer the probability that a hypothesis (a parameter) may be true through Markov chain Monte Carlo (MCMC) iterations. A detailed description of MCMC methods is beyond the scope of this paper (see suggested references on Bayesian statistics and for more in-depth treatment of this topic, Gamerman and Lopes 2006). In brief, a Markov chain is a sequence of events where an event at time point t is only influenced by an event at $t - 1$, whereas the Monte Carlo process is a simulation using random number generators (i.e., random sampling). As a result of these 2 processes working together, MCMC methods provide a posterior distribution of each parameter from which we can easily obtain means, SEs, and 95% credible intervals (a Bayesian version of CIs). The samples from the posterior distribution of a parameter are parameter values that are likely given the data, and the

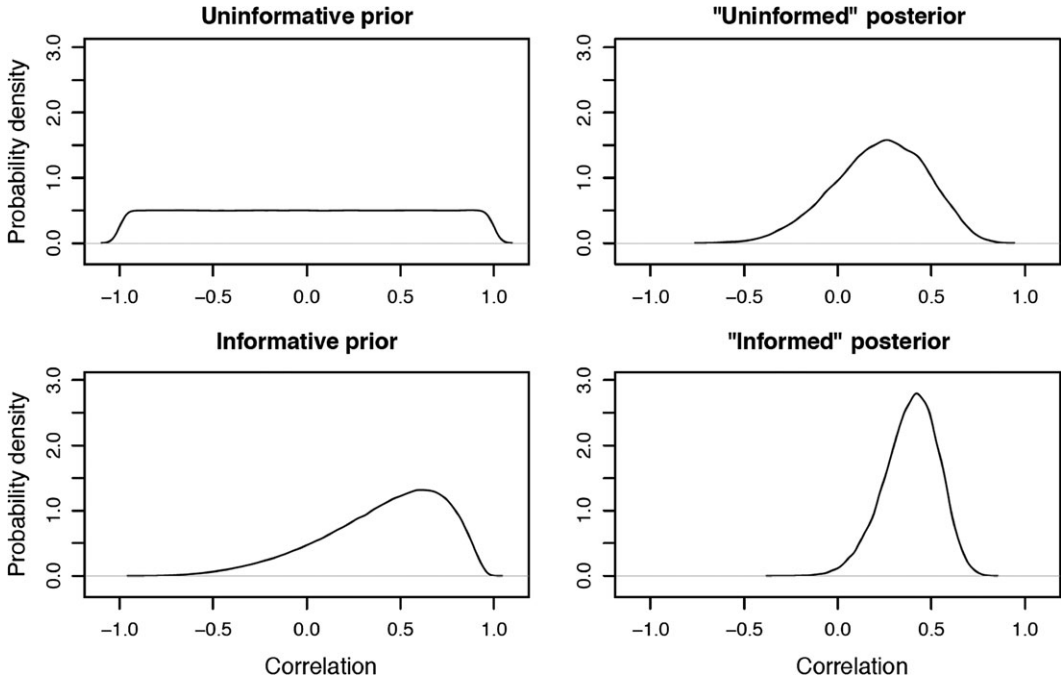


Figure 3
Bayesian versus NHT approaches: Incorporating previous results via a prior. Nakagawa et al. (2007) conducted a series of meta-analyses to assess the function of bib size in male house sparrows (*Passer domesticus*). The results showed that there was a strong correlation between male bib size and status (competitive ability) after integrating 15 effect size estimates from 12 different populations ($r = 0.463$; 95% CI = 0.290–0.608). This constitutes a good example for a "badge of status" in the house sparrow. Now imagine that a researcher tested this relationship in an island population of house sparrows. She observed aggressive encounters among 15 male sparrows and measured their bib sizes. The statistical result was nonsignificant according to NHT ($r = 0.266$, $t_{13} = 0.991$, $P = 0.339$). She might be tempted to conclude that the badge of status hypothesis does not hold in this island population. This, however, is the fallacy of interpreting lack of evidence against the null hypothesis as evidence for it. Although this might seem obvious in this particular case, this fallacy is an inherent issue in NHT. The conclusion could have been different if she had been a Bayesian. The correlation between bib size and status after incorporating the above meta-analytic result supports the badge of status hypothesis (posterior mode $r = 0.422$, 95% credible interval = 0.102–0.657). Note that this estimate does not simply reflect the prior information but that incorporates the limited amount of new data from the island, which results in relatively large credibility intervals. Notably, if she uses an uninformative prior, the Bayesian estimation would have been similar to the classical correlation (posterior mode $r = 0.262$, 95% credible interval = −0.269 to 0.669; see figure). She could report both results with the informative and uninformative priors and discuss the implications of both results. Such a formal integration of prior knowledge will provide more careful assessments of the current result in relation to the previous research. Hence, the Bayesian method is essentially meta-analytic, that is, reflecting new data in the light of prior information, and naturally shifts the focus to effect size rather than significance thresholds. The figure shows the visual presentations of probability distributions of the priors and posteriors.

density distribution of the samples shows values of the parameter that are more likely than others. The Bayesian 95% credible interval contains the true value of the parameter with a probability equal to 0.95, given the model, the data, and the prior (for accounts of subtle and clear differences between confidence and credible intervals see Hilborn and Mangel 1997; McCarthy 2007). Bayesian results, therefore, can be easily interpreted in a manner most behavioral ecologists are already used to.

To obtain a posterior distribution from a Markov chain, "priors" are required. Priors are used to establish initial probability distributions for the parameters in the model, and they set out the parameter space where the Markov chain is allowed to explore. Therefore, in a Bayesian framework, the evaluation of hypotheses is fundamentally linked to preceding information and assumptions (see example in Figure 3). By contrast, the assimilation of previous information into NHT approaches is subjective, as we focus on post hoc explanations of unexpected results that contradict our predictions and previously available information. The careful choice of priors in Bayesian statistics by incorporating knowledge from previous findings can increase the precision at a lower sample size (note that an analogous issue is termed as "statistical" power in a NHT framework). For example, previous studies suggest treating males with testosterone reduces paternal care in many bird species (Wingfield et al. 1987), and we may reasonably predict a prior probability distribution for reduction in care in a bird species subjected to testosterone treatment. Then, we may be able to reduce the number of birds involved in a study where effects of testosterone on paternal care and associated questions are investigated. This is because the prior increases the precision of the posterior estimate (or reduces its SE) provided that data more or less support the prior evidence and that the variance associated with the prior is small compared with variance associated with data. Therefore, the appropriate use of priors increases scientific efficiency and also has welfare implications.

However, choosing and finding appropriate priors is probably the most contentious issue among Bayesian statisticians (Gelman et al. 1995; Gill 2002). If we choose incorrect priors for parameters of interest, such choice will lead to biased parameter estimates, incorrect SEs, and thus possibly incorrect conclusions, especially when the sample size is small. Moreover, we may have difficulty in defining prior distributions for certain parameters because, say, we work on a species, which has never been studied or we investigate totally new aspects of behavior using a new technique. In such cases, values referred to as uninformative priors can be used; these

priors have "flat" probability distributions with equal probability assigned to a large range of parameter values (or hypotheses). When uninformative priors are used, estimates from frequentist and Bayesian methods are usually similar. However, Bayesian statistics may often be the only solution for problems that cannot be traced in a classical framework (Gill 2002; McCarthy 2007) because of their flexibility in model building. Figure 3 demonstrates how prior information can be efficiently incorporated into a Bayesian framework.

Bayesian statistics outperforms frequentist methods in several respects. In a frequentist approach, parameter estimates are usually to a large extent restricted by the assumed probability distributions (hidden in the assumptions of statistical models) and can thus be unreliable (Gelman et al. 1995; Gill 2002; Gelman and Hill 2007). For example, in the classical framework, SEs for variances are usually approximated assuming normal distributions of these variances (given the fact that variances cannot go below zero, variances are usually not normally distributed). The Bayesian approach, on the other hand, is less restricted by certain probability distributions. Furthermore, parameter estimates can easily incorporate various sources of uncertainties. For example, posterior and prior distributions depict stochastic variations, by which variations in trait values caused by measurement errors or within-individual fluctuations are captured (van Dongen 2001). Prior distributions can deal with uncertainties around biological assumptions and predictions, for which previous knowledge can be taken into account. Posterior distributions allow statements about the probability of a hypothesis or that a parameter falls within a particular range (Gill 2002; Congdon 2003). Moreover, in a comparative study of trait variation across species, the application of the Bayesian approach can treat uncertainties about estimating phylogenetic relationships (Pagel et al. 2004; Pagel and Meade 2006). In this context, posterior distributions are obtained from evolutionary models fitted to millions of statistically supported phylogenetic trees. Consequently, with Bayesian methods more reliable statistical modeling is possible for various and complex biological problems, even when nonnormal data and small sample sizes are used (Carlin and Louis 2000). Furthermore, model selection can be conducted using criterion-based approaches described in the previous section although criteria different from AIC such as BIC and DIC or Bayes Factors (BF) are more often used (Congdon 2003, 2005, 2006). Finally, Bayesian approaches can be employed to effectively deal with missing data and zero-inflated distributions (see below). In fact, the estimation of unobserved data is an inherent feature of the Bayesian technique, as it is the by-product of MCMC modeling, by which the posterior distributions of the parameters are obtained.

CIs ALONG WITH EFFECT SIZES

Statistically significant results may be demonstrated even when the effects are negligibly small biologically, given sufficiently large sample sizes. This problem is relatively rare in behavioral ecology, where samples tend to be relatively small and any statistically significant result is almost certain to represent a meaningful effect. Nevertheless, it is important not to confuse the P value with the magnitude of effects of interest, as P is sensitive to sample size (Nakagawa and Cuthill 2007). The importance of an effect and the precision of its estimate are statistically characterized by effect size and the associated CIs (Cohen 1994; Rosenthal 1994; Grafen and Hails 2002). Effect sizes describe biological patterns along a continuum by using a common currency metric (often in units of standard deviations [SDs]), which makes results easily interpretable and comparable across studies. Effect sizes are estimated from samples, and the robustness of the estimates is manifested in their CIs. In this framework, small effects (that would be nonsignificant by using NHT) can be informative. If they are surrounded by narrow CIs, then the researchers can have high confidence that the true biological effect is weak. In contrast, a result with a very large CI on the effect size (even if it is significant in NHT), which can be of intermediate or large magnitude, cannot reasonably be used to conclude that the effect is large. This is because broad CIs may also imply that the true effect is actually weak, but based on the available sample it can be estimated with considerable uncertainty (see further differences between interpretations based on NHT and effect size theorem in Figures 2 and 4). Hence CIs should always accompany effect sizes. Accordingly, reporting effect sizes and their CIs has become a recommended practice in biology, although this recommendation has not permeated general statistical practices in behavioral ecology (Nakagawa 2004; Garamszegi 2006).

There are 2 broad categories in what is referred to as effect size: unstandardized effect sizes (e.g., regression coefficients and mean differences) and standardized effect sizes (e.g., correlation coefficients, Cohen's d and Hedges' d) (Cohen 1988; Rosenthal 1994; Nakagawa and Cuthill 2007). Standardized effect sizes or dimensionless effect statistics are particularly useful because these statistics are comparable across studies and are easy to calculate even with a calculator or spreadsheet (for an extended discussion on problems associated with effect size calculations see Nakagawa and Cuthill 2007). However, the correct calculation of CIs around standardized effect sizes may appear a challenging task for behavioral ecologists because this necessitates the use of noncentral t and F distributions, which are far from being practically used in our field and in conventional statistical packages. The approximate width of 95% CIs for an effect size can be derived from the asymptotic SE for the effect size, which is the most commonly used estimation method in practice. Bootstrap resampling provides a simple and robust method for calculating statistics that do not have simple sampling distributions (Efron and Tibshirani 1993). Given random sampling and sufficiently large sample size, the distribution of effect sizes in bootstrap samples simulates the probability distribution of effect sizes in the population. One can therefore calculate the effect size for thousands of bootstrap samples drawn from the original empirical sample and then obtain the 95% CI by determining the range of frequency distribution which includes 95% of those effect sizes (Manly 1991; Kelley 2005). Routine presentation of (standardized) effect sizes and their CIs will encourage researchers to view their results in the context of previous research because this statistic is independent of the scale on which variables were measured and the statistical design they came from (Thompson 2002; Figure 4).

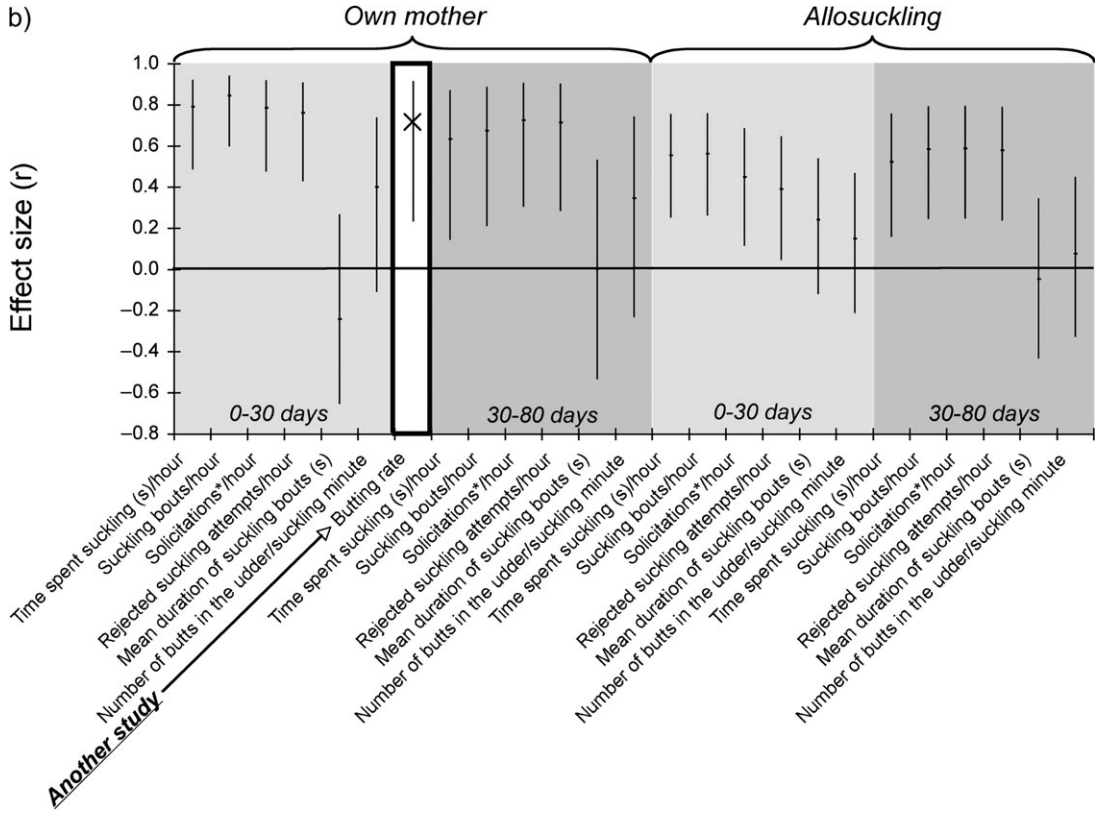
Reporting effect sizes with SEs (or CIs) will also facilitate the incorporation of results into future meta-analyses (Lajeunesse and Forbes 2003), which has become the standard method for synthesizing published research in biology (Arnquist and Wooster 1995). However, an emerging challenge for ecological meta-analysis in biology is how to generalize research and pool effect sizes when research is replicated across a diversity of taxa. The problem is that research outcomes of closely related taxa may not represent independent pieces of information (Felsenstein 1985). This may threaten the validity of quantitative reviews because the statistical assumption of independence in meta-analysis is violated (Hedges and Olkin 1985). Meta-analysis also assumes homogeneity of variances among effect sizes, but effect size data with a phylogenetic structure can violate this assumption because taxa may have evolved at different rates (Harvey and Pagel 1991). Recent statistical developments that account for phylogenetic

Figure 4
Interpreting experimental results by using the effect size theorem. Therrien et al. (2008) investigated the trade-off between maintenance and reproduction in the white-tailed deer (*Odocoileus virginianus*). In this experimental study, the researchers compared nursing behavior of fawns between a food-restricted and a control group. Based on the significance of the treatment effect in the statistical analyses, the study concluded that 4 of 6 variables (total time spent suckling, suckling frequency, number of rejected attempts, and number of solicitations) were affected by the treatment, whereas the remaining 2 variables (mean duration of suckling bouts and the number of butts per suckling bout) did not differ between experimental groups (a). This pattern was consistent between periods and suckling conditions.

a) (modified from Therrien en al. 2008)

	Own mother			Allosuckling		
	F	*df*	*P*	*F*	*df*	*P*
0–30 days						
Time spent suckling (s)/h	22.73	1,14	<0.001	13.08	1,30	0.001
Suckling bouts/h	33.61	1,14	<0.001	13.53	1,30	<0.001
Solicitations/h	21.78	1,14	<0.001	7.41	1,30	0.01
Rejected suckling attempts/h	18.68	1,14	<0.001	5.26	1,30	0.03
Mean duration of suckling bouts (s)	0.88	1,14	0.37	1.81	1,30	0.33
Number of butts in the udder/suckling minute	2.61	1,14	0.13	0.65	1,30	0.43
30–80 days						
Time spent suckling (s)/h	7.21	1,11	0.02	8.42	1,23	0.01
Suckling bouts/h	8.95	1,11	0.01	11.7	1,23	<0.01
Solicitations/h	11.91	1,11	0.01	11.87	1,23	<0.01
Rejected suckling attempts/h	11.19	1,11	0.01	11.34	1,23	<0.01
Mean duration of suckling bouts (s)	0	1,11	0.98	0.06	1,23	0.82
Number of butts in the udder/suckling minute	1.47	1,11	0.26	0.12	1,23	0.74

However, interpretations based on effect sizes and 95% CI would allow more careful conclusions (b). Effect sizes and their CIs for the former 4 behavioral traits fall within the range of medium to strong effects sensu Cohen (1988). However, these effects tend to be systematically weaker in the first period (0–30 days of lactation) in the case of allosuckling. Moreover, CIs cover quite broad ranges when data are limited (e.g., see the cases when fawns are nursed by their own mother), which raises uncertainty around the estimates of weaker effects. Accordingly, it would be premature to conclude that the number of butts per suckling bout was not affected by the treatment in every circumstance. The study indeed fails to show a very strong effect for this relationship, but based on the range of CIs, it is equally likely that a future study will find a zero or strong effect. With this respect, the authors contrasted their results with that of another study (Haley et al. 1998) that showed a significant relationship between butting rate and milk flow in the domestic cow (*Bos taurus*). However, when the comparison is based on effect sizes, the difference does not seem very robust, as the 95% CIs largely overlap (b).



The relevance of effects sizes along a continuous scale is obvious in the first period of allosuckling, in which effect sizes decline with the order of variables. The difference between effect sizes for rejected suckling attempts and for solicitation is similar to the difference between effect sizes for rejected suckling attempts and for mean duration of suckling bout. A NHT-based interpretation would qualitatively distinguish these similar differences because it would disregard the former difference (as both effect sizes correspond to a significant association), whereas it would only contrast the nonsignificant treatment effect for the duration of suckling bout with the significant patterns found for the other 2 traits.

nonindependence have improved the estimation of pooled effect sizes and CIs without bias (Verdú and Traveset 2005; Adams 2008; Lajeunesse 2009). These statistics intergrate phylogenetic information into all the traditional meta-analytical tools, such as using fixed- and random-effects models for pooling effect sizes and calculating CIs and testing for homogeneity of variances (Lajeunesse 2009). In addition, these statistics emphasize a generalized least squares approach that uses AIC scores to fit different evolutionary hypotheses (e.g., Brownian motion or Ornstein–Uhlenbeck process) to the meta-analytic data.

DATA QUALITY AND REPEATABILITY

Statistical analyses rely on available data. Therefore, the choices of statistical approaches and the interpretation of results should consider not only analytical issues but also data quality. If the statistical analysis does not account for the requirements imposed by the nature of the data (e.g., assumptions concerning the distribution of data), the statistical outcome will be biased or false. If the analyses use unreliable data that do not reflect the biological phenomenon of interest (e.g., the use of absolute brain size without correcting for body size to reflect cognitive abilities, see Martin and Harvey 1985), the results of the study will provide misleading conclusions even if they seem statistically meaningful.

When studying animal behavior, we generally rely on the strong assumption that a behavior, or its predictor, is an individual-specific attribute. The statistical consequence of this assumption is that mean individual values of traits are used in the analyses. However, within-individual, or more broadly, within-subject variation cannot be inherently neglected, as it can have biological meaning on one hand and can also invalidate statistical results that are based on mean values on the other hand.

Statistically, within-subject variation can be described by repeatability (e.g., between observers, between measurements, between data sources, within individuals or species), which influences the replicability of the main findings and the extent to which we can trust subject-specific mean values (Lessells and Boag 1987). Repeatability approximates the amount to which between-subject (e.g., between-individual or between-species) variation relates to total variation. High repeatability means that individuals always produce more or less the same measured value (Hayes and Jenkins 1997), whereas low repeatability indicates that the trait displays considerable within-subject variation or that our measurement is prone to sampling errors (see below). When using NHT, the significance of repeatability will be the probability value associated with the subject factor in the ANOVA table, although the repeatability value itself is a derived metric that requires further calculations (Lessells and Boag 1987). Additional formulas exist to calculate SEs and 95% CIs around repeatability (Becker 1984).

In behavioral ecology, within-subject variation may increase due to several biological and technical reasons, so calculating repeatability and balancing between number of subjects and the number of trials/measurements within subjects seem particularly important tasks. First, observed behaviors in animals are the results of extremely complex mechanisms, as they are influenced by several intrinsic (i.e., neural, endocrine, and genetic effects) and extrinsic (i.e., physical and social environments) factors (Danchin et al. 2008). Therefore, behavioral traits like features of song, cognitive performance traits, foraging or personality traits are usually displayed with great individual flexibility and variability even across consecutive observations, which result in lower repeatability than in the case of morphological traits (Garamszegi et al. 2006a). A recent meta-analysis relying on more than 700 repeatability estimates of behavioral traits revealed that the average repeatability across all estimates is below 0.4 (Bell et al. 2009). Hence, modest repeatability is an inherent and expected feature of many variables used in our field. It is intriguing that although many studies in behavioral ecology characterize the determinants of the mean expression of individual behaviors, few focus on within-individual variation. This suggests that the conceptual questions of interest have concerned the average behavior across individuals, rather than the variation within these individuals (i.e., behavioral characterization of populations rather than individuals). Second, traits can vary with time because they may be exhibited differently during different times of the day or during different parts of the breeding season, or depending on the individual's state, all of which increase within-individual variation. Similarly, behavioral variation can occur due to spatial heterogeneity. Third, in the interspecific context of comparative studies, within-species trends are of importance for shaping within-subject variation (Harvey and Pagel 1991). For example, differences between populations, sexes, age-classes, or individuals can all contribute to within-species variations and thus reduce repeatability. In fact, the problem caused by within-species variation in interspecific studies currently receives considerable attention in evolutionary biology (Harmon and Losos 2005; Ives et al. 2007; Felsenstein 2008). Simulations have demonstrated that when the data are structured by phylogenetic relationships, low repeatability can cause type I errors (i.e., spurious relationships; Harmon and Losos 2005). Finally, measurement errors are also important causes of unwanted variation and noise, and they are of little, if any, biological relevance. Measurement error may not only be caused by instrumental constraints, but we may also commit mistakes and simplifications when we make calculations from the raw measurements (Calhim and Birkhead 2007). Moreover, simulation shows that metrics that depict the extent to which within- and between-subject variations relate to each other are sensitive to perturbations in sampling design (Pollard KA, unpublished data). We suspect that it is usually a challenge to distinguish measurement error from biological factors which reduce repeatability although there have been developments in statistical methods dealing with measurements errors (Congdon 2006).

Low repeatability is thus likely to represent an important problem in behavioral ecology that deserves statistical treatment. This is actually a dual task. First, we need statistical approaches that are able to handle variation within the studied objects. For example, formulas exist to correct for the downward bias that low repeatability due to within-subject variation raises in the calculation of correlation coefficients and regression slopes (Fan 2003; McArdle 2003; Adolph and Hardin 2007; see also Figure 5). If both dependent and independent variables can be estimated at the case by case level within subjects (i.e., multiple measurements are available within individuals), mixed-effects models allow analyses with the raw data in which subject-specific effects can be followed through the corresponding main factor without the need of calculating means at the subject level (van de Pol and Wright 2009). Alternatively, one may calculate different measures of within-subject variation and include them in statistical models together with mean values (e.g., van de Pol and Verhulst 2006; Byers 2007; Dochtermann and Jenkins 2007). A further complication is that not only mean values have a repeatability but also slopes and this sometimes needs to be considered (Schielzeth and Forstmeier 2009). Moreover, for problems arising in a phylogenetic context, recently developed comparative methods allow for the incorporation of within-species variation into the evolutionary models (Ives et al. 2007; Felsenstein 2008). These can be used to study interspecific patterns of behaviors while simultaneously controlling for

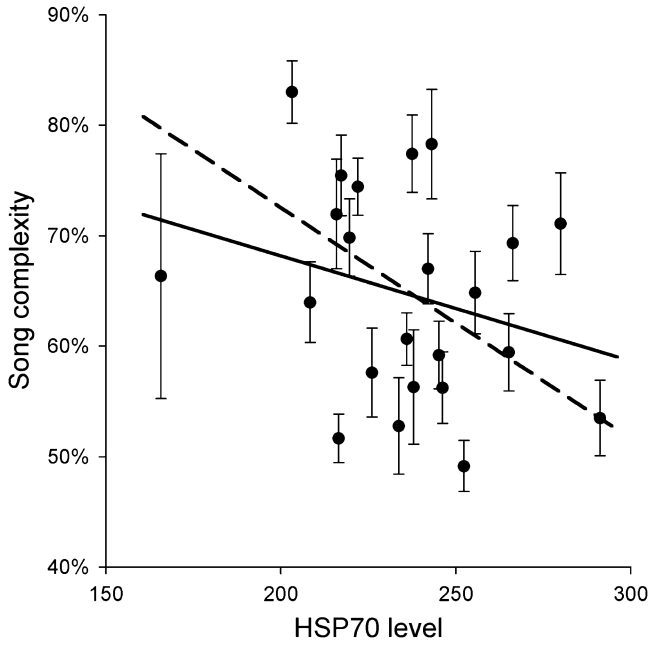


Figure 5
Adjusting for heterogeneity in data quality. Garamszegi et al. (2006b) studied the physiological consequences of the production of sexually selected traits using the collared flycatcher (*Ficedula albicollis*) as a model to analyze the relationship between song traits and levels of heat shock proteins (HSP 70) that mediate stress response. One focal variable was song complexity, which is estimated as the number of syllable types relative to the number of syllables within each song produced. A field record usually contains several songs for each individual, thus allowing multiple sampling within subjects. In the study, 20 songs were obtained for each male. Subsequent tests were performed by using the average of these measurements for each male (dots). A correlation relying on individuals as the unit of analysis did not reveal strong relationship between song complexity and levels of HSP 70 ($r = -0.271$, $N = 23$, 95% CI $= -0.615/0.159$, $P = 0.211$, solid line). However, song complexity has only a moderate repeatability ($R = 0.373$, $F_{22,437} = 8.142$, $P < 0.001$) indicating that the trait displays a substantial within-individual variation, which is neglected when using individual-specific means. Applying a correction formula on the correlation coefficient that takes repeatability into account (Adolph and Hardin 2007), the relationship appears stronger ($r = -0.399$, $N = 23$, 95% CI $= -0.697/0.016$, $P = 0.059$, when using $R = 0.5$ for the repeatability of HSP 70 based on Morales J, unpublished data on females). Supposing that available sample size varies across males (e.g., for some individuals less than ten songs could be recorded), the reliability of data units differs. This is because trait values can be estimated with larger errors if only few measurements are available (error bars). If one corrects for this heterogeneity in data quality by using weighted regression based on within-subject sample size, the outcome might be different. In this example, we arbitrarily created lower within-subject sample size for some males (by randomly deleting some raw measurements) that deviate from the suspected pattern and then downweighted their role in the regression analysis by using statistical weights. This resulted in a stronger relationship between the focal variables at the individual level ($r = -0.478$, $N = 23$, 95% CI $= -0.744/-0.082$, $P = 0.021$, dashed line). Note that in this example, within-individual sample sizes were artificially modified for illustrative purposes. It could equally occur that a correction for unbalanced sampling has an effect in the other direction, in which the weighted effect size is smaller than the effect sizes obtained based on the individual specific means.

different sources of biases, such as phylogeny and within-species variation.

Second, we must also deal with the fact that data quality can vary across observations. A common underlying assumption of most statistical approaches is that each data point provides equally precise information about the deterministic part of total process variation, that is, the error term is constant over all values of the predictor or explanatory variables (Sokal and Rohlf 1995). If repeatability is modest, mean estimates derived from a few observations will be less reliable than mean estimates from larger within-subject samples. Therefore, if within-subject sample size differs across subjects, it may be expected that heterogeneity in data quality will be introduced in the data set that uses mean values. The standard solution to violations of assumptions by heterogeneous data quality is to weight each observation by sample size or another measure of sampling effort (Draper and Smith 1981; Neter et al. 1996; see Figure 5) or to use mixed models. Interestingly, this is analogous and also statistically similar to what meta-analysis does (i.e., weighting each effect size with corresponding sample size from which effect size is calculated).

Another common violation of statistical assumptions is heterogeneous data quality caused by the nonstandard data distributions, such as a large proportion of zero values included in the data. For example, parasite prevalence is a case in point because many animals have no parasites at all, whereas others may have heavy parasite load (Jovani and Tella 2006). Data transformations in these cases do not work for normalization, so it is required to employ special statistical methods to analyze zero-inflated data (Martin et al. 2005). Since the first report on the analytic methods of the zero-inflated data (Lambert 1992), several methods have been made available (Martin et al. 2005). One of these, the 2-component model (also called "hurdle" model) conducts 2 separate analyses—the first analysis investigates factors predicting whether the dependent term is zero or nonzero and the second one investigates factors predicting nonzero values of the dependent term using a truncated Poisson distribution (see Cockburn et al. 2008). A different solution is to model the zero-inflated data by using the discrete mixture model (Lambert 1992; Welsh et al. 1996), constructing a single distribution out of 2 different discrete distributions—a mixture of the Bernoulli distribution and a Poisson distribution (see Charpentier et al. 2008 as an example; Lambert 1992; Welsh et al. 1996). As discussed in a previous section, Bayesian methods are particularly suitable for implementing complex models such as the discrete mixture model and may more effectively process variation in non-Gaussian data without relying on transformations.

The problem of missing data, another aspect of data quality, is a neglected topic in the field of ecology and evolution (Hadfield 2008; Nakagawa and Freckleton 2008), although missing data issues are probably present in most data sets behavioral ecologists deal with. We usually delete missing observations and work with complete cases. However, such complete case analyses often lead to reduced statistical power. In behavioral ecology, data sets are often small and further reduction of power may be the last thing behavioral ecologists need. Additionally, case-wise deletion can also cause biases in parameter estimates, especially if missing data occur nonrandomly as they are "missing" due to biological reasons. For example, older animals or one specific sex may be harder to catch, but even behavioral types can cause trap-shy or trap-happy effects and thus biased sampling (Biro and Dingemanse 2009; Garamszegi et al. 2009). Furthermore, missing data threaten the validity of model selection, whichever selection methods are used. Fortunately, there have been recent statistical advances in handling missing data to alleviate these problems. Techniques such as multiple imputations and data augmentation have become well accepted in the statistical literature (Allison 2002; Little and Rubin 2002). Moreover, the Bayesian framework also offers

approaches to treat missing data. We refer readers to a recent article on missing data and associated statistical techniques (see Nakagawa and Freckleton 2008; and references therein).

TAKE-HOME MESSAGE AND SUGGESTIONS FOR FUTURE DIRECTIONS

We have discussed some major statistical issues that practicing behavioral ecologists may frequently encounter. These issues share at least 4 important features, along which our field is likely to develop. First, all of these incorporate the common philosophy that moving away from heavy reliance on statistical significance is necessary, whereas more attention should be paid to biological relevance with the appreciation that uncertainty is an inherent feature of biological data. These uncertainties can be handled through model averaging, posterior distributions, CIs, and repeatability. Second, "new" approaches offer variable tools to amalgamate previous findings or knowledge with the current analysis. Theoretical or observational evidence can drive decisions about the initial model set in an IT framework or the prior distributions in Bayesian statistics, whereas effect sizes stimulate meta-analytic design to summarize related research findings. Third, each method still involves methodological challenges and brings up new problems to be solved. Particularly, the drawbacks and benefits of available techniques are not fully understood in the context of the evolutionary study of behavior, and we currently lack a consistent statistical philosophy in behavioral ecology. Fourth, none of the approaches should be overwhelmingly and generally supported over each other or even traditional approaches, as they provide means to treat particular but not all statistical problems. Some features have not been widely explored in association with data sets typical for behavioral ecology, thus some care is needed when applying novel methods in our field. Additionally, more than one approach may often seem applicable to a given analytical design, and the researcher is left with the task of selecting among the available methods.

Therefore, the methods added recently to the analytical arsenal of behavioral ecologists bring new and pluralistic statistical concepts into our research focus. As the statistics chosen can have strong implications for the biological conclusions, we would like to stimulate the community to prepare for changes in ecological statistics by statistical training and careful implementation and encourage researchers to test the applicability of "new" methods in the specific designs we adopt in our field. To enhance this process, we recommend some potentially useful routes along which behavioral ecologists can improve the integration of novel statistical concepts into their research and reports. Readers may agree or disagree with these suggestions, but at the least we advocate that researchers in our field prepare for the changes we experience and expect in ecological statistics.

In general, research practice should ideally echo the key statistical concepts emerging in ecological statistics. To be able to make an objective statistical choice, we need to understand the pros and cons of different methodologies. Instead of blindly following a fashion, it is our responsibility to carefully evaluate the available analytical approaches (both old and new) while taking into account the question, the assumptions and the data at hand. To achieve this, the task of researchers is that they perpetually train themselves.

To make the analyses transparent, there are ways by which we can "significantly" improve how we report data. Most importantly, data presentation and interpretation can be made more objective if they reflect the uncertainty with which estimations of biological effects are associated. This may involve the presentation of all statistical results, instead of the preferential report of those beyond some threshold, which can help the readers' interpretation. For example, 1) researchers using IT may want to report the full list of evaluated and ranked models with the associated AIC scores (or parameter weights); 2) summary statistics for the posterior distribution (e.g., mean, SD, credible interval) from a Bayesian Markov chain can be reported (together with figures showing chain convergence), from which the exploration of parameter space becomes clear; 3) for any presented biological effect (even in the IT and Bayesian frameworks), both the standardized effect sizes and the corresponding CIs can be given; and 4) the estimation of within-group repeatability of some parameters is valuable and maybe of interest, especially if data structure constrains us to eliminate within-subject variance by calculating mean values. Furthermore, we can enhance data presentation to help comparisons with previous and future findings (e.g., in a meta-analysis). Electronic appendices are now widely available for publishing excessive data or result details.

We could also show progress in clarifying our statistical decisions by providing clear reasoning behind each statistical choice followed and by the careful examination of the underlying assumptions. If multiple approaches seem equally applicable, their outcomes can be reported together (again electronic appendix material can be used), and thus, the robustness of the results can be assessed. This may also apply in cases when different settings are similarly plausible (i.e., different information criteria, candidate model sets, and prior settings). Electronic appendices can also be used for the inclusion of raw data sets, which is not conventional in the field of ecology and evolution but is so in other fields (American Psychological Association 2001). Given the diversity of statistical methods, making raw data available seems reasonable and provides a potential solution to misinterpretations. Also, such data depositions in public enrich our field and help it progress, encouraging the testing of alternative explanations and discouraging scientific fraud. This process seems obvious in relation to phylogenetic comparative studies, in which the corresponding inter-specific data sets are becoming generally accessible. Another use of electronic appendices can be that authors of theoretical papers can illustrate how their new theories can be incorporated into real statistical models, which rarely seems to happen.

To help statistical integration at different levels, we have created BeStat (http://bestat.ecoinformatics.org) to encourage the dissemination of statistical development among behavioral ecologists. The aim of this web-project is to synchronize statistical discussion and training based on a user-built information source. This platform offers several functions which can potentially help the transfer of knowledge, but its content is left to be assembled by the entire community. We have opened spaces for any kind of online discussion (Stat-Chat), for the standard broadcasting of any statistical issue via lay summaries and references (Stat-Sum), for the building of electronic tutorials and examples (Stat-Wiki), and for hosting statistical programs, resources, and links (Stat-Prog).

We emphasize that statistics are only tools to aid our interpretations of data. As behavioral ecologists, we should remember to integrate knowledge of the biology and ecology of the study species into statistical practices at all analytical stages.

FUNDING

Postdoctoral fellowship from the Research Foundation, Flanders (Fonds Wetenschappelijk Onderzoek, Vlaanderen, Belgium) (to L.Z.G.); "Ramon y Cajal" research grant from the Spanish National Research Council (Consejo Superior de

Investigaciones Científicas, Spain) (to L.Z.G.); Travel grant, University of Otago (to S.N.); Australian Research Council (to M.R.E.S.); Deutsche Forschungsgemeinschaft (FO 340/2 to H.S.).

We are very grateful to National Center for Ecological Analysis and Synthesis for hosting BeStat, with special thanks to J. Regetz and S. Walbridge for their practical help. M. Elgar and 2 anonymous reviewers provided constructive comments. We are indebted to S. Vehrencamp for her assistance in organizing the symposium.

REFERENCES

Adams DC. 2008. Phylogenetic meta-analysis. Evolution. 62:567–572.
Adolph SC, Hardin JS. 2007. Estimating phenotypic correlations: correcting for bias due to intraindividual variability. Funct Ecol. 21: 178–184.
Allison PD. 2002. Missing data. Thousand Oaks (CA): Sage.
American Psychological Association. 2001. Publication manual of the American Psychological Association. 5th ed. Washington (DC): American Psychological Association.
Anderson DR. 2008. Model based inference in the life sciences: a primer on evidence. New York: Springer.
Anderson DR, Burnham KP, Thompson WL. 2000. Null hypothesis testing: problems, prevalence, and an alternative. J Wildl Manage. 64:912–923.
Arnquist G, Wooster D. 1995. Meta-analysis: synthesizing research findings in ecology and evolution. Trends Ecol Evol. 10:236–240.
Becker WA. 1984. Manual of quantitative genetics. Pullman (WA): Academic Enterprises.
Bell AM, Hankison SJ, Laskowski KL. 2009. The repeatability of behavior: a meta-analysis. Anim Behav. 77:771–783.
Biro PA, Dingemanse NJ. 2009. Sampling bias resulting from animal personality. Trends Ecol Evol. 24:66–67.
Burnham KP, Anderson DR. 2002. Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer-Verlag.
Byers BE. 2007. Extra-pair paternity in chestnut-sided warblers is correlated with consistent vocal performance. Behav Ecol. 18: 130–136.
Calhim S, Birkhead TR. 2007. Testes size in birds: quality versus quantity—assumptions, errors, and estimates. Behav Ecol. 18:271–275.
Carlin BP, Louis TA. 2000. Bayes and empirical Bayes methods for data analysis. London: Chapman & Hall.
Charpentier MJE, Prugnolle F, Gimenez O, Widdig A. 2008. Genetic heterozygosity and sociality in a primate species. Behav Genet. 38:151–158.
Claeskens C, Hjort NL. 2008. Model selection and model averaging. Cambridge: Cambridge University Press.
Clark JS. 2005. Why environmental scientists are becoming Bayesians. Ecol Lett. 8:2–14.
Clark JS, Gelfand AE. 2006. A future for models and data in environmental science. Trends Ecol Evol. 21:375–380.
Cockburn A, Sims RA, Osmond HL, Green DJ, Double MC, Mulder RA. 2008. Can we measure the benefits of help in cooperatively breeding birds: the case of superb fairy-wrens *Malurus cyaneus*? J Anim Ecol. 77:430–438.
Cohen J. 1988. Statistical power analysis for the behavioural sciences. 2nd ed. Hillsdale (NJ): Lawrence Erlbaum Associates.
Cohen J. 1994. The earth is round ($p < .05$). Am Psychol. 49: 997–1003.
Congdon P. 2003. Applied Bayesian modelling. Chichester (UK): Wiley.
Congdon P. 2005. Bayesian models for categorical data. Chichester (UK): Wiley.
Congdon P. 2006. Bayesian statistical modelling. 2nd ed. Chichester (UK): Wiley.
Danchin É, Giraldeau L-A, Cézilly F. 2008. Behavioural ecology: an evolutionary perspective on behaviour. Oxford: Oxford University Press.
Dochtermann NA, Jenkins SH. 2007. Behavioural syndromes in Merriam's kangaroo rats (*Dipodomys merriami*): a test of competing hypotheses. Proc R Soc B Biol Sci. 274:2343–2349.
Draper NR, Smith H. 1981. Applied regression analysis. 2nd ed. New York: John Wiley.
Eberhardt LL. 2003. What should we do about hypothesis testing? J Wildlife Manage. 67:241–247.
Efron B, Tibshirani RJ. 1993. An introduction to the bootstrap. Monographs on statistics and applied probability 57. New York: Chapman & Hall.
Fan XT. 2003. Two approaches for correcting correlation attenuation caused by measurement error: implications for research practice. Educ Psychol Meas. 63:915–930.
Felsenstein J. 1985. Phylogenies and the comparative method. Am Nat. 125:1–15.
Felsenstein J. 2008. Comparative methods with sampling error and within-species variation: contrasts revisited and revised. Am Nat. 171:713–725.
Forster MR. 2000. Key concepts in model selection: performance and generalizability. J Math Psychol. 44:205–231.
Gamerman D, Lopes HF. 2006. Markov chain Monte Carlo: stochastic simulation for Bayesian inference. Boca Raton (FL): CRC.
Garamszegi LZ. 2006. Comparing effect sizes across variables: generalization without the need for Bonferroni correction. Behav Ecol. 17:682–687.
Garamszegi LZ. Forthcoming 2010. Information-theoretic approaches to statistical analysis in behavioural ecology: an introduction. Behav Ecol Sociobiol.
Garamszegi LZ, Eens M, Török J. 2009. Behavioural syndromes and trappability in free-living collared flycatchers, *Ficedula albicollis*. Anim Behav. 77:803–812.
Garamszegi LZ, Hegyi G, Heylen D, Ninni P, de Lope F, Eens M, Møller AP. 2006. The design of complex sexual traits in male barn swallows: associations between signal attributes. J Evol Biol. 19: 2052–2066.
Garamszegi LZ, Merino S, Török J, Eens M, Martínez J. 2006. Indicators of physiological stress and the elaboration of sexual traits in the collared flycatcher. Behav Ecol. 17:399–404.
Garamszegi LZ, Møller AP, Török J, Michl G, Péczely P, Richard M. 2004. Immune challenge mediates vocal communication in a passerine bird: an experiment. Behav Ecol. 15:148–157.
Gelman A, Carlin JB, Stern HS, Rubin DB. 1995. Bayesian data analysis. London: Chapman & Hall.
Gelman A, Hill J. 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge: Cambridge University Press.
Gill J. 2002. Bayesian methods: a social and behavioral sciences approach. Boca Raton (FL): CRC Press.
Ginzburg LR, Jensen CXJ. 2004. Rules of thumb for judging ecological theories. Trends Ecol Evol. 19:121–126.
Grafen A, Hails RS. 2002. Modern statistics for the life sciences. Oxford: Oxford University Press.
Guthery FS, Brennan LA, Peterson MJ, Lusk JJ. 2005. Information theory in wildlife science: critique and viewpoint. J Wildl Manage. 69:457–465.
Hadfield JD. 2008. Estimating evolutionary parameters when viability selection is operating. Proc R Soc B Biol Sci. 275:723–734.
Haley DB, Rushen J, Duncan IJH, Widowski TM, De Passillé AM. 1998. Butting by calves, *Bos taurus*, and rate of milk flow. Anim Behav. 56:1545–1551.
Harmon LJ, Losos JB. 2005. The effect of intraspecific sample size on type I and type II error rates in comparative studies. Evolution. 59:2705–2710.
Harvey PH, Pagel MD. 1991. The comparative method in evolutionary biology. Oxford: Oxford University Press.
Hayes JP, Jenkins SH. 1997. Individual variation in mammals. J Mammal. 78:274–293.
Hedges LV, Olkin I. 1985. Statistical methods for meta-analysis. London: Academic Press.
Hilborn R, Mangel M. 1997. The ecological detective: confronting models with data. Princeton (NJ): Princeton University Press.
Ives AR, Midford PE, Garland T. 2007. Within-species variation and measurement error in phylogenetic comparative methods. Syst Biol. 56:252–270.
Johnson JB, Omland KS. 2004. Model selection in ecology and evolution. Trends Ecol Evol. 19:101–108.
Jovani R, Tella JL. 2006. Parasite prevalence and sample size: misconceptions and solutions. Trends Parasitol. 22:214–218.
Popper KR. 1963. Conjectures and refutations. London: Routledge and Keagan Paul.

Kelley K. 2005. The effects of nonnormal distributions on confidence intervals around the standardized mean difference: bootstrap and parametric confidence intervals. Educ Psychol Meas. 65: 51–69.
Lajeunesse MJ. 2009. Meta-analysis and the comparative phylogenetic method. Am Nat. 174:369–381.
Lajeunesse MJ, Forbes MR. 2003. Variable reporting and quantitative reviews: a comparison of three meta-analytical techniques. Ecol Lett. 6:448–454.
Lambert D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 34:1–14.
Lessells CM, Boag PT. 1987. Unrepeatable repeatabilities: a common mistake. Auk. 104:116–121.
Little RJA, Rubin DB. 2002. Statistical analysis with missing data. 2nd ed. New York: Wiley.
Lukacs PM, Thompson WL, Kendall WL, Gould WR, Doherty PF, Burnham KP, Anderson DR. 2007. Concerns regarding a call for pluralism of information theory and hypothesis testing. J Appl Ecol. 44:456–460.
Manly BFJ. 1991. Randomization, bootstrap and Monte Carlo methods in biology. New York: Chapman & Hall.
Martin RD, Harvey PH. 1985. Brain size allometry: ontogeny and phylogeny. In: Jungers WL, editor. Size and scaling in primate biology. New York: Plenum Press. p. 147–173.
Martin TG, Wintle BA, Rhodes JR, Kuhnert PM, Field SA, Low-Choy SJ, Tyre AJ, Possingham HP. 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecol Lett. 8:1235–1246.
McArdle BH. 2003. Lines, models, and errors: regression in the field. Limnol Oceanogr. 48:1363–1366.
McCarthy MA. 2007. Bayesian methods for ecology. Cambridge: Cambridge University Press.
Miller AJ. 1992. Subset selection in regression. Boca Raton (FL): Chapman and Hall.
Møller AP, Jennions MD. 2001. Testing and adjusting for publication bias. Trends Ecol Evol. 16:580–586.
Mundry R, Nunn CL. 2009. Stepwise model fitting and statistical inference: turning noise into signal pollution. Am Nat. 173: 119–123.
Murtaugh PA. 2009. Performance of several variable-selection methods applied to real ecological data. Ecol Lett. 12:1061–1068.
Nakagawa S. 2004. A farewell to Bonferroni: the problems of low statistical power and publication bias. Behav Ecol. 15: 1044–1045.
Nakagawa S, Cuthill IC. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev. 82: 591–605.
Nakagawa S, Freckleton R. 2008. Missing inaction: the dangers of ignoring missing data. Trends Ecol Evol. 23:592–596.
Nakagawa S, Ockendon N, Gillespie DOS, Hatchwell BJ, Burke T. 2007. Assessing the function of house sparrows' bib size using a flexible meta-analysis method. Behav Ecol. 18:831–840.
Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. 1996. Applied linear statistical models. Chicago (IL): Irwin.
Pagel M, Meade A. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am Nat. 167:808–825.
Pagel M, Meade A, Barker D. 2004. Bayesian estimation of ancestral character states on phylogenies. Syst Biol. 53:673–684.
Pan W. 1999. Bootstrapping likelihood for model selection with small samples. J Comput Graph Stat. 8:687–698.
Quinn GP, Keough MJ. 2002. Experimental design and data analysis for biologists. Cambridge: Cambridge University Press.
Richards SA. 2005. Testing ecological theory using the information-theoretic approach: examples and cautionary results. Ecology. 86: 2805–2814.
Ronquist F. 2004. Bayesian inference of character evolution. Trends Ecol Evol. 19:475–481.
Rosenthal R. 1979. The "file drawer problem" and tolerance for null results. Psychol Bull. 86:638–641.
Rosenthal R. 1994. Parametric measures of effect size. In: Cooper H, Hedges LV, editors. The handbook of research synthesis. New York: Russell Sage Foundation. p. 231–244.
Rushton SP, Ormerod SJ, Kerby G. 2004. New paradigms for modelling species distributions? J Appl Ecol. 41:193–200.
Ruxton GD. 2006. The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. Behav Ecol. 17:688–690.
Schielzeth H, Forstmeier W. 2009. Conclusions beyond support: overconfident estimates in mixed models. Behav Ecol. 20:416–420.
Seghouane AK. 2006. A note on overfitting properties of KIC and KICc. Signal Processing. 86:3055–3060.
Sokal RR, Rohlf FJ. 1995. Biometry. 3rd ed. New York: W.H. Freeman & Co.
Stephens PA, Buskirk SW, del Rio CM. 2007. Inference in ecology and evolution. Trends Ecol Evol. 22:192–197.
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, Hayward GD, del Rio CM. 2007. A call for statistical pluralism answered. J Appl Ecol. 44:461–463.
Symonds MRE, Johnson CN. 2006. Determinants of local abundance in a major radiation of Australian passerines (Aves: Meliphagoidea). J Biogeogr. 33:794–802.
Therrien JF, Cote SD, Festa-Bianchet M, Ouellet JP. 2008. Maternal care in whitetailed deer: trade-off between maintenance and reproduction under food restriction. Anim Behav. 75:235–243.
Thompson B. 2002. What future quantitative social science research could look like: confidence intervals for effect sizes. Educ Res. 31: 25–32.
van de Pol M, Verhulst S. 2006. Age-dependent traits: a new statistical model to separate within- and between-individual effects. Am Nat. 167:766–773.
van de Pol MV, Wright J. 2009. A simple method for distinguishing within- versus between-subject effects using mixed models. Anim Behav. 77:753–758.
van Dongen S. 2001. Modelling developmental instability in relation to individual fitness: a fully Bayesian latent variable model approach. J Evol Biol. 14:552–563.
Verdú M, Traveset A. 2005. Early emergence enhances plant fitness: a phylogenetically controlled meta-analysis. Ecology. 86:1385–1394.
Ward EJ. 2008. A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. Ecol Model. 211:1–10.
Welsh AH, Cunningham RB, Donnelly CF, Lindenmayer DB. 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecol Model. 88:297–308.
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.
Wingfield JC, Ball GF, Duffy AMJ, Hegner RE, Ramenofsky M. 1987. Testosterone and agression in birds. Am Sci. 75:602–608.