Ecology, 95(3), 2014, pp. 637–642

In defense of *P* values: comment on the statistical methods actually used by ecologists

John Stanton-Geddes,[1,3] Cintia Gomes de Freitas,[2] and Cristian de Sales Dambros[1]

[1]*Department of Biology, University of Vermont, 109 Carrigan Drive, Burlington, Vermont 05405 USA*
[2]*Department of Plant Biology, University of Vermont, 63 Carrigan Drive, Burlington, Vermont 05405 USA*

Introduction

In recent years, a persuasive argument has been made for the use of information criterion (IC) model selection in place of null hypothesis significance testing (NHST) based on *P* values (Johnson 1999, Burnham and Anderson 2002, Johnson and Omland 2004). In this issue, Murtaugh (2014) questions the basis for this argument. We comment on this paper from the perspective of early-career ecologists and present the results of an informal survey of our colleagues on their choice of statistical methods. Specifically, we ask to what extent the IC approach has supplanted traditional hypothesis testing. Finally, we address issues related to the use and interpretation of *P* values, the Akaike information criterion (AIC), and effect sizes in ecological studies.

What Are *P* Values For?

Statistical models often are used in a NHST framework to find the factors "explaining" a certain pattern. Increasingly, statistical models also are used in an exploratory analysis or for data mining, in which many predictors are examined without a priori hypotheses, and the "significant" results are considered candidates for follow-up study (e.g., genome-wide association studies and climatic effects on species distribution). As practicing ecologists, we use *P* values or AIC to determine whether a specific factor (e.g., water quality) is an important predictor for an ecological outcome (e.g., fish abundance). *P* values lead to binary decision making (Fisher 1973 as cited in Murtaugh 2014). While this yes/no outcome may be desirable for management outcomes, it is exactly what IC approaches try to avoid. While the past 15 years have seen a strong push for the use of AIC in ecological studies to avoid this binary decision making, in practice, threshold values of change in AIC (ΔAIC) are often used in a similar way as are *P* values: to assess significance of a predictor. This practice is one of the arguments that Murtaugh (2014) uses to question the criticism of NHST.

Specifically, for nested linear models with Gaussian errors, Murtaugh (2014) demonstrates that *P* values, confidence intervals, and AIC are mathematically equivalent and therefore provide different approaches to reporting the same statistical information. While the equivalence of *P* values and confidence intervals is by definition true and should be no surprise to any student of statistics, the relationship between *P* values and AIC is not as intuitive. The proponents of AIC cited by Murtaugh and others (e.g., Whittingham et al. 2006) have made strong statements regarding null hypothesis testing that appear to be ill founded in light of Murtaugh's results. In particular, demonstrating that the choice of a threshold for ΔAIC is as arbitrary as a chosen significance (α) level for *P* values challenges the idea the ΔAIC is always the preferable method.

In practice, the choice of statistical method is constrained by experimental design. Specifically, as explained by Murtaugh (2014), null hypothesis testing is appropriate to test "the effects of treatments in a randomized experiment" whereas AIC is "useful in other situations involving the comparison of non-nested statistical models." Thus, for designed experiments with few parameters, there is no clear reason to use AIC over NHST, whereas in studies with many parameters and potential interactions, AIC is preferable. Moreover, AIC has the advantage that it can be used for non-nested models. Given that for many studies, using AIC as opposed to *P* values to select significant predictors is primarily a matter of choice, we were interested in the extent to which ecologists chose AIC or conventional NHST in the analysis of a simple data set.

What Methods Are Early-Career Ecologists Using?

To evaluate the extent to which the IC approach has supplanted the use of *P* values, we downloaded a typical *observational* data set from Ecological Archives (Koenig and Knops 2013) consisting of a single response (acorn count), three designed effects (species, site, and year) and 14 environmental variables, from which we selected a subset of 7 for simplicity. We recruited early-career

Manuscript received 18 June 2013; revised 16 September 2013; accepted 18 September 2013. Corresponding Editor: A. M. Ellison. For reprints of this Forum, see footnote 1, p. 609.

[3] E-mail: johnsg@uvm.edu

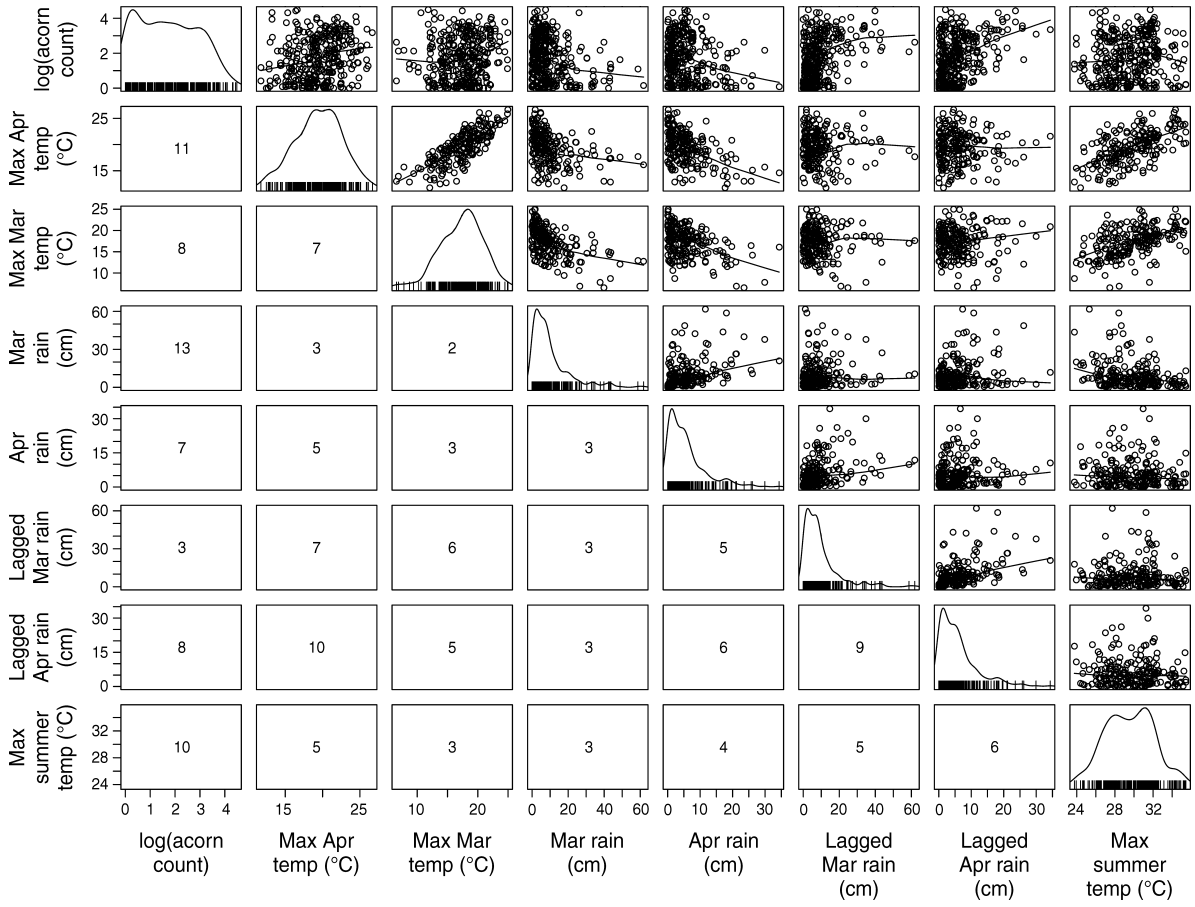

FIG. 1. Plot used for exploratory data analysis of the example data. The diagonal shows density plots representing the distribution of each variable. The upper right triangle of panels shows bivariate scatterplots of all variables, with a lowess smooth function. The lower left triangle of panels is a summary of the final models from the survey participants, with the number of models out of 20 that included each variable (rows) under the first column (acorn count) and the number of time each combination of variables occurred on the off-diagonals. The order of the rows and columns is acorn count (number per 30 seconds of counting, log-transformed); mean maximum (max) April temp; mean max March temp; March rain; April rain; March rain lagged 1 yr; April rain lagged 1 year; mean max summer temp. Full description of the variables is available in Koenig and Knops (2013).

ecologists, who will have had the most exposure to the AIC literature in their academic training, using personal e-mail, the ecolog list serve, and the early-career ecologists blog (*available online*).[4,5] We asked them to "explain the variation in the response variable (acorn count) using the predictors available" (full details in Supplement). We received responses from a skilled (average self-reported statistical expertise of 6.7 on scale of 1 [low] to 10 [high]) diverse group of 24 ecologists representing 7 countries. Of these responses, 10 participants used *P* values, 10 used AIC, and four used alternative (e.g., Bayesian) approaches. Thus, it appears that even among early-career ecologists, there is a lack of clear consensus of which method is more appropriate.

Starting with the same data set, participants came to surprisingly different conclusions. Of the participants who reported some type of model selection, no two final models included exactly the same set of predictors. Moreover, of the 10 potential predictor variables, not a single one was included in every final model (Fig. 1, lower left panels). While the final models differed in the number of predictors they contained, each term was retained in roughly the same proportion of models selected by *P* values or AIC. Moreover, most final models had similar predictive power and there was no qualitative improvement in prediction after four parameters were included in the model (Fig. 2) emphasizing the point that "Regression is for prediction and not explanation." We further explored how model selection influenced prediction by dividing the data into trial (70% of observations) and test (30% of observations) data sets. For each of the 20 final models provided by survey participants, we fit linear models on 400 trial data sets and calculated the squared error for each model as the deviation of the predicted values from the observed

[4] https://listserv.umd.edu/archives/ecolog-l.html
[5] https://earlycareerecologists.wordpress.com/

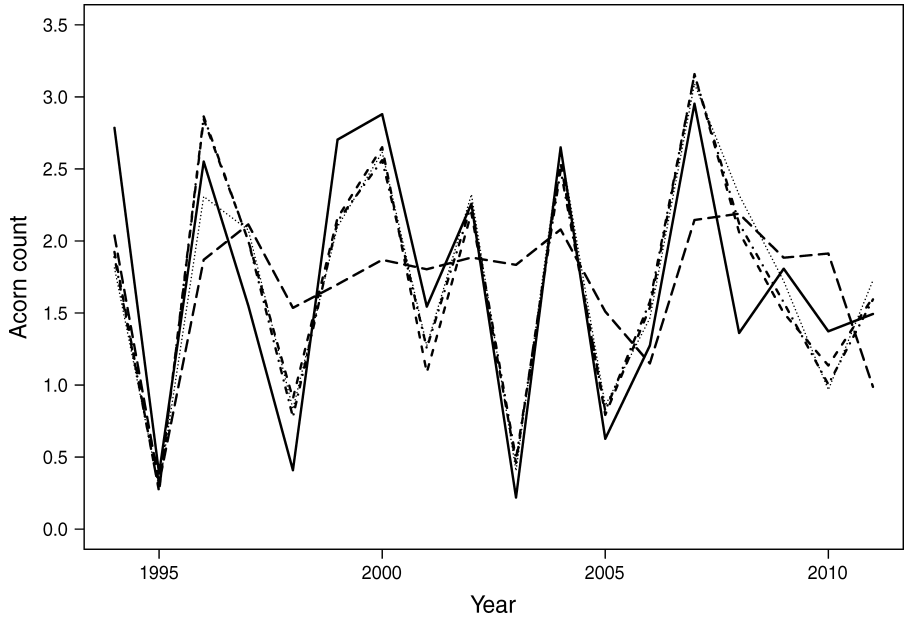

FIG. 2. Time series (solid line) of acorn count, averaged across species and sites. Predictions from models including a single parameter (long-dashed line), four parameters selected by null hypothesis significance testing (NHST; dotted line), five parameters selected by Akaike's information criterion (AIC; dot-dashed line), and all parameters (short-dashed line) are also shown.

values in the test data set (Fig. 3). Additionally, we created models by randomly picking one to all of the variables and testing their predictive ability (Fig. 3, gray-shaded region and black line). We found that model selection improves prediction when few parameters are included in the model, but with four or more parameters there is no difference between randomly selecting parameters and using model selection (Fig. 3; point estimates for selected models fall along the line for randomly selected models). Moreover, there was no clear difference in predictive ability among models selected by AIC (solid circles) or *P* values (open squares), though models selected by AIC tended to include fewer parameters.

This variation in results was surprising, given that all participants used the same general approach. The majority, 88%, of participants performed exploratory analysis, typically graphical exploration of correlations among the environmental variables (Fig. 1, upper right panels). In many cases, they often selected a single one of the highly correlated variables (e.g., $r = 0.78$ between mean maximum March and mean maximum April temperatures) to include in the models. Thus, the choice of which variables to include as predictors in the model is one explanation for why the final models differed among responses. Subsequently, participants fit a wide range of statistical models, with 96% using R (R Development Core Team 2013), as we encouraged in our initial request to facilitate reproducibility of the analysis. The methods used included standard multiple linear regression (lm), mixed-effects models (lme), generalized linear mixed models (glmer), autoregressive-moving-average (gls), boosted regression trees (gbm), and Bayesian methods (JAGS). In addition, three participants suggested using cross-validation methods. From this anecdotal sample, it is clear that there is little consensus about the standard accepted practice for ecological data analysis. Consequently, ecologists tend to use the methods with which they are most familiar. This lack of standardization in the statistical methods led to a range of conclusions about the importance of individual predictors from a single data set.

Given our instructions, our preferred analysis to explain variation in acorn production was a mixed-effects model with site and year as random effects and the remaining terms as fixed. After stepwise model selection, three terms were retained at $P < 0.05$, but two of these were marginally significant ($P = 0.049$) and would be removed after correcting for multiple testing ($P < 0.05/7 = 0.007$). In contrast, ΔAIC retained only the single highly significant predictor. Alternatively, to focus on interannual variation in acorn production, a time-series analysis could be performed (Fig. 2). Using either NHST or AIC, this approach retained more terms in the final model (four and five terms, respectively), including the one term (April rain lagged 1 year) retained by both NHST and AIC in the mixed-effects analysis. The two terms marginally significant ($P = 0.049$) by NHST in the mixed-effects analysis were both significant when performed as a time-series analysis, indicating that this method is more powerful for detecting significant environmental effects.

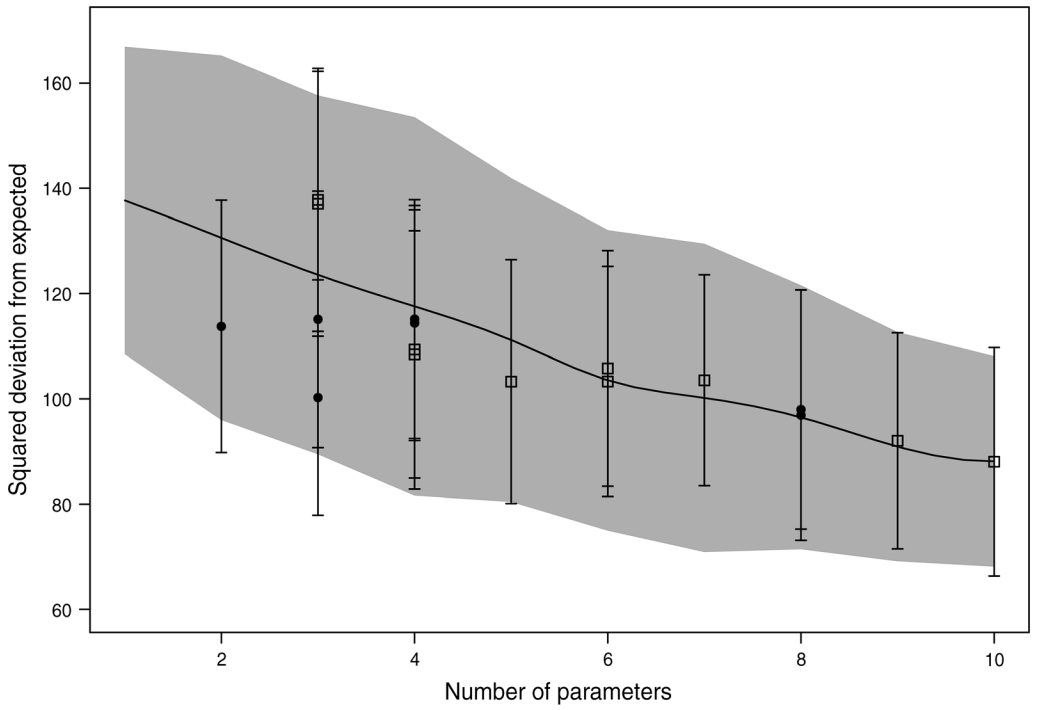

FIG. 3. Predictive ability of the final models selected by survey participants shown as the squared deviation from expected with 30% of observations as test data for 400 observations. Solid circles represent final models selected by AIC, while open squares are final models selected by NHST. Error bars are 95% confidence intervals of the 400 replicates. The gray-shaded region and black line are the 95% CI and mean of randomly selected variables for each number of parameters.

P VALUES, EFFECT SIZE, AND ECOLOGICAL INFERENCE

Murtaugh (2014) touches on three points that warrant further consideration. First, statistical tests using small sample sizes often lack power to reject the null hypothesis, even when the differences among means are large (Murtaugh 2014: Eq. 3). Without explicit consideration of the false negative rate, ecologists may fail to reject a false null hypothesis (a Type II or "consumer" error) when, in fact, they lack sufficient power to reject it. Arguably, the consequences of a consumer error are greater than rejecting a true null hypothesis (e.g., error on the side of caution), yet for small sample sizes typical of ecological studies, the standard $\alpha = 0.05$ makes a consumer error more likely than falsely rejecting a true null hypothesis (a Type I or "producer" error; Fig. 1 in Mudge 2012*a*). Murtaugh (2014) alludes to methods that allow the relative costs of consumer and producer errors to be made explicit, prior to the analysis (Mapstone 1995, Mudge 2012*a*, *b*). For example, Mudge et al. (2012*b*) reanalyze data from the Canadian Environmental Effects Monitoring Program, setting an optimal α that balances the cost of making a producer or consumer error for a given sample size. From this analysis, they found that 8–31% of the tests would have resulted in different management outcomes if an optimal α level (ranging from 6.6×10^{-5} to 0.309 with median = 0.073) had been used (Mudge et al. 2012*b*). Whereas the choice of the critical effect size and optimal consumer vs. producer error cost is somewhat subjective, interpretation of results and management decisions likely will be improved by explicit consideration of these parameters.

Second, Murtaugh (2014) states that "ΔAIC-based comparisons of nested models are often much more conservative than conventional hypothesis test done at the 0.05 level..." Our analysis is consistent with this, although we note that after correcting for multiple testing the *P* value based likelihood ratio test approach gives the same result as using ΔAIC. One of the challenges of AIC is that ecological studies frequently report only the "best" model extracted from automated AIC selection procedures, even though the others are likely to be as good as well (Whittingham et al. 2006). Burnham and Anderson (2002) have advocated in favor of reporting multiple models or performing model averaging. Competing with these approaches, a multitude of methods exist for *P* value correction including sequential Bonferroni (Rice 1989, but see Moran 2003) and false discovery rate (García 2004). With data sets of ever-increasing size being collected, it is becoming more common for the number of variables to exceed the number of observations, and correct application of these methods is imperative to avoid false positive associations.

Finally, ecologists in general pay more attention to the *P* values than to the parameters of biological interest; the effect sizes. We support the position of

Murtaugh (2014) and Nakagawa and Cuthill (2007) that estimated effect sizes should always be published alongside of *P* values. However, just as *P* values have limitations, the effect size is itself an estimate that is susceptible to bias. This is explicit in the bias–variance trade-off (Burnham and Anderson 2001:31) where increasing the predictive ability of a model (decreasing bias) increases the variance; model selection optimizes this trade-off. Moreover, due to the "winner's curse" (Zollner and Pritchard 2007, Button et al. 2013), a parameter that is found to be significant, especially in an under-powered study, is quite likely to have an exaggerated estimate of its effect size. This problem (e.g., the "Beavis effect" in plant genetics; Beavis 1994, Xu 2003) plagues the field of mapping phenotype to genotype as the effect size of significant quantitative trait loci (or nucleotides) are overestimated in initial studies and are found to be much smaller upon validation (Larsson et al. 2013). Thus, reporting effect sizes is imperative, but understanding that effect sizes can be biased is crucial for their interpretation.

Conclusion

This is not the first time that ecologists are being reminded that there are few laws of statistics. Thirteen years ago, Stewart-Oaten (1995) wrote in *Ecology* that "judgments based on opinions become laws of what 'should be done'," which echoes the sentiment of Murtaugh regarding *P* values and AIC (Murtaugh 2014). In face of this repeated problem, how are students of ecology supposed to learn the difference between an opinion and a law? Ellison and Dennis (2010) recommend ecologists gain statistical fluency through calculus and statistics. Houle et al. (2011) have argued that there is "... a systemic lack of respect for measurement and models in biology" and similarly calls for increased awareness of and education in quantitative skills. All ecologists may not be able to take courses in statistical theory, but there are ample opportunities for self-teaching of statistics in the practice of research (Ellison and Dennis 2010).

One suggestion that we have is that ecologists should more fully embrace the spirit of reproducible research (Gentleman and Lang 2004, Ellison 2010). In addition to archiving their raw data, which is now required by many journals, authors should make the source code freely available. In fact, this is now required practice at *Ecology*. R scripts can be archived at Ecological Archives or Dryad with data files, or made available through resources such as GitHub, which has the advantage of allowing for version control and collaboration. If readers are skeptical of the statistical analyses performed by the authors, they will be able to reanalyze the data, applying the methods they find most appropriate.

To conclude, notwithstanding 10 years calling for the abandonment of NHST using *P* values, we find that early-career ecologists continue to use *P* values, in addition to a battery of other statistical tools including AIC. We find this encouraging, as it was clear in the responses to our survey that ecologists are actively trying to use the best statistical methods possible in the face of uncertain and contradictory statistical advice. With colleagues such as these, we look forward to more robust and nuanced uses of statistics for addressing the major questions of ecology.

Acknowledgments

We thank the 24 ecologists who responded to our request to conduct a statistical analysis, and Nick Gotelli, Ruth Shaw, Charlie Geyer, and an anonymous reviewer for comments on the manuscript. Financial support was provided by the National Science Foundation DEB #1136717 to J. Stanton-Geddes and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Estágio Pós-Doutoral 3138135 and BEX 5366100) to C. G. Freitas and C. S. Dambros.

Literature Cited

Beavis, W. D. 1994. The power and deceit of QTL experiments: lessons from comparative QTL studies. Pages 250–266 *in* Proceedings of the 49th Annual Corn and Sorghum Research Conference. American Seed Trade Association, Chicago, Illinois, USA.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multi-model inference: a practical information-theoretic approach. Second edition. Springer, New York, New York, USA.

Button, K. S., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò. 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience 14:365–376.

Ellison, A. M. 2010. Repeatability and transparency in ecological research. Ecology 91:2536–2539.

Ellison, A. M., and B. Dennis. 2010. Paths to statistical fluency for ecologists. Frontiers in Ecology and the Environment 8:362–370.

Fisher, R. A. 1973. Statistical methods for research workers. 14th edition. Hafner Publishing, New York, New York, USA.

García, L. V. 2004. Escaping the Bonferroni iron claw in ecological studies. Oikos 105:657–663.

Gentleman, R., and D. Lang. 2004. Statistical analyses and reproducible research. Bioconductor Project Working Papers. Working Paper 2. Berkeley Electronic Press, Berkeley, California, USA.

Houle, D., C. Pélabon, G. P. Wagner, and T. F. Hansen. 2011. Measurement and meaning in biology. Quarterly Review of Biology 86:3–34.

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

Johnson, J. B., and K. S. Omland. 2004. Model selection in ecology and evolution. Trends in Ecology and Evolution 19:101–108.

Koenig, W. D., and J. M. H. Knops. 2013. Large-scale spatial synchrony and cross-synchrony in acorn production by two California oaks. Ecology 94:83–93.

Larsson, S. J., A. E. Lipka, and E. S. Buckler. 2013. Lessons from Dwarf8 on the strengths and weaknesses of structured association mapping. PLoS Genetics 9:e1003246.

Mapstone, B. D. 1995. Scalable decision rules for environmental impact studies: effect size, Type I and Type II errors. Ecological Applications 5:401–410.

Moran, M. D. 2003. Arguments for rejecting the sequential Bonferroni in ecological studies. Oikos 100:403–405.

Mudge, J. F., L. F. Baker, C. B. Edge, and J. E. Houlahan. 2012*a*. Setting an optimal α that minimizes errors in null hypothesis significance tests. PLoS ONE 7:e32734.

Mudge, J. F., T. J. Barrett, K. R. Munkittrick, and J. E. Houlahan. 2012*b*. Negative consequences of using $\alpha = 0.05$ for environmental monitoring decisions: a case study from a decade of Canada's Environmental Effects Monitoring Program. Environmental Science and Technology 46:9249–9255.

Murtaugh, P. A. 2014. In defense of *P* values. Ecology 95:611–617.

Nakagawa, S., and I. C. Cuthill. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Reviews 82:591–605.

R Development Core Team. 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.r-project.org

Rice, W. R. 1989. Analyzing tables of statistical tests. Evolution 43:223–225.

Stewart-Oaten, A. 1995. Rules and judgments in statistics: three examples. Ecology 76:2001–2009.

Whittingham, M. J., P. A. Stephens, R. B. Bradbury, and R. P. Freckleton. 2006. Why do we still use stepwise modelling in ecology and behaviour? Journal of Animal Ecology 75:1182–1189.

Xu, S. 2003. Theoretical basis of the Beavis effect. Genetics 165:2259–2268.

Zollner, S., and J. K. Pritchard. 2007. Overcoming the winner's curse: estimating penetrance parameters from case-control data. American Journal of Human Genetics 80:605–615.

Supplemental Material

Supplement

R script, results and example data set provided to early-career ecologists for a survey of statistical methods used in analysis of an ecological data set (*Ecological Archives* E095-054-S1).

Ecology, 95(3), 2014, pp. 642–645

Comment on Murtaugh

Michael Lavine[1]

University of Massachusetts, Amherst, Massachusetts 01003 USA

Murtaugh (2014) argues, "Since *P* values, confidence intervals, and ΔAIC are based on the same statistical information, all have their places in modern statistical practice. The choice of which to use should be stylistic . . ." and "To say that one of these metrics is always best ignores the complexities of ecological data analysis, as well as the mathematical relationships among the metrics."

On the whole, I agree. I will use this Comment to discuss some technical issues and to argue that *P* values, confidence intervals, and change in Akaike's information criterion (ΔAIC) should be viewed as descriptive statistics, not as formal quantifications of evidence.

Binary declarations of significance

I agree with Murtaugh that "One resolution of the problem of the arbitrariness of a cut-off . . . is to abandon the idea of the binary decision rule entirely and instead simply report the *P* value." However, most accept/reject declarations have no consequences, so I disagree with calling them decisions. To illustrate, after a medical trial, doctors must decide whether to prescribe a treatment and patients must decide whether to take it. But doctors' and patients' decisions need not agree with each other and need not agree with the original investigators' declaration of significance. It's not the investigators who decide; it's doctors and patients. Their decisions have consequences whose probabilities and utilities should guide the decisions.

Most accept/reject declarations have no consequences, are not guided by the probabilities and utilities of consequences, and cannot be recommended as substitutes for subsequent decisions. Though some authors explain accept/reject declarations in terms of 0–1 utility functions, those functions are chosen for explanatory value, not for realism. Where Murtaugh advises "instead simply report the *P* value," I argue that the declaration is not a useful entity that needs something else in its stead.

That we can abandon the declaration but still report a *P* value, confidence interval, or ΔAIC shows that the arbitrariness of 0.05 is an argument against the declaration, not against the *P* value, CI, or ΔAIC.

Manuscript received 12 June 2013; revised 23 July 2013; accepted 27 July 2013; final version received 16 August 2013. Corresponding Editor: A. M. Ellison. For reprints of this Forum, see footnote 1, p. 609.

[1] E-mail: lavine@math.umass.edu