Fisher, R. A. 1925. Statistical methods for research workers. Oliver and Boyd, Edinburgh, UK.

Fisher, R. A. 1955. Statistical methods and scientific induction. Journal of the Royal Statistical Society B 17:69–78.

Lehmann, E. L. 1986. Testing statistical hypotheses. Second edition. Wiley, New York, New York, USA.

Lindley, D. V. 1957. A statistical paradox. Biometrika 44:187–192.

Mayo, D. G. 1996. Error and the growth of experimental knowledge. University of Chicago Press, Chicago, Illinois, USA.

Mayo, D. G., and A. Spanos. 2004. Methodology in practice: statistical misspecification testing. Philosophy of Science 71:1007–1025.

Mayo, D. G., and A. Spanos. 2006. Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. British Journal for the Philosophy of Science 57:323–357.

Mayo, D. G., and A. Spanos. 2011. Error statistics. Pages 151–196 *in* D. Gabbay, P. Thagard, and J. Woods, editors. The handbook of philosophy of science, volume 7: philosophy of statistics. Elsevier, Amsterdam, The Netherlands.

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

Neyman, J. 1956. Note on an article by Sir Ronald Fisher. Journal of the Royal Statistical Society B 18:288–294.

Neyman, J., and E. S. Pearson. 1933. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society A 231:289–337.

Pearson, E. S. 1955. Statistical concepts in the relation to reality. Journal of the Royal Statistical Society B 17:204–207.

Spanos, A. 2007. Curve-fitting, the reliability of inductive inference and the error-statistical approach. Philosophy of Science 74:1046–1066.

Spanos, A. 2010*a*. Is frequentist testing vulnerable to the base-rate fallacy? Philosophy of Science 77:565–583.

Spanos, A. 2010*b*. Akaike-type criteria and the reliability of inference: model selection vs. statistical model specification. Journal of Econometrics 158:204–220.

Spanos, A. 2011. Misplaced criticisms of Neyman-Pearson (N-P) testing in the case of two simple hypotheses. Advances and Applications in Statistical Science 6:229–242.

Spanos, A. 2013. Who should be afraid of the Jeffreys-Lindley paradox? Philosophy of Science 80:73–93.

Ecology, 95(3), 2014, pp. 651–653
© 2014 by the Ecological Society of America

Rejoinder

PAUL A. MURTAUGH[1]

Department of Statistics, Oregon State University, Corvallis, Oregon 97331 USA

I thank the editors of *Ecology* for their interest in my paper, and the discussants for their extensive comments. I found myself agreeing with most of what was said, so I will make just a few observations.

Burnham and Anderson (2014) are mistaken when they say that the relationship between *P* values and AIC differences "holds only for the simplest case (i.e., comparison of two nested models differing by only one parameter)." As shown in Murtaugh (2014) Eqs. 5 and 6, the relationship holds for any k, i.e., for nested models differing by any number of parameters. It is also worth pointing out that the relationship holds for not only for nested linear models with Gaussian errors, as stated by Stanton-Geddes et al. (2014), but also for nested models with non-Gaussian errors if n is large (Murtaugh 2014: Eq. 5).

Burnham and Anderson (2014) comment that information-theoretic methods are "free from arbitrary cutoff values," yet they and others have published arbitrary guidelines for deciding how large a value of ΔAIC is needed for one model to be preferred over another (see Table 1). In any case, it is clear that both the *P* value and ΔAIC are continuous metrics, the interpretation of which is necessarily subjective (see my original Figs. 1 and 3).

De Valpine (2013) comments on the oft-repeated criticism that the *P* value is based on unobserved data, because it is the probability of obtaining a statistic at least as extreme as the observed statistic, given that the null hypothesis is true. As he suggests, any statistical method involving likelihoods is grounded in the assumption that a particular statistical distribution underlies both the observed and unobserved, hypothetical data, so that "part and parcel of that model are the probabilities associated with the unobserved data." I would add that Bayesians working with subjective priors also depend quite heavily on unobserved data.

It seems foolish to discard useful statistical tools because they are old (Burnham and Anderson 2014), or because they can only be applied in certain settings. I think it is healthy that the ecologists challenged by Stanton-Geddes et al. (2014) used a variety of methods to do their analyses, although it is disconcerting that the "participants came to surprisingly different conclusions." I wholeheartedly agree with Stanton-Geddes et

Manuscript received 1 October 2013; accepted 3 October 2013. Corresponding Editor: A. M. Ellison. For reprints of this Forum, see footnote 1, p. 609.

[1] E-mail: murtaugh@science.oregonstate.edu

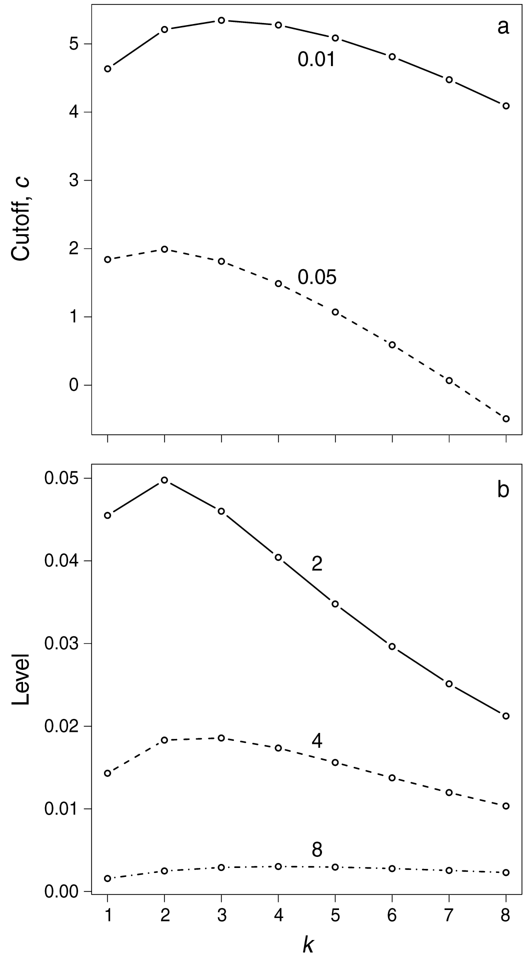

FIG. 1. For a test in which a reduced model is rejected in favor of a full model when $\Delta\text{AIC} = \text{AIC}_R - \text{AIC}_F > c$ (where AIC_R is AIC for the reduced model, AIC_F is AIC for the full model, and c is a cutoff value) (a) for a given level (i.e., probability of rejecting the reduced model when it is correct, set at 0.01 or 0.05), the relationship between the cutoff c needed to achieve that level, and the number of parameters k distinguishing the full and reduced models; and (b) for a given cutoff c (set at 2, 4, or 8), the relationship between the level of the test and k. The relationships are from Eq. 7 of Murtaugh (2014).

al. (2014) that "ecologists should more fully embrace the spirit of reproducible research," and I hope that recent attempts to increase the availability of raw data, combined with clearer explanations of statistical methodology, will help us understand why different analyses sometimes lead to different conclusions.

Burnham and Anderson (2014) express a common sentiment when they write that "step-up, step-down, and step-wise regression analyses represent perhaps the worst of these historical methods due partially to their reliance on a sequence of *P* values." In simulations (Murtaugh 1998) and cross-validation with real data sets (Murtaugh 2009), I failed to find support for this view. Methods based on *P* values and information-theoretic criteria performed comparably, which is not surprising since they are just different transformations of the likelihood ratio. It is perhaps more surprising that the algorithm used to compare these criteria among models, stepwise variable selection or all-subsets selection, also had little effect on the results (Murtaugh 2009).

As Lavine (2014) points out, the relationship between the *P* value and ΔAIC changes with *k*, the difference in the number of parameters between full and reduced models. That is, the value of ΔAIC corresponding to a particular *P* value, and vice-versa, changes with *k* (Murtaugh 2014: Eq. 7). Fig. 1 in this paper shows (a) how the ΔAIC cutoff needed to achieve a given level changes with *k*, and (b) how, for a given cutoff, the level of the test changes with *k*. Interestingly, these relationships are non-monotonic.

As seen in Fig. 1, ΔAIC is usually more conservative than the *P* value in comparing nested models, and the difference increases with the disparity in the sizes of the full and reduced models. There is nothing "wrong" with this; it simply reflects the philosophy embedded in AIC that the penalty for model complexity should be more severe than that inherent in the *P* value.

Lavine (2014) and Barber and Ogle (2014) discuss Schervish's (1996) interesting observation that the *P* value is "incoherent" in special cases, i.e., for two hypotheses, one of which is a subset of the other, the *P* value can indicate stronger support for the narrower hypothesis. In practice, we usually consider strength of evidence against a fixed null hypothesis for hypothetically variable data, rather than comparing the strength of evidence against two null hypotheses for a fixed set of data. Still, Schervish's result does add an important technical qualification to the general statement that *P* values indicate strength of evidence against the null hypothesis. As Lavine (2014) points out, a similar logical inconsistency arises with the use of ΔAIC in certain situations.

In my paper, I purposely avoided comparisons between hypothesis testing and Bayesian inference, partly because they stray from my main point and partly because it is difficult to compare the different currencies of the two approaches (but see, for example, Berger 2003). After an historical period of tension, frequentists and Bayesians now comfortably cohabit the pages of statistical journals, at least, and many scientists have argued for the value of both approaches in data analysis (e.g., see Breslow 1990, Efron 2005). But many ecologists still take the "either/or" approach, typically arguing for Bayesian approaches as a necessary improvement over the tired ideas of frequentists (e.g., see Hobbs and Hilborn 2006).

I couldn't agree more with Lavine's (2014) comments about the need for plots in conjunction with statistical summaries. The longer I have worked in statistics, the more convinced I have become that statistical analyses should be viewed as confirmations of patterns suggested by plots or other descriptive summaries, rather than as prima facie, stand-alone evidence of important associations. This is heresy to many of my colleagues and students, and there are, admittedly, applications where postulated patterns cannot be easily visualized in plots. But I am always skeptical of statistically significant associations, e.g., interactions between predictors in a regression model, for which I cannot find graphical evidence (e.g., see Murtaugh 2008).

In other comments, Spanos (2014) contrasts *P* values with other procedures in a broader historical and philosophical context than I provided, and he sensibly suggests that the choice between different procedures "should depend on the questions of interest, the answers sought, and the reliability of the procedures." Aho et al. (2014) discuss the Bayesian point of view and consider the relative strengths and appropriateness of the use of AIC and the Bayesian information criterion in different situations.

In summary, I reiterate that, in comparisons of nested linear models, *P* values and ΔAIC are just different transformations of the likelihood ratio, so that one metric cannot be 'better' than the other at discriminating between models. Unlike the *P* value, ΔAIC can be used to compare non-nested models. When either metric can be used, individual analysts may find the probability scale of the *P* value easier to understand than the Kullbach-Leibler information of ΔAIC, or vice-versa, but that is a matter of preference, not scientific legitimacy. Both approaches have long traditions of usefulness in data analysis, and it seems pointless to urge practitioners to abandon one in favor of the other.

Acknowledgments

I thank Claudio Fuentes, Bruce McCune, and Bryan Wright for encouraging me to follow through with this project when I was ready to shelve it, and C. Fuentes for helpful discussions along the way. None of these colleagues necessarily shares the views I have expressed.

Literature Cited

Aho, K., D. Derryberry, and T. Peterson. 2013. Model selection for ecologists: the worldviews of AIC and BIC. Ecology 95:631–636.

Barber, J. J., and K. Ogle. 2014. To *P* or not to *P*? Ecology 95:621–626.

Berger, J. O. 2003. Could Fisher, Jeffreys and Neyman have agreed on testing? Statistical Science 18:1–32.

Breslow, N. 1990. Biostatistics and Bayes. Statistical Science 6:269–298.

Burnham, K. P., and D. R. Anderson. 2014. *P* values are only an index to evidence: 20th- vs. 21st-century statistical science. Ecology 95:627–630.

de Valpine, P. 2014. The common sense of *P* values. Ecology 95:617–621.

Efron, B. 2005. Bayesians, frequentists, and scientists. Journal of the American Statistical Association 100:1–5.

Hobbs, N. T., and R. Hilborn. 2006. Alternatives to statistical hypothesis testing in ecology: a guide to self teaching. Ecological Applications 16:5–19.

Lavine, M. 2014. Comment on Murtaugh. Ecology 95:642–645.

Murtaugh, P. A. 1998. Methods of variable selection in regression modeling. Communications in Statistics—Simulation and Computation 27:711–734.

Murtaugh, P. A. 2008. No evidence for an interactive effect of herbivore and predator diversity on herbivore abundance in the experimental mesocosms of Douglass et al. (2008). Ecology Letters 11:E6–E8.

Murtaugh, P. A. 2009. Performance of several variable-selection methods applied to real ecological data. Ecology Letters 12:1061–1068.

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

Schervish, M. J. 1996. *P* values: what they are and what they are not. American Statistician 50:203–206.

Spanos, A. 2014. Recurring controversies about *P* values and confidence intervals revisited. Ecology 95:645–651.

Stanton-Geddes, J., C. G. de Freitas, and C. de Sales Dambros. 2014. In defense of *P* values: comment on the statistical methods actually used by ecologists. Ecology 95:637–642.