Ecology, 95(3), 2014, pp. 611–617

In defense of *P* values

Paul A. Murtaugh[1]

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

Abstract. Statistical hypothesis testing has been widely criticized by ecologists in recent years. I review some of the more persistent criticisms of *P* values and argue that most stem from misunderstandings or incorrect interpretations, rather than from intrinsic shortcomings of the *P* value. I show that *P* values are intimately linked to confidence intervals and to differences in Akaike's information criterion (ΔAIC), two metrics that have been advocated as replacements for the *P* value. The choice of a threshold value of ΔAIC that breaks ties among competing models is as arbitrary as the choice of the probability of a Type I error in hypothesis testing, and several other criticisms of the *P* value apply equally to ΔAIC. 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, dictated by details of the application rather than by dogmatic, a priori considerations.

Key words: *AIC; confidence interval; null hypothesis;* P *value*; *significance testing.*

In the 1970s, a number of authors argued for the systematic use of null and alternative hypotheses when framing research questions in ecology (e.g., see Strong 1980). They were later rebutted by others who judged this approach was overly restrictive and potentially misleading (Quinn and Dunham 1983, Loehle 1987). An interesting analogue to that history has occurred more recently in the realm of statistical hypothesis testing in ecology. Long a mainstay in ecological data analysis, the use of hypothesis testing has been increasingly frowned upon in recent years (Johnson 1999, Anderson et al. 2000, Burnham and Anderson 2002, Gerrodette 2011).

The tone of the criticisms has been surprisingly vehement, accompanied by much hand wringing about the future of a science that is still so burdened with statistical hypothesis testing (e.g., see Anderson et al. 2001, Fidler et al. 2006, Martinez-Abrain 2008, Gerrodette 2011). Anderson et al. (2000) generalize their criticisms beyond ecology, commenting that "tests of statistical null hypotheses have relatively little utility in science ... ." Most of the critics of significance testing advocate alternative approaches based on information-theoretic criteria or Bayesian statistics (Johnson 1999, Burnham and Anderson 2002, Hobbs and Hilborn 2006, Lukacs et al. 2007).

Stephens et al. (2005) summarize, and respond to, recent criticisms of statistical hypothesis testing in ecology, arguing that some are unfounded and others stem from misuse of these procedures by practitioners. Hurlbert and Lombardi (2009) also consider criticisms that have been leveled against significance testing, noting that most of them "concern the misuse and misinterpretation of significance and *P* values by investigators and not the inherent properties ... of the tests or *P* values themselves," and they make suggestions for the appropriate use and interpretation of *P* values. Mundry (2011) discusses some of the limitations of information-theoretic methods and argues for a balanced approach in which both those methods and hypothesis testing are used, with the choice of method dictated by the circumstances of the analysis.

In this paper, I review and comment on some of the more persistent criticisms of statistical hypothesis testing in ecology, focusing on the centerpiece of that approach, the *P* value. Addressing suggestions that confidence intervals and information-theoretic criteria are superior to *P* values, I argue that, since all three tools are based on the same statistical information, the choice of which summary to present should be largely stylistic, depending on details of the application at hand. I conclude that *P* values, confidence intervals, and information-theoretic criteria all have their places in sound statistical practice, and that none of them should be excluded based on dogmatic, a priori considerations.

The definition and interpretation of the P *value*

Consider the comparison of two nested linear models, i.e., two models such that one (the "reduced" model) is a special case of the other (the "full" model). Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ be the vector of unknown parameters for the full model, and assume that the

Manuscript received 28 March 2013; accepted 25 April 2013; final version received 29 May 2013. Corresponding Editor: A. M. Ellison. For reprints of this Forum, see footnote 1, p. 609.

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

reduced model is obtained by setting the first k parameters equal to zero.

Based on a set of independent observations, $y_1, \ldots, y_n$, we can test the null hypothesis that $\theta_1 = \theta_2 = \cdots = \theta_k = 0$ using the likelihood ratio statistic

$$\Lambda = -2 \log\left\{\mathcal{L}(\hat{\boldsymbol{\theta}}_0)/\mathcal{L}(\hat{\boldsymbol{\theta}})\right\}$$

where $\hat{\boldsymbol{\theta}}$ is the vector of maximum likelihood estimates (MLEs) for the full model, $\hat{\boldsymbol{\theta}}_0$ is the vector of constrained MLEs under the null hypothesis, and $\mathcal{L}(\cdot)$ is the likelihood, i.e., the joint probability density function of the data, expressed as a function of the parameters.

The P value (P) is the probability of obtaining a statistic at least as extreme as the observed statistic, given that the null hypothesis is true. For a broad array of distributions of the data, Λ will have a χ^2 distribution with k degrees of freedom for large n, if the null hypothesis is true. Therefore, for a particular observed value of the statistic, Λ^*,

$$P = \Pr(\chi^2_k > \Lambda^*). \quad (1)$$

The smaller the P value, the more evidence we have against the null hypothesis.

Comparisons of nested linear models are ubiquitous in statistical practice, occurring in the contexts of the two-sample comparison, one- and multi-way analysis of variance, simple and multiple linear regression, generalized linear models, the χ^2 test for contingency tables, survival analysis, and many other applications.

In the special case of nested linear models with Gaussian errors, the MLE of $\boldsymbol{\theta}$ coincides with the least-squares estimate, i.e., the value that minimizes the error sum of squares, $\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, where $\hat{y}_i$ is the fitted value for the ith observation. An exact P value can be obtained from the extra-sum-of-squares F statistic:

$$F^* = \frac{(\text{SSE}_\text{R} - \text{SSE}_\text{F})/k}{\text{SSE}_\text{F}/(n-p+1)}$$

$$P = \Pr(F_{k,\,n-p+1} > F^*), \quad (2)$$

where SSE_F and SSE_R are the error sums of squares for the full and reduced models, respectively, and $F_{k,\,n-p+1}$ is a random variable from the F distribution with k and $n - p + 1$ degrees of freedom.

To understand the factors influencing the P value, consider the simple example of the comparison of two population means, μ_1 and μ_2, based on two independent samples of size n_1 and n_2. Let y_{ij} be the jth observation in group i ($i = 1, 2$; $j = 1, \ldots, n_i$); let $\bar{y}_{i\cdot}$ be the average of the n_i observations in group i ($i = 1, 2$); and let $s_\text{p}^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2/(n_1 + n_2 - 2)$ be the pooled sample variance. The equal-variances t statistic is

$$T^* = \frac{\bar{y}_{2\cdot} - \bar{y}_{1\cdot}}{\sqrt{s_\text{p}^2(1/n_1 + 1/n_2)}}.$$

It can be shown that $(T^*)^2$ is equal to the extra-sum-of-squares F statistic from Eq. 2. An exact P value for testing the equality of means, identical to that from Eq. 2, is

$$P = 2 \times \Pr(t_{n_1+n_2-2} > |T^*|)$$

$$= 2 \times \Pr\left\{t_{n_1+n_2-2} > \frac{|\bar{y}_{2\cdot} - \bar{y}_{1\cdot}|}{\sqrt{s_\text{p}^2(1/n_1 + 1/n_2)}}\right\} \quad (3)$$

where $t_{n_1+n_2-2}$ is a random variable from the t distribution with $n_1 + n_2 - 2$ degrees of freedom.

A very small P value indicates that the data are not consistent with the null hypothesis, leading us to prefer the alternative hypothesis that the two populations have different means. Note from Eq. 3 that a small P value can result from a small denominator of the t statistic, as well as from a large numerator ($|\bar{y}_{2\cdot} - \bar{y}_{1\cdot}|$). That is, the P value decreases as the pooled sample variance, s_p^2, decreases and as the sample sizes, n_1 and n_2, increase. Hence, it is pointless to report a P value without also reporting the observed difference between means; depending on the variance and sample size, it is possible to obtain small P values for practically unimportant differences between means, and large P values for large differences between means.

Some persistent criticisms of the P *value*

The 0.05 level is arbitrary.—Discussing the use of the standard normal distribution in hypothesis testing, R. A. Fisher (1973:44) wrote, "The value for which $P = 0.05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant."

Fisher's thinking on this subject evolved over his lifetime, as he became more sympathetic to the idea of reporting exact P values, rather than adhering to the binary decision rule (Hurlbert and Lombardi 2009). Nevertheless, Fisher's recipe for interpreting the results of hypothesis tests was adopted with enthusiasm by the scientific community, to the extent that many authors appear to believe that (1) there is a firm cutoff between significant and nonsignificant results, with P values just above the cutoff to be interpreted differently from P values just below the cutoff, and (2) 0.05 is the sole reasonable choice for this cutoff. The arbitrariness of the choice of the cutoff and the rigidity with which it is often applied has been pointed out by many authors (Johnson 1999, Anderson et al. 2000, Rinella and James 2010).

In hypothesis testing, one can mistakenly reject a true null hypothesis (a Type I error, occurring with probability α) or fail to reject a false null hypothesis (a Type II error). Even though the practice of setting α equal to 0.05 is firmly entrenched in the scientific literature, a case can be made that the "acceptable" rate of Type I errors should be allowed to vary from

application to application, depending on the cost of such errors or, perhaps, the relative costs of Type I and Type II errors (Mapstone 1995, Johnson 1999, Hanson 2011, Mudge et al. 2012).

One resolution of the problem of the arbitrariness of a cutoff for statistical significance is to abandon the idea of the binary decision rule entirely and instead simply report the *P* value, along with the estimated effect size, of course (Ramsey and Schafer 2002:47; Hurlbert and Lombardi 2009). The *P* value is a continuous measure of the strength of evidence against the null hypothesis, with very small values indicating strong evidence of a difference between means (in the two-sample comparison), large values indicating little or no evidence of a difference, and intermediate values indicating something in between, as shown in Fig. 1.

It is clear that a decision rule leading to very different interpretations of *P* values of 0.049 and 0.051 is not very rational. The prevalence of that view in the scientific literature is a fault not of the conceptual basis of hypothesis testing, but rather of practitioners adhering too rigidly to the suggestions of R. A. Fisher many decades ago.

Confidence intervals are better than P *values*.—Many authors have advocated the use of confidence intervals instead of *P* values (Johnson 1999, Di Stefano 2004, Nakagawa and Cuthill 2007, Rinella and James 2010). In fact, the two are based on identical information: the point estimate of a parameter, the standard error of the estimate, and a statistical distribution that applies when the null hypothesis is true. A $100(1 - \alpha)\%$ confidence interval for a parameter θ is the set of all values θ^* for which we would fail to reject the null hypothesis $\theta = \theta^*$ at level α. (This relationship is not exact if the standard error depends on θ, as in the case of a Bernoulli random variable.)

So, *P* values and confidence intervals are just different ways of summarizing the same information. As mentioned earlier, a point estimate of the effect or association of interest should always be provided. Whether a *P* value or confidence interval is the more appropriate adjunct to the estimate depends on the setting. If a particular null hypothesis is of interest (e.g., that there is no effect of some experimental treatment on a response), a *P* value might be the most pertinent summary of the uncertainty of the estimate, reflecting its distance from the null-hypothesized value. But, if the focus is on describing an association for which there is no particular null-hypothesized value (in an observational study, for example), a confidence interval gives a succinct summary of the precision of the estimate.

Some authors have implied that the arbitrariness of selecting the 0.05 level in hypothesis testing can be skirted by using confidence intervals (Nakagawa and Cuthill 2007). But, of course, the choice of the coverage of the confidence interval (usually 95%) is every bit as arbitrary as the choice of the level of a hypothesis test.

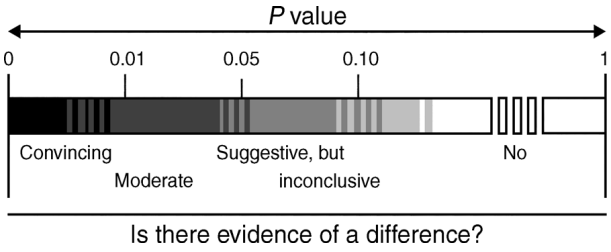

FIG. 1. Interpretation of the *P* value. Reprinted with permission from Ramsey and Schafer (2002).

Statistical significance does not imply practical significance.—This is of course true, and, as mentioned earlier, a *P* value should not be reported without also reporting the observed effect size. Nevertheless, authors continue to denigrate the *P* value because, in isolation, it does not specify the effect size (Fidler et al. 2006, Nakagawa and Cuthill 2007, Rinella and James 2010, Gerrodette 2011), or because statistical significance of a result is often confused with practical significance (Yoccoz 1991, Johnson 1999, Anderson et al. 2000, Martinez-Abrain 2008).

The null hypothesis is usually false.—Many authors have commented that many or most null hypotheses in ecology are known to be false (Anderson et al. 2000, Burnham et al. 2011, Gerrodette 2011). Null hypotheses are probably most useful in analyzing data from randomized experiments (Eberhardt 2003), in which the null hypothesis would be literally true if the treatment(s) had no effect on the response of interest. In observational studies, null hypotheses often do seem a priori implausible (but see Stephens et al. 2005, and Mundry 2011), in which case it is always an option to test for the existence of some small, marginally meaningful association. Or, it might be preferable to report a confidence interval for the difference, based on the same information that would be used in the hypothesis test.

P *values don't tell us what we want to know*.—*P* values continue to be maligned because they are sometimes mistakenly interpreted as the probability that the null hypothesis is true, or because one minus the *P* value is wrongly interpreted as the probability that the alternative hypothesis is true (Johnson 1999, Rinella and James 2010, Gerrodette 2011). Many appear to wish that the *P* value *would* give the probability that the null hypothesis is true, given the data, instead of the probability of the data given the null hypothesis. Hobbs and Hilborn (2006) comment that "... *P* values associated with traditional statistical tests do not assess the strength of evidence supporting a hypothesis or model." This is literally true, since the *P* value summarizes the strength of evidence *against* the null hypothesis. (Puzzlingly, Martinez-Abrain [2008] writes that "By no means is it true that the smaller the *P* value the bigger the evidence against the null hypothesis.") But this seems like an

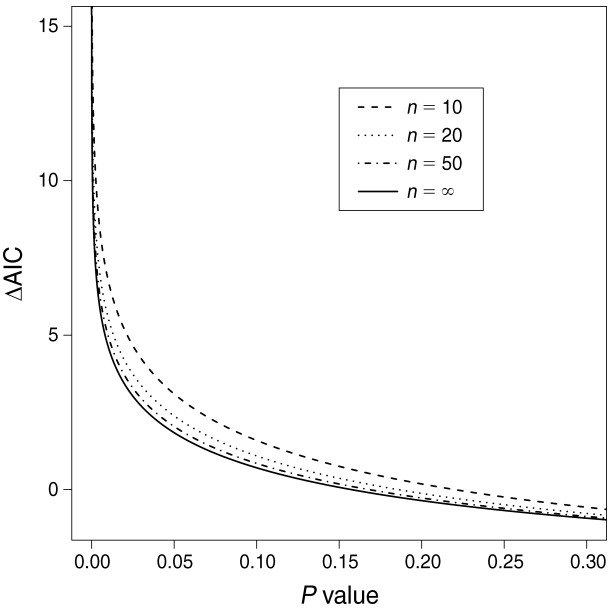

FIG. 2. The relationship between ΔAIC (as defined in Eq. 4) and the *P* value in a comparison of two models differing with respect to one parameter (as in a two-sample comparison, or simple linear regression), for different total sample sizes (*n*). The lines for finite *n* are based on the least-squares case (Eq. 6), and the line for $n = \infty$ is based on the asympotic distribution of the likelihood ratio statistic (Eq. 5).

unfair criticism, because (1) in a head-to-head comparison of nested models, which I have been emphasizing here, evidence against the simpler model necessarily weighs in favor of the more complicated model, and (2) the *P* value is based on the same information used by the information-theoretic criteria favored by Hobbs and Hilborn (2006), as I will discuss next.

The relationship between the P *value and Akaike's information criterion*

Information-theoretic criteria like Akaike's information criterion (AIC) have been widely touted as superior tools for deciding among statistical models, as compared to the *P* value (Anderson et al. 2000, Burnham and Anderson 2002, Gerrodette 2011). While proponents of AIC are loath to use it in a hypothesis-testing framework or as a tool for judging when one model is "significantly" better than another (Burnham and Anderson 2002), it is instructive to compare the conventional hypothesis-testing approach to a ranking of two candidate models based on AIC (e.g., see Mundry 2011).

For a model with *p* parameters, Akaike's information criterion is

$$\mathrm{AIC} = -2\log\mathcal{L}(\hat{\boldsymbol{\theta}}) + 2p,$$

where $\mathcal{L}(\hat{\boldsymbol{\theta}})$ is the maximized likelihood. The difference in AIC for a full model containing *p* parameters and a reduced model obtained by setting *k* of those parameters equal to zero is

$$\begin{aligned}\Delta\mathrm{AIC} &= \mathrm{AIC_R} - \mathrm{AIC_F} = -2\log\left\{\mathcal{L}(\hat{\boldsymbol{\theta}}_0)/\mathcal{L}(\hat{\boldsymbol{\theta}})\right\} - 2k \\ &= \Lambda - 2k, \end{aligned} \quad (4)$$

where $\mathcal{L}(\hat{\boldsymbol{\theta}}_0)$ and $\mathcal{L}(\hat{\boldsymbol{\theta}})$ are the maximized likelihoods of the data under the null and alternative hypotheses, respectively, and Λ is the likelihood ratio test statistic. This result implies that the relative likelihood of the full and reduced models is

$$\mathcal{L}(\hat{\boldsymbol{\theta}})/\mathcal{L}(\hat{\boldsymbol{\theta}}_0) = \exp\left(\frac{\Delta\mathrm{AIC}}{2} + k\right).$$

Eqs. 1 and 4 imply the following relationships between the *P* value and ΔAIC:

$$P = \Pr(\chi^2_k > \Delta\mathrm{AIC} + 2k)$$

and

$$\Delta\mathrm{AIC} = F^{-1}_{\chi^2_k}(1-p) - 2k, \quad (5)$$

where χ^2_k is a chi-square random variable with *k* degrees of freedom, and

$$F^{-1}_{\chi^2_k}(1-p)$$

is the $(1 - p)$ quantile of the χ^2_k distribution. This relationship between the *P* value and ΔAIC is shown graphically in Fig. 2 (solid line).

In the special case of nested linear models with Gaussian errors, it can be shown that $\Delta AIC = n\log(\mathrm{SSE_R}/\mathrm{SSE_F}) - 2k$. Combined with Eq. 2, this leads to the following relationships:

$$P = \Pr\left\{F_{k,\,n-p+1} > \frac{n-p+1}{k} \times \left[\exp\left(\frac{\Delta\mathrm{AIC}+2k}{n}\right) - 1\right]\right\}$$

$$\Delta\mathrm{AIC} = n\log\left[\frac{k}{n-p+1} \times F^{-1}_{F_{k,\,n-p+1}}(1-P) + 1\right] - 2k, \quad (6)$$

where

$$F^{-1}_{F_{k,\,n-p+1}}(1-P)$$

is the $(1 - P)$ quantile of an *F* distribution with *k* and $n - p + 1$ degrees of freedom. For large *n*, these relationships are approximately equivalent to those based on the likelihood ratio statistic (Eq. 5). Fig. 2 shows some examples.

Suppose we decide to reject the reduced model in favor of the full model when ΔAIC exceeds some positive cutoff, *c*. This decision rule is equivalent to a conventional hypothesis test done at a level determined by *c* and by the number of parameters *k* differing between the full and reduced models. If α is the level (i.e., the probability of rejecting the reduced model when it is "correct"), it follows from Eq. 5 that

TABLE 1. Interpretations of ΔAIC by different authors.

$AIC_i - AIC_j$	Relative likelihood (j:i)	Interpretation
Reference 1		
>1–2	>1.6–2.7	significant difference between models i and j
Reference 2		
4.2	8	strong enough difference to be of general scientific interest
6.9	32	"quite strong" evidence in favor of model j
Reference 3		
0–4.6	1–10	limited support for model j
4.6–9.2	10–100	moderate support
9.2–13.8	100–1000	strong support
>13.8	>1000	very strong support
Reference 4		
0–2	1–2.7	substantial support of model i
4–7	7.4–33.1	considerably less support
>10	>148	essentially no support
Reference 5		
0 to 4–7	1 to 7.4–33.1	model i is plausible
7–14	33.1–1097	value judgments for hypotheses in this region are equivocal
>14	>1097	model i is implausible

Notes: In my discussion, model i is the reduced model and model j is the full model. The greater the values of ($AIC_i - AIC_j$) and the relative likelihood, the greater the support for the full model. References are 1, Sakamoto et al. (1986:84); 2, Royall (1997:89–90); 3, Evett and Weir (1998), as quoted in Lukacs et al. (2007); 4, Burnham and Anderson (2002:70); 5, Burnham et al. (2011:25).

$$\alpha = \Pr(\chi^2_k > 2k + c) \quad \text{and} \quad c = F^{-1}_{\chi^2_k}(1 - \alpha) - 2k. \quad (7)$$

As shown in Table 1 and Fig. 3, the choice of a "critical" value of ΔAIC appears to be even more subjective than the choice of the probability of a Type I error in hypothesis testing. The value of ΔAIC considered large enough to break ties among competing models ranges from as little as 1 (relative likelihood of 1.6) to as high as 14 (relative likelihood of 1097). The most modern prescription, from Burnham et al. (2011), suggests that two models with a relative likelihood as high as 33 are still essentially indistinguishable, and that the superior model must be at least 1097 times as likely as the inferior competitor, before the latter can be confidently discarded. It is not clear to me what guides these recommendations and why they vary so much between authors and even within authors writing at different times.

Eq. 7 implies that, for threshold values of ΔAIC that are currently in use, ΔAIC-based comparisons of nested models are often much more conservative than conventional hypothesis tests done at the 0.05 level, a direct consequence of the extra penalty for model complexity that ΔAIC imposes, compared to P value based directly on the likelihood ratio statistic. For example, for a ΔAIC cutoff of 7, the corresponding significance level is 0.003 when $k = 1$ (as in a two-sample comparison, or simple linear regression); it reaches a maximum value of 0.005 when $k = 4$; and it approaches zero as k increases beyond 4.

Because of the one-to-one relationship between the P value and ΔAIC (Fig. 2), several of the criticisms leveled at the P value also apply to ΔAIC. In particular, the choice of 4 or 7 (or 1 or 14) as the threshold for declaring one model superior to another is just as arbitrary as the choice of $P = 0.05$ as the cutoff for statistical significance in a hypothesis test; ΔAIC does not include an estimate of the effect size; and a value of ΔAIC exceeding the chosen threshold does not imply that the difference between models is practically important. As I did for the P value, I would argue that none of these issues is an inherent problem of ΔAIC, which, when used properly, produces a comparison between models that is as informative as that provided by a hypothesis test or confidence interval.

CONCLUSIONS

P values, confidence intervals, and information-theoretic criteria are just different ways of summarizing

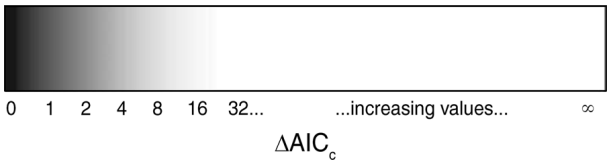

FIG. 3. Interpretation of ΔAIC, from Burnham et al. (2011). "Plausible hypotheses are identified by a narrow region in the continuum where Δ < perhaps four to seven (black and dark gray). The evidence in the light grey area is inconclusive and value judgments for hypotheses in this region are equivocal. Implausible models are shown in white, Δ > about 14." (The authors define Δ, or ΔAIC_c, as the difference between the value of AIC_c for a focal model and the minimum value of AIC_c in a group of models, where AIC_c is a modification of AIC that includes a correction for small sample size.)

the same statistical information. The intimate link between P values and confidence intervals is obvious: a $100(1 - \alpha)\%$ confidence interval is the set of parameter values that would yield P values less than or equal to α in two-sided hypothesis tests.

The connection between P values and ΔAIC is just as direct, as shown in Fig. 2. This is perhaps surprising, because the two approaches use apparently different yardsticks in comparing models: a P value from an F test is a probability based on a specific distribution that the test statistic will follow if the null hypothesis is true, while ΔAIC is simply based on the relative likelihood of the data under two different models, penalized by the disparity in model complexity. Nonetheless, deciding how small a P value is needed for us to prefer the more complicated model is equivalent to deciding how large a ratio of likelihoods indicates a convincing difference between models.

The comparison of P values and ΔAIC considered here is set in the home territory of the P value, namely a head-to-head comparison of two models, one of which is a special case of the other. An important advantage of the information-theoretic criteria over the P value is their ability to rank two or more models that are *not* nested in this way. In comparisons of nested models, however, many practitioners will find the scale of the P value—which expresses the probability of data, given the null hypothesis—easier to understand and interpret than the scale of ΔAIC, in units of Kullback-Leibler distance between models. This is reflected by the order-of-magnitude variation in the range of suggested "critical" values of ΔAIC (Table 1), compared to the relatively narrow range of levels that are used in conventional hypothesis testing.

Consideration of the close relationships among P values, confidence intervals and ΔAIC leads to the unsurprising conclusion that all of these metrics have their places in modern statistical practice. A test of the effects of treatments in a randomized experiment is a natural setting for a P value from the analysis of variance. The summary of a difference in some response between groups in an observational study is often well-accomplished with a confidence interval. ΔAIC can be used in either of the above settings, and it can be useful in other situations involving the comparison of non-nested statistical models, where the P value is of no help. 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.

Data analysis can be always be redone with different statistical tools. The suitability of the data for answering a particular scientific question, however, cannot be improved upon once a study is completed. In my opinion, it would benefit the science if more time and effort were spent on designing effective studies with adequate replication (Hurlbert 1984), and less on advocacy for particular tools to be used in summarizing the data.

Acknowledgments

I thank Sarah Emerson and two anonymous reviewers for constructive criticisms of an earlier version of the manuscript, and K. Zongo for finding an error in one of my derivations.

Literature Cited

Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64:912–923.

Anderson, D. R., W. A. Link, D. H. Johnson, and K. P. Burnham. 2001. Suggestions for presenting the results of data analyses. Journal of Wildlife Management 65:373–378.

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

Burnham, K. P., D. R. Anderson, and K. P. Huyvaert. 2011. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behavioral Ecology and Sociobiology 65:23–35.

Di Stefano, J. 2004. A confidence interval approach to data analysis. Forest Ecology and Management 187:173–183.

Eberhardt, L. L. 2003. What should we do about hypothesis testing? Journal of Wildlife Management 67:241–247.

Evett, I. W., and B. S. Weir. 1998. Interpreting DNA evidence: statistical genetics for forensic scientists. Sinauer Associates, Sunderland, Massachusetts, USA.

Fidler, F., M. A. Burgman, G. Cumming, R. Buttrose, and N. Thomason. 2006. Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. Conservation Biology 20:1539–1544.

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

Gerrodette, T. 2011. Inference without significance: measuring support for hypotheses rather than rejecting them. Marine Ecology: An Evolutionary Perspective 32:404–418.

Hanson, N. 2011. Using biological data from field studies with multiple reference sites as a basis for environmental management: the risks for false positives and false negatives. Journal of Environmental Management 92:610–619.

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

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54:187–211.

Hurlbert, S. H., and C. Lombardi. 2009. Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. Annales Zoologici Fennici 46:311–349.

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

Loehle, C. 1987. Hypothesis testing in ecology: psychological aspects and the importance of theory maturation. Quarterly Review of Biology 62:397–409.

Lukacs, P. M., W. L. Thompson, W. L. Kendall, W. R. Gould, P. F. Doherty, Jr., K. P. Burnham, and D. R. Anderson. 2007. Concerns regarding a call for pluralism of information theory and hypothesis testing. Journal of Applied Ecology 44:456–460.

Martinez-Abrain, A. 2008. Statistical significance and biological relevance: a call for a more cautious interpretation of results in ecology. Acta Oecologica—International Journal of Ecology 34:9–11.

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

Mudge, J. F., L. F. Baker, C. B. Edge, and J. E. Houlahan. 2012. Setting an optimal α that minimizes errors in null hypothesis significance tests. PLoS ONE 7(2):e32734.

Mundry, R. 2011. Issues in information theory-based statistical inference—a commentary from a frequentist's perspective. Behavioral Ecology and Sociobiology 65:57–68.
Nakagawa, S., and I. C. Cuthill. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Reviews 82:591–605.
Quinn, J. F., and A. E. Dunham. 1983. On hypothesis testing in ecology and evolution. American Naturalist 122:602–617.
Ramsey, F., and D. Schafer. 2002. The statistical sleuth: a course in methods of data analysis. Second edition. Duxbury Press, Belmont, California, USA.
Rinella, M. J., and J. J. James. 2010. Invasive plant researchers should calculate effect sizes, not *P*-values. Invasive Plant Science and Management 3:106–112.
Royall, R. M. 1997. Statistical evidence: a likelihood paradigm. Chapman and Hall, New York, New York, USA.
Sakamoto, Y., M. Ishiguro, and G. Kitagawa. 1986. Akaike information criterion statistics. D. Reidel Publishing, Hingham, Massachusetts, USA.
Stephens, P. A., S. W. Buskirk, G. D. Hayward, and C. Martinez del Rio. 2005. Information theory and hypothesis testing: a call for pluralism. Journal of Applied Ecology 42:4–12.
Strong, D. R. 1980. Null hypotheses in ecology. Synthese 43:271–285.
Yoccoz, N. G. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. Bulletin of the Ecological Society of America 72:106–111.

Ecology, 95(3), 2014, pp. 617–621

The common sense of *P* values

Perry de Valpine[1]

Department of Environmental Science, Policy and Management, 130 Mulford Hall #3114, University of California, Berkeley, California 94720-3114 USA

When perplexed graduate students ask me about the anti-*P*-value arguments they've heard, I offer them many of the same responses as Murtaugh (2014), and some others as well. Now I can start by having them read his paper. In this comment, I will support his basic message but dig more deeply into some of the issues.

What are *P* values for? The purpose of *P* values is to convince a skeptic that a pattern in data is real. Or, when you are the skeptic, the purpose of *P* values is to convince others that a pattern in data could plausibly have arisen by chance alone. When there is a scientific need for skeptical reasoning with noisy data, the logic of *P* values is inevitable.

Say there is concern that the chemical *gobsmackene* is toxic to frogs, but *gobsmackene* is an effective insecticide. The proponents of *gobsmackene* are vehement skeptics of its toxicity to frogs. You run an experiment, and the resulting *P* value is 0.001 against the null hypothesis that *gobsmackene* has no effect on frogs. As the expert witness in a trial, you explain to the judge that, if *gobsmackene* is not toxic to frogs, the chances of obtaining data at least as extreme as yours just by a fluke are tiny: just one in a thousand. The judge interprets this probabilistic statement about your evidence and bans *gobsmackene*.

Manuscript received 2 July 2013; revised 10 September 2013; accepted 10 September 2013. Corresponding Editor: A. M. Ellison. For reprints of this Forum, see footnote 1, p. 609.

[1] E-mail: pdevalpine@berkeley.edu

Now take an example from the other side, where you are the skeptic. Suppose someone claims to have a treatment that neutralizes a soil contaminant. She presents experimental data from 20 control and 20 treated plots, and the treated plots have 30% less of the contaminant than the control plots. You examine the data and determine that the variation between replicates is so large that, even if the treatment really has no effect, there would be a 20% chance of reporting an effect at least that big, i.e., $P = 0.20$. Since that is more likely than having three children turn out to be all boys, you are not convinced that their treatment is really effective. Of course, one doesn't need good-guy/bad-guy cartoons to imagine the kind of serious skepticism for which *P* value reasoning is useful.

These uses of *P* value reasoning seem like common sense. Why, then, is there so much controversy about such reasoning? I agree with Murtaugh (2014) that many anti-*P*-value arguments boil down to frustrations with practice rather than principle. For example, the arbitrariness of the conventional 0.05 threshold for significance is an example of the "fallacy of the beard." How many whiskers does it take to make a beard? Because it is impossible to give a precise answer that doesn't admit exceptions, should you choose never to discuss beards? Similarly, the arbitrariness of 0.05 is unavoidable, but that doesn't mean we shouldn't consider *P* values as one way to interpret evidence against a null hypothesis. And if a null hypothesis is silly, there will be no skeptics of the alternative, so *P* values are unnecessary.