P values, hypothesis testing, and model selection: it's déjà vu all over again[1]

It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind.
...
And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were in the wrong!

So, oft in theologic wars
The disputants, I ween,
Rail on in utter ignorance
Of what each other mean,
And prate about an Elephant
Not one of them has seen!

—From *The Blind Men and the Elephant: A Hindoo Fable*, by John Godfrey Saxe (1872)

Even if you didn't immediately skip over this page (or the entire Forum in this issue of *Ecology*), you may still be asking yourself, "Haven't I seen this before? Do we really need another Forum on *P* values, hypothesis testing, and model selection?" So please bear with us; this elephant is still in the room. We thank Paul Murtaugh for the reminder and the invited commentators for their varying perspectives on the current shape of statistical testing and inference in ecology.

Those of us who went through graduate school in the 1970s, 1980s, and 1990s remember attempting to coax another 0.001 out of SAS's $P = 0.051$ output (maybe if I just rounded to two decimal places ...), raising a toast to $P = 0.0499$ (and the invention of floating point processors), or desperately searching the back pages of Sokal and Rohlf for a different test that would cross the finish line and satisfy our dissertation committee. The $P = 0.05$ "red line in the sand" partly motivated the ecological Bayesian wars of the late 1990s and the model-selection detente of the early 2000s. The introduction of Markov chain Monte Carlo (MCMC) integration to statistical modeling and inference led many of us to hope that we could capture, or at least model, ecological elephants.

Murtaugh revisits a familiar analysis in which an ecologist is trying to decide how many parameters are needed for a model that provides the "best" fit to a set of observations. For a specific, albeit widespread, case—two or more nested general linear models—*P* values, confidence intervals, and differences in Akaike's information criterion (ΔAIC) are based on identical statistical information and are mathematically interchangeable (this is not the case for non-nested models). Thus, whether one calls it a tree, a snake, or a fan, it's still describing the same elephant. More formally, these methods all provide some measure of the probability or likelihood of the observed data y (and, in some cases, data more extreme than the observed data) *given* a particular model (defined by a set of parameters $\boldsymbol{\theta}$): $P(y \mid \boldsymbol{\theta}) \equiv \mathcal{L}(\boldsymbol{\theta} \mid y)$.

Like John Saxe, we began by asking six individuals to comment on Murtaugh's elephant; we explicitly included the Bayesian perspective with the commentary by Barber and Ogle. We rounded out the forum with Aho et al.'s commentary, which had been submitted concurrently but independently to *Ecological Applications*. Several common themes appear in the submitted commentaries.

The starting point of this safari is an important, but often neglected question: Is the interest in $P(\text{data} \mid \text{model})$ or $P(\text{model} \mid \text{data})$? Murtaugh and the other elephant hunters are explicit that frequentist *P* values quantify the probability of the observed data *and more extreme, but unobserved data* given a specific model: $P(y \geq y_{\text{obs}} \mid \boldsymbol{\theta})$. Further, when calculating a *P* value, the model $\boldsymbol{\theta}$ that is conditioned on is typically the null hypothesis (H_0): a parsimonious sampling model that is rejected easily with real ecological data, especially if sample sizes are large. But as more than one commentary points out, *P* values by themselves *provide no information* on the probability or

[1] Reprints of this 44-page Forum are available for $10 each, either as PDF files or as hard copy. Prepayment is required. Order reprints from the Ecological Society of America, Attention: Reprint Department, 1990 M Street, N.W., Suite 700, Washington, DC 20036 (e-mail: esaHQ@esa.org).

acceptability of the alternative hypothesis or hypotheses. Part of the problem is that ecologists rarely do more than express such alternatives as qualitative statements of expected pattern in the data that simply present alternative hypotheses as trivial negations of the null (e.g., "elephant browsing changes tree density").

In contrast to the fairly straightforward interpretation of a P value associated with a simple null hypothesis, the interpretation of likelihood is less clear. Somewhat like a P value, the likelihood ($\mathcal{L}$) quantifies the probability of data given a model. But $\mathcal{L}$ uses only the observed data, *not* the more extreme but unobserved data: $\mathcal{L}(\theta \mid y_{\text{obs}}) \propto P(y_{\text{obs}} \mid \theta)$. Thus, the choice of whether to use a likelihood or a P value should be, at least in part, determined by one's stance on the "sample-space argument" (see commentaries by de Valpine, and Barber and Ogle). Note also that P values are conveniently scaled between 0 and 1, whereas likelihoods are not probabilities and have no natural scaling. As Murtaugh illustrates, there is a nonlinear negative relationship between a P value and a ΔAIC, and there is no objective cut-point to determine when data significantly depart from the null expectation or when one model should be preferred over another. We don't gain anything by changing from $P \leq 0.05$ to $\Delta\text{AIC} \geq 7$ (or 10 or 14). Burnham and Anderson argue that likelihood-based model selection defines "21st-century science"; we hope this assertion rests on the strength of comparing multiple non-nested models, not simply an exchange of P values for ΔAICs.

Aho et al. identify two world views that clarify the role of inference in interpreting both experimental and observational data. On one hand (Aho et al.'s simulation A), processes giving rise to observed data are complex and poorly understood; replicated experiments to probe these processes would be difficult to devise; sample sizes are unlikely to ever approach the parameter space of the process(es); and we never expect our own models to be the "true" model. On the other hand (simulation B), relatively simple processes give rise to observed data; replicated experiments could be used to test the processes; sample sizes easily can exceed the parameter space of the process; and we expect that at least one of our models is an accurate representation of the underlying process. AIC is appropriate for simulation A; P values, Bayes factors, and Bayesian information criteria (BIC, an asymptotic approximation to the Bayes factor) are appropriate for simulation B. We note that analysis of Big Data—complex processes, surprisingly small sample sizes (e.g., genomes from only a few individuals, but millions of observations [expressed sequence tags] per sample)—falls squarely in simulation A. Yet, as Stanton-Geddes et al. clearly illustrate, even small, relative simple data sets can be interpreted and analyzed in many different ways.

An elephantine wrinkle in Aho et al.'s dichotomy is that P values, ΔAIC, and Bayes factors all suffer from "incoherence" (see commentaries by Lavine, and Barber and Ogle). Given two hypotheses H_1 and H_2, if H_1 implies H_2 then a "coherent" test that rejects H_2 also should always reject H_1. P values, ΔAIC, and Bayes factors all fail to satisfy this criterion; the jury is still out on the coherence of the severity evaluation described by Spanos. Like P values, however, severity violates the likelihood principle by including unobserved data. More informative interpretations of P values, ΔAIC, and severity all depend not only on the data at hand but also on their broader context.

Despite continued disagreements about appropriate use of P values, ΔAIC, and Bayesian posterior probabilities, most of the authors agree that emphasis should be on estimation and evidence, not binary decisions. Most importantly, the mantra to visualize data should be emblazoned on all of our monitors. We have all seen statistically "significant" results explain virtually none of the variation in the data and that are unconvincing when plotted. Fortunately, it is now commonplace to see plots or tables of summary statistics along with significance values. Yet, it is still surprising how often published abstracts fail to report measured effect sizes (as a simple percentage or difference in means) of statistically significant results. Even in the absence of a complex analysis of quantitative model predictions, ecologists can still do a much better job of plotting, reporting, and discussing effects sizes than we have so far.

We also need to remember that "statistics" is an active research discipline, not a static tool-box to be opened once and used repeatedly. Stanton-Geddes et al. clearly illustrate that many ecologists only use methods they learned early in their careers. Such habits of mind need to change! Continual new developments in statistics allow not only for reexamination of existing data sets and conclusions drawn from their analysis, but also for inclusion of new data in drawing more informative scientific inferences. Applying a plurality of methods to more, and better, data is a better way to model an elephant. But don't forget to include its script file with your manuscript!

—Aaron M. Ellison
—Nicholas J. Gotelli
—Brian D. Inouye
—Donald R. Strong
Editors

Key words: *Bayesian inference; hypothesis testing; model selection;* P *value.*

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 $s^2_\text{p} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2/(n_1 + n_2 - 2)$ be the pooled sample variance. The equal-variances t statistic is

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

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

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

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

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^2_p, 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.

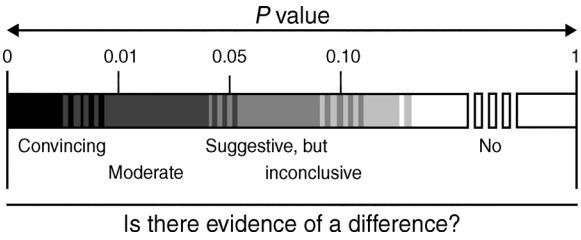

FIG. 1. Interpretation of the *P* value. Reprinted with permission from Ramsey and Schafer (2002).

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

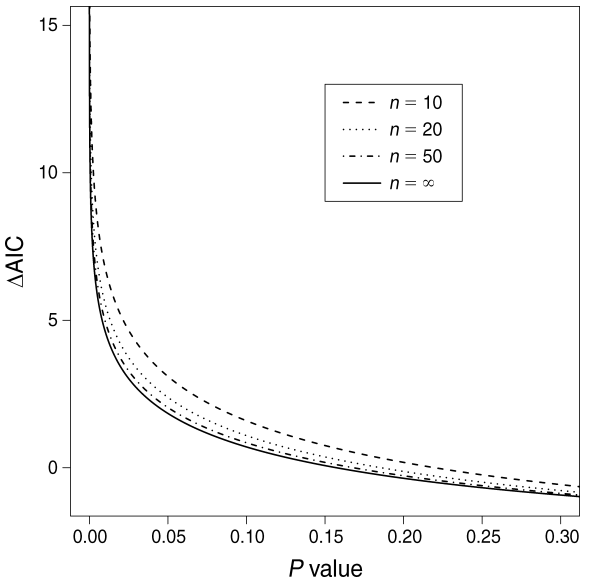

FIG. 2. The relationship between ΔAIC (as defined in Eq. 4) and the *P* value in a comparison of two models differing with respect to one parameter (as in a two-sample comparison, or simple linear regression), for different total sample sizes (*n*). The lines for finite *n* are based on the least-squares 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

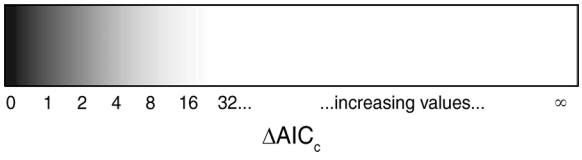

FIG. 3. Interpretation of ΔAIC, from Burnham et al. (2011). "Plausible hypotheses are identified by a narrow region in the continuum where Δ < perhaps four to seven (black and dark gray). The evidence in the light grey area is inconclusive and value judgments for hypotheses in this region are equivocal. 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.

However, other anti-*P*-value arguments are worth taking more seriously, and Murtaugh (2014) does not do them justice. Specifically, it is worth examining what people mean when they argue that *P* values do not measure evidence. Such arguments are the starting point for dismissing the use of *P* values, which motivates the use of information theoretic and Bayesian methods. An excellent, ecologically oriented volume on scientific evidence was edited by Taper and Lele (2004).

I will argue that while *P* values are not a general measure of evidence, they can provide valuable interpretations of evidence. A useful set of distinctions that emerges from the debates about statistical philosophy is among "model fit," "evidence," and "error probabilities" (Lele 2004, Mayo 2004). For purposes here, likelihoods measure "model fit," likelihood ratios compare "evidence" between two models, and *P* values are "error probabilities." Thus, likelihood ratios (to which *F* ratios are closely related) are the central quantity for comparing models, and *P* values are one way to interpret them. Separating these ideas goes a long way towards clarifying a healthy role for *P* values as one type of reasoning about hypotheses. In this light, commonly used *P* values such as for linear models, analysis of variance, generalized linear (and mixed) models, and other likelihood ratio tests all make sense and are philosophically sound, even if they are not the tool one needs for every analysis.

Sample space arguments against P *values*

What are the philosophical arguments against *P* values? Many dismissals of *P* values are built on claims that sample spaces must by force of logic be irrelevant to measuring evidence, from which it follows that P values cannot measure evidence. For example, Murtaugh (2014) quotes Hobbs and Hilborn (2006) on this point, who in turn cited Royall (1997). The sample space of a probability model is the mathematical "space" of all possible data sets. A *P* value is the probability of observing evidence at least as strong against the null hypothesis as the actual evidence. Therefore, a *P* value is a summation (or integration) over probabilities of data that could have been observed but weren't, i.e., a summation over the sample space. If sample space probabilities can have no role in reasoning about hypotheses, then *P* values are useless. This is considered to be an implication of the "likelihood principle," which states that "all the information about [parameters] ... is contained in the likelihood function" (Berger and Wolpert 1988:19). Under the likelihood principle, likelihood ratios alone—but not the corresponding *P* value—compare hypotheses.

Before considering the arguments against sample space probabilities, it is important to see why the *P* value concept *alone* is considered inadequate as a general measure of evidence for comparing hypotheses (Berger and Wolpert 1988, Royall 1997, Lele 2004). Hypothetically, one could construct *P* values (or confidence intervals) that are technically valid but bizarre. For example, one could make a procedure where the confidence interval is calculated in randomly chosen ways yet still rejects the null at the nominal probability level (Royall 1997). Such examples have almost no bearing on statistical practice, but they do establish that the *P* value concept alone is not enough. We need *P* values based on a good measure of evidence (or test statistic), generally likelihood ratios, which illustrates why it is useful to separate "evidence" (likelihood ratio) from "error probabilities" (e.g., *P* values).

Now we are ready to consider the arguments against sample space probabilities. Two common types of hypothetical examples are those for stopping rules and those for multiple testing (Berger and Wolpert 1988, Royall 1997). In a stopping-rule example, one considers data collected under two protocols. Either the sample size is pre-determined, or intermediate analyses of some of the data may be used to decide whether to continue the study, i.e., there is a "stopping rule." Now suppose that at the end of each study, each protocol happens to have generated exactly the same data. Then, it is asserted, it makes no sense to reach different conclusions about various hypotheses.

However, if one uses *P* values, the second study must be viewed as a stochastic process in which the sample space includes different times at which the study might have been stopped based on intermediate results. Therefore, the probability of obtaining evidence at least as strong against a null hypothesis involves a different sample space probability for the two studies. This violates the likelihood principle by considering not just the final data but also the decisions used to obtain the data. To be sure, some thought experiments are less blatant, or more bizarre, than this one, but they all involve the same data generated by different protocols (Royall 1997).

The other type of argument against sample-space probabilities involves multiple testing. Again one compares two people who obtain the same data in different ways (Royall 1997). Say both are studying human traits associated with myopia (near-sightedness), and both use a sample of 100 people. In addition to measuring myopia, one person measures only birth month and the other measures birth month and 999 other variables. If both obtain the same data for birth month and myopia, they should have the same evidence about that specific relationship. It should not matter that the second person measured other variables, but that is exactly what a multiple-testing correction (e.g., Bonferroni) would enforce.

The flaw in both types of examples is that they dismiss the possibility that how a study is conducted really can impact the probability of obtaining spurious evidence. In the stopping-rule example, if you tell experimentalists they are allowed to check their data at every step and stop when they have a result they like, some really would

do so, and that protocol really would shape the probabilities of spurious outcomes. In the multiple-testing example, if you tell someone they can inspect 1000 relationships separately and then write headline news about whichever one has the strongest evidence while ignoring the fact that they started with 1000 variables, they have a much higher chance of reporting a spurious result than does someone who looked at only one relationship. Mayo (2004) argues that this disconnect between philosophy and practice arose because many philosophers of science begin their reasoning once the data are in hand, and the process of obtaining data may seem irrelevant.

The situation is not helped by the rhetoric of irritation: It seems plainly ridiculous that if you "peek" at your data, you are automatically invalidating your results, or that if you decide to measure some extra variables in addition to your primary variable (e.g., birth month) you must change you interpretations about birth month (Royall 1997). But those misconstrue the logic of *P* values: if you merely peek and then proceed, or if you treat birth month as an a priori hypothesis no matter what else you measure, then you have caused no harm. But if you stop when you decide your data look good, or if you study many variables and report whichever has the strongest evidence, you are de facto changing the probability of obtaining a strong result. In summary, the anti-sample-space arguments against *P* values ignore the fact that how data are collected can influence the probabilities of the final data in hand.

Another problem with the multiple-testing argument is that it pretends one must control the family-wise error rate (e.g., Bonferroni correction). However, alternatives include presenting each result separately and letting the reader evaluate them or presenting a false discovery rate. In other words, disliking Bonferroni corrections is unrelated to whether *P* value logic is sound.

Model fit, evidence, and error probabilities

Making the distinctions between model fit, evidence, and error probabilities can unwind a lot of tension in the above debates, or at least focus them more narrowly. In both cases above, if the likelihood ratio (comparison of model fits) represents "evidence," then the likelihood principle is satisfied. This relies upon all the statistical foundations about why likelihoods are so useful (Lele 2004). Then, as an aid to deciding what to believe from the evidence, the error probabilities represented by *P* values are one useful type of reasoning. Indeed, even Berger and Wolpert (1988), in their seminal anti-*P*-value work, state that "most classical procedures work very well much of the time," even if they would disagree about why. Such classical procedures would seem to include *P* values and confidence intervals based on likelihood ratios or the closely related *F* ratios, i.e., the vast majority of *P* values that ecologists generate.

Using these distinctions, the two investigators in the examples above might have the same evidence, but different error probabilities. One might have obtained the evidence by a procedure more likely than the other to generate spurious results, such as a stopping rule or multiple testing. In these distinctions, the usage of "evidence" is construed in a narrow sense to mean "likelihood ratio." It may be confusing that a broad dictionary definition of "evidence" would include anything used to decide what to believe, which could encompass *P* values based on likelihood ratios.

These distinctions also clarify that likelihoods are more fundamental, and *P* values are one way to use likelihoods. Indeed, Murtaugh (2014) is implicitly talking not about any kind of hypothetical *P* values but rather about *P* values *from likelihood ratios* (or *F* ratios). In this case, the two have a monotonic relationship, and one can see why he chooses to speak of *P* values directly as "evidence." However, after so much philosophical debate has gone into separating the two concepts, it strikes me as confusing to try to put them back together. In essence, Murtaugh (2014) is using the broad sense of "evidence," but statisticians have gone to great lengths to posit the narrow sense of "evidence" to mean likelihood ratios. By discussing *P* values based on likelihood ratios, Murtaugh has blurred this distinction, although his fundamental points remain sound.

Unfortunately an occasional rhetorical follow-on to the dismissal of sample space probabilities is to describe them derisively as probabilities of unobserved data. This makes *P* values sound contrary to common sense because rather than focusing on the data actually observed, they consider possible unobserved data, which sounds foolish. This is misleading. If one is using a probability model for the data, which is the basis for likelihoods in the first place, then part and parcel of that model are the probabilities associated with unobserved data. A likelihood calculation *would not exist* if the model didn't describe a distribution for other hypothetical data. Using sample space probabilities is not on its face ridiculous.

Another problem with the arguments against *P* values is to treat their validity as akin to a mathematical theorem: it must be universally either true or false. But there is no reason to dismiss a principle that works in some situations and not others; doing so should violate ecologists' healthy sense of pragmatism. The types of hypotheticals used for these philosophical debates typically have no bearing on, say, finding the *P* value of one parameter in a model for simply collected data. Indeed, most anti-sample-space thought experiments do not consider the common situation of nested models, and hence don't address the fact that a "bigger" model will always fit the data better and so we often need to apply skeptical reasoning (Forster and Sober 2004).

A final argument against *P* values as evidence is tied to the relationship between *P* values and accept/reject hypothesis testing. In a nutshell, the statement "we found no evidence ($P = 0.06$)" appears to offend common sense. If $P = 0.06$, there is certainly evidence

against the null, and to state otherwise sounds Orwellian. This is not a flaw with *P* values, but rather with presentation customs. I interpret "we found no evidence ($P = 0.06$)" to be shorthand for a more elaborate explanation that the evidence was not strong enough to convince a skeptic and hence no claim about it will be attempted. Unfortunately this has created a feeling that *P* value *practices* enforce ridiculous statements even though the principles are sound. Again, distinguishing the concepts of evidence and error probabilities clarifies the issue.

In summary, the philosophical arguments to dismiss *P* value logic fail to appreciate its role in skeptical reasoning about serious hypotheses. Skepticism is fundamental to science, and *P* values are fundamental to skeptical reasoning.

Bayesian methods and P *values*

Murtaugh (2014) does not go far into Bayesian arguments against *P* values, but there is much to say there. It is common for a pro-Bayesian argument to begin with a rejection of frequentism as represented by sample space probabilities and *P* values. However, there are also "Bayesian *P* values" and credible intervals (as Murtaugh points out, confidence intervals come from a continuous application of *P* values, so without *P* values there would be no confidence intervals), so one must ask if these are really more meaningful than their frequentist counterparts. For example, a common pro-Bayesian argument is that frequentist confidence intervals are approximate while Bayesian credible intervals are exact. I will argue next that this is a misleading comparison.

The reason a confidence interval is approximate is that the procedure is trying to do something objective but hard: cover the correct parameter value in 95% of data sets. In a Bayesian analysis, a 95% credible interval is defined with "degree of belief" probability, so it is meaningless to call it "accurate" vs. "approximate." Sure, it is "exact" in the sense of a correct execution of obtaining the posterior distribution, but we could just as well say the frequentist confidence interval is exact because we obtained it with mathematically correct profile likelihood calculations. So the pro-Bayesian argument amounts to saying "we prefer an exact calculation of something subjective to an approximation of something objective."

Let us make a more stark example. Suppose a weather forecaster predicts a 10% probability of rain tomorrow. What should that mean? The frequentist answer is that 10% means "1 out of 10 times on average": out of many "10%" predictions, it should have actually rained 10% of the time. This could be tested with data. The Bayesian definition of 10% *probability* of rain tomorrow is 10% *degree-of-belief* in rain, which can mean whatever the forecaster wants it to. If you collect data and determine that "10% degree-of-belief" corresponds to rain 20% of the time, you would have done nothing to change what the forecaster wants 10% to mean, nor could you. Are you happier with an "exact" 10% degree-of-belief that can mean anything or an "approximate" 10% frequentist probability that aims for objectivity?

If this example makes you uncomfortable, you are in good company. In a seminal paper, Rubin (1984) argued that to be scientists, we must have objective criteria that allow the possibility of rejecting a model (or hypothesis) from data. Since a pure Bayesian outlook does not provide this possibility, we should seek Bayesian results that are "calibrated" to frequentist probability. In other words, Bayesians still need frequentism. In practice, many Bayesian results will be *approximately* calibrated to frequentist interpretations because of asymptotic likelihood theory. In summary, the need for skeptical reasoning in science leads to *P* values, and the objectivity desired in such reasoning should make one uncomfortable with a pure Bayesian stance.

Model selection and P *values*

Murtaugh's (2014) emphasis on the relationship between AIC model selection and *P* values is good medicine for the way they are often viewed in opposition to each other. However, they are derived to solve different problems: AIC is for finding the best model for predicting new data, and for all one knew it might have led somewhere unrelated to *P* values. They turn out to be closely related (for nested models), which is what Murtaugh emphasizes, but the different problems they solve make the use of one vs. the other more than a "stylistic" difference. A more subtle point is that the derivation of AIC is valid even when the models are not nested and do not include the "correct" model (Burnham and Anderson 2002), while likelihood ratio tests rely upon nested, well-chosen models.

Murtaugh highlighted that some conventions for interpreting AIC differences, such as thresholds of 4 and 7, would be conservative by hypothesis-testing standards. On the other hand, I think there is a temptation to over-interpret the AIC winner between two models, which is a liberal threshold by hypothesis testing standards. If model F has one more parameter than model R, it must have a maximum log likelihood more than 1.0 higher than R in order to have a better AIC (Murtaugh 2014: Eq. 4). The *P* value from the corresponding likelihood ratio test is 0.16 (Murtaugh 2014: Eq. 5; $\Delta\text{AIC} = 0$ in Murtaugh 2014: Fig. 2), liberal rather than conservative. An amazing outcome of the AIC derivation is that it actually puts meaning on this particular *P* value threshold. (There is a deeper reason to be careful mixing hypothesis tests with model selection, which is that a test comparing the top two models from a larger set does not incorporate the stochasticity of the AIC ranks.).

History, headaches, and P *values*

I suspect that many ecologists are happy to see *P* values get philosophically whacked because they are

such a bloody pain. Long ago, ecology was a science of storytelling. It was an uphill battle to infuse hypothesis testing into the science, but eventually it became the gold standard for presenting results. Unfortunately, that meant that authors and journals over-emphasized *P* values and under-emphasized effect size, biological significance, and statistical prediction. Against that background, those who pushed forward model selection, model averaging, and Bayesian methods faced another uphill battle to broaden the scope of ecological statistics beyond *P* values. As a result, *P* values have sometimes been bashed rather than put in healthy perspective.

In complicated situations, simply obtaining a valid *P* value (or confidence interval) can be so difficult that sometimes practice is confused with principle. It would be nice if we could dismiss them as unimportant when they are impractical. Ironically, however, some of the methods that arose behind anti-*P*-value arguments are useful for obtaining better *P* values for difficult problems. For example, one use of model-averaging with AIC weights is to obtain more accurate *P* values and confidence intervals by incorporating uncertainty due to model selection. Similarly, there are claims that sometimes Bayesian procedures can provide more accurate frequentist coverage.

Suppose you are backed up against a wall by a gang of wild raving pure scientists, and they are forcing you to *decide what you believe*. Under such duress, objective statements about error probabilities can provide one useful line of reasoning for Bayesians and frequentists alike. What would science be if one can't argue objectively that someone else's claims are statistically spurious?

Acknowledgments

I thank Daniel Turek and Subhash Lele for helpful comments.

Literature Cited

Berger, J. O., and R. L. Wolpert. 1988. The likelihood principle: a review, generalizations, and statistical implications. Second edition. IMS lecture notes. Monograph series, volume 6. Institute of Mathematical Statistics, Hayward, California, USA.

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

Forster, M., and E. Sober. 2004. Why likelihood? Pages 153–165 *in* M. L. Taper and S. R. Lele, editors. The nature of scientific evidence. University of Chicago Press, Chicago, Illinois, USA.

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

Lele, S. R. 2004. Evidence functions and the optimality of the law of likelihood. Pages 191–216 *in* M. L. Taper and S. R. Lele, editors. The nature of scientific evidence. University of Chicago Press, Chicago, Illinois, USA.

Mayo, D. G. 2004. An error-statistical philosophy of evidence. Pages 79–97 *in* M. L. Taper and S. R. Lele, editors. The nature of scientific evidence. University of Chicago Press, Chicago, Illinois, USA.

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

Royall, R. M. 1997. Statistical evidence: a likelihood paradigm. Chapman and Hall, New York, New York, USA.

Rubin, D. B. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Annals of Statistics 12(4):1151–1172.

Taper, M. L., and S. R. Lele. 2004. The nature of scientific evidence. University of Chicago Press, Chicago, Illinois, USA.

Ecology, 95(3), 2014, pp. 621–626

To *P* or not to *P*?

Jarrett J. Barber[1,3] and Kiona Ogle[2]

[1]*School of Mathematical and Statistical Sciences, Arizona State University, Tempe, Arizona 85287-1804 USA*
[2]*School of Life Sciences, Arizona State University, Tempe, Arizona 85287-6505 USA*

Introduction

We appreciate Murtaugh's (2014) very readable defense of *P* values. Murtaugh argues that most of the criticisms of *P* values arise more from misunderstanding or misinterpretation than from intrinsic shortcomings of *P* values. After an introductory musing on a familiar definition of the *P* value, we discuss what appears to be an "intrinsic shortcoming" of the *P* value, rather than a misunderstanding or misinterpretation; the *P* value lacks what might be considered a very reasonable property to be desired in measures of evidence of hypotheses (Schervish 1996). Then, we attempt to provide a sense of the misunderstanding of *P* values as posterior probabilities of hypotheses (Murtaugh's fifth criticism) or as error probabilities; *P* values can often be much

Manuscript received 22 July 2013; revised 23 August 2013; accepted 26 August 2013. Corresponding Editor: A. M. Ellison. For reprints of this Forum, see footnote 1, p. 609.

[3] E-mail: Jarrett.Barber@asu.edu

lower than the probability of the null hypothesis being true, a result commonly referred to as Lindley's Paradox (Lindley 1957, Jeffreys 1961), which suggests that typical "significant" P values (e.g., 0.001–0.1) may not be so significant. We also present a "calibrated P value" (Sellke et al. 2001), as a sort of reconciliation of the P value and Bayesian posterior probability, to be reported instead of, or in addition to, the P value; our discussion touches on the related alternatives of *conditional* (Type I or Type II) error probabilities. In addition, we briefly broaden the discussion to view the P value as a factor contributing to lower than expected rates at which ecological and other reported scientific studies are reproduced. Finally, we summarize and provide a final remark on interpreting P values in the context of their variability.

Our intent is to provide the reader with a better sense of the P value controversy by presenting some of its underlying details, most of which may be found in Murtaugh's references or related literature. (However, the coherence property [Gabriel 1969, Schervish 1996, Lavine and Schervish 1999], which we discuss in *An intrinsic shortcoming*, seems relatively underrepresented among the criticisms and may be new to many readers.) Our presentation is not necessarily directed at Murtaugh's focus on the P value from the likelihood ratio test of nested hypotheses. Indeed, we suspect that the use of the likelihood ratio test statistic ameliorates some P value criticisms. Like Murtaugh, we largely refrain from discussing the degree to which general criticisms may or may not apply to the particular case of the likelihood ratio test statistic or to particular ecological examples.

On the Definition of the P Value

As provided by Murtaugh, the P value is the probability of obtaining a result, i.e., data or test statistic, at least as extreme as the observed result, assuming that the null hypothesis is true; see Murtaugh's Eqs. 1 and 2. Thus, the P value is computed not only using the *observed* result, but also all *unobserved* results more extreme than the observed result. It seems sensible to us to include the observed result in a measure of evidence. But, to include all more extreme results that are somehow hypothetically observable, but are actually unobserved, seems unusual to us. Our intuition suggests that they would somehow lead to bias against the null, a notion that we discuss more in *Understanding* P *values*. This oddity is echoed in a comment by Harold Jeffreys, a physicist and well-known Bayesian statistician, who said, "An hypothesis, that may be true, may be rejected because it has not predicted observable results that have not occurred" (Jeffreys 1961). See Berger and Wolpert (1988: Section 4.4) for examples of what can go wrong with inference when we include more extreme values when using P values. Incidentally, a strict Bayesian approach conditions on the observed data to avoid problems arising from unobserved results.

Further, we believe it is worthwhile to note that, as in Murtaugh's paper, the P value is often stressed to be a measure of evidence against the null (though the P value formed from the likelihood ratio test statistic does reflect a particular alternative). But, this careful wording is not necessarily required for measures of evidence. In particular, if we use the probability of hypotheses, as in a Bayesian approach to testing, then we may speak freely of probability as being a measure for or against a hypothesis. For example, a probability of 0.05 for/against a hypothesis gives a probability of 0.95 against/for it, whereas a P value of 0.05 against the null does not imply a measure of evidence of 0.95 for an alternative.

Aside from not having a definition or set of underlying principles for "measures of evidence," the P value as evidence against the null is often an appropriate interpretation in the sense that the P value is often independent of an alternative hypothesis. The P value often provides only a measure of how well the observed data fit the null hypothesis; often no connection to an alternative exists. This seems useful for comparing data sets relative to a single hypothesis, but, alas, we typically have only one data set and multiple hypotheses. In the context of the likelihood ratio, however, which is Murtaugh's focus, an alternative hypothesis is "built in," so that the P value for the likelihood ratio test statistic generally depends on an alternative hypothesis, which seems appropriate to us if we wish to compare hypotheses.

An Intrinsic Shortcoming

Murtaugh attributes criticism of P values mainly to misunderstanding or misinterpretation, not mainly to inherent shortcomings. We find it difficult to know what is meant by "shortcoming" without a formal set of principles or desired properties by which to evaluate measures of evidence. The property of coherence, however, seems to us a compelling property, which is not possessed by the P value.

Suppose that one hypothesis implies another, e.g., $H_1 : \theta \in (-\infty, 0]$ implies $H_2 : \theta \in (-\infty, 10]$ because $(-\infty, 0]$ is a subset of $(-\infty, 10]$. Then, we say that tests of H_1 and H_2 are *coherent* if rejection of H_2 always entails rejection of H_1 (Gabriel 1969), and we say that a measure of support for hypotheses is coherent if, whenever H_1 implies H_2, the measure of support for H_2 is at least as great as that for H_1 (Schervish 1996). For example, a coherent measure's support for $H_2 : \theta \in (-\infty, 10]$ is at least as large as its support for $H_1 : \theta \in (-\infty, 0]$.

Schervish (1996) shows that P values are incoherent for point hypotheses, for one–sided hypotheses and for hypotheses of the bounded interval type. As a simple example of the latter type, assume $X \sim N(\theta, 1)$ (e.g., view X as a difference of averages from two samples that has been scaled to have unit variance for simplicity, and view θ as a difference of means), and consider the hypotheses, $H_1 : \theta \in [-0.5, 0.5]$ and $H_2 : \theta \in [-0.82,$

0.52]. For $X = 2.18$, Schervish (1996) shows that the P value of the first hypothesis is 0.0502 and that of the second is 0.0498, thus, the P value indicates more support for $H_1 : \theta \in [-0.5, 0.5]$ than for $H_2 : \theta \in [-0.82, 0.52]$ despite $[-0.5, 0.5]$ being a subset of $[-0.82, 0.52]$. In other words, if we were to test each hypothesis at an error rate of $\alpha = 0.05$ (perhaps using a Bonferroni adjustment to bound the overall rate at or below 0.1), then we would reject that θ is in $[-0.82, 0.52]$ but would not reject that θ is in $[-0.5, 0.5]$! Thus, inasmuch as coherency is a property of measures of support or evidence for hypotheses, then P values are invalid measures of support. Note that the insignificance of significance in this example—0.0498 vs. 0.0502—may lessen concern for coherence, but we still see coherence as a compelling property to require of measures of evidence of hypotheses; see Murtaugh's first criticism and our comments on the variability of P values (Boos and Stefansky 2011) in *Closing remarks*.

For those who believe this example is artificial because ecologists rarely use interval hypotheses, we note that the P value for the null hypothesis, H: $\theta = 0.5$, is 0.0930, larger than for H_1 and H_2, despite H being contained in both H_1 and H_2. Also, the P value for the one-sided null, $H : \theta \leq 0.5$, we know, is half that of $H : \theta = 0.5$. Further, we argue that ecologists may want to use interval hypotheses, which is suggested by Murtaugh in his fourth criticism, that the null hypothesis is usually false; he suggests a test for a "small, marginally meaningful association," which we argue could be performed by specifying a null consisting of an interval of values that are not considered ecologically meaningful or important.

Schervish (1996) asks, what *do* P values measure? If we look at the P value as a function of the data given a single, fixed hypothesis, then we may interpret the P values for different data sets as indicating different degrees to which the different data sets support the hypothesis. But, then, it appears that we cannot acknowledge other hypotheses without concluding that the P values for the different hypotheses are on different scales of support. For example, similar to the previous examples in this section, Schervish (1996) computes P values of 0.0718 for $H_1 : \theta = 0.9$ and 0.0446 for $H_2 : \theta \leq 1$ based on an observation of $X = 2.7$ from $X \sim N(\theta, 1)$. Because we are compelled to think that 2.7 must indicate more support for $H_2 : \theta \leq 1$ than $H_1 : \theta = 0.9$, we are thus compelled to think that the P values must be on different scales, that 0.0718 must indicate lower support for $H_1 : \theta = 0.9$ than 0.0446 does for $H_2 : \theta \leq 1$; the P value is not a coherent measure of support for hypotheses.

If these arguments are compelling in simple examples of point, one-sided, and interval hypotheses in one dimension (Schervish 1996), as discussed here, we may also expect them to be compelling in more typical ecological analyses.

Understanding P Values

As a way to help us understand P values, consider the following simulation (Sellke et al. [2001] and the applet cited within). Consider $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$, where θ is, say, the difference between mean responses under two different treatments. (We assume a difference of zero is a plausible hypothesis or is a plausible approximation to an interval about zero.) Randomly generate several "data" sets, each of size n, from either $N(0,1)$, the null distribution with (mean difference) $\theta = 0$, or from an alternative, $N(\theta, 1)$, $\theta \neq 0$. For each data set, whether from the null or an alternative, we compute a familiar test statistic $z = (\bar{x} - 0)/(1/\sqrt{n}) = \sqrt{n}\bar{x}$, where $\bar{x}$ is the sample average of n data values, to obtain several test statistics generated under H_0 and several under H_1, depending on the proportion of times we generate a data set from each hypothesis. Then, for each result, z, that gives a P value of about 0.05 (say between 0.049 and 0.051), we record the result's hypothesis of origin, H_0 or H_1, thus computing the proportion of results that are associated with the null or the alternative, given that the result is associated with a P value of about 0.05. For what proportion of results, having P value of 0.05, is the null hypothesis true?

Before obtaining an answer from our simulation, we must first decide what proportion of data sets to generate under each of H_0 and H_1, and, further, how to choose the θ values for generating under H_1. Attempting impartiality, we choose 50% of the θ values to be zero, under H_0, and 50% from H_1. We may simply pick a single value of θ in H_1, so that the alternative becomes a point hypothesis, or we may somehow spread out the values of θ under the alternative hypothesis. (We found results to be remarkably consistent over a wide variety of ways of distributing θ values over H_1.) For the sake of proceeding, we choose to sample the alternative θ values from a uniform$(-3, 3)$ (with zero probability of being exactly zero). Finally, we choose $n = 20$ per simulated data set.

We obtained 2000 simulations with P values between 0.049 and 0.051. Of these, 1221, or about 61%, were associated with the null. Roughly speaking, for our 50/50 choice with uniform spreading of θ over H_1, the probability of the null being true is about 0.61, despite the P value being 0.05. In other words, given that we reject the null hypothesis when observing $P = 0.05$, we do so falsely about 61% of the time. So, we should not view the P value as the probability of a Type I error after (or before) conditioning on the observed $P = 0.05$, even though it is equal, in this example, to the pre-experimental (before looking at the data) probability of a Type I error of 0.05. It turns out that, for point null hypotheses (Sellke et al. 2001), as considered in this simulation, the proportion of nulls is generally and remarkably higher than the P value.

The reader may have noticed that we can interpret this simulation from a Bayesian point of view. We can see the prior probability of 1/2 given to each hypothesis,

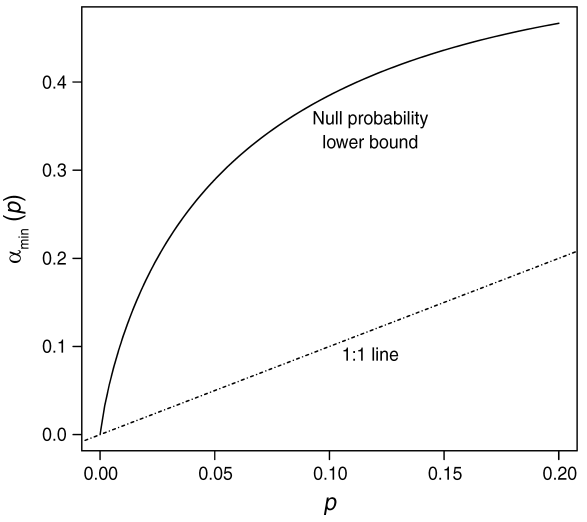

FIG. 1. The parameter $\alpha_{min}(p)$, solid line, is the lower bound of the probability of a true null as a function of the P value, p.

with the 1/2 probability mass on the alternative being spread according to the uniform(−3, 3) prior, and, given data that result in a P value of 0.05, the (posterior) probability of the null is remarkably greater than the P value. But, we do not need to adopt a Bayesian perspective to see how the simulation illustrates Murtaugh's fifth criticism, which warns us not to interpret the P value as the probability of the null being true. Incidentally, this large disparity between the Bayesian probability of a point null being true and the P value is often referred to as Lindley's Paradox (Lindley 1957, Jeffreys 1961); the disparity is not nearly as remarkable with one-sided alternatives (Casella and Berger 1987).

Upon further consideration, we might object to calling this a paradox. As usual, we are to interpret the P value as saying roughly that H_0 is not a compelling choice for the observed data (given a P value of about 0.05). That the Bayesian result is so different may not be surprising when we consider that our uniform prior on H_1 puts substantial support on many values of θ that are a very poor explanation of the data: pick a value of θ in H_0 or H_1, then H_0 can be no farther than three from the selected value, but potentially many values in H_1 can be a distance of three to six from the selected value. Thus, we might expect the Bayesian approach to favor H_0 over H_1 compared to the P value: no paradox.

This suggests the questions: Does posterior probability tend to favor H_0 too much? Could the large disparity observed in our simulation be a fluke due to our particular choice of how to distribute θ values in H_1? What is the *least* support for H_0 that we can have in terms of posterior probability? That is, given an observed P value, how small can the probability of H_0 be, and is the P value still far from this smallest posterior probability? Sellke et al. (2001) address this last question for point nulls in a theoretical extension of the simulation just performed. Loosely speaking, they find the least favorable way (prior) to distribute θ values over H_1, with mass 1/2, as in our simulation, to arrive at a lower bound on the posterior probability of H_0. Essentially, no matter how we spread values of θ values over H_1, the probability of H_0 can be no smaller than the lower bound given by Sellke et al. (2001).

Generally, our simulation results will differ depending on the P value, and, relatedly, the theoretically based lower bound of Sellke et al. (2001) is a function of the P value (p)

$$\alpha_{\min}(p) = \frac{1}{1 - 1/(e \times p \log(p))}$$

where $p < e^{-1}$ and $e = \exp(1) \approx 2.7$; see Fig. 1. For our simulation, we chose P value ≈ 0.05, thus $\alpha_{\min}(0.05) = 0.289$, which is much smaller than our simulation result of 0.61 for the probability of the null. Evidently, the uniform prior that we chose in our simulation is not least favorable to the null. Still, the disparity between 0.05 and 0.289, or more generally between the 1:1 line and the lower bound line in Fig. 1, shows a substantial disparity between P values and posterior probabilities for the point null case. In other words, as in our simulation, given that we reject the null when observing P value $= p$, we see that we do so in error with probability at least $\alpha_{\min}(p)$. So, again, we see, now theoretically, that the P value is not the probability of a Type I error after (or before) conditioning on the observed P value $= p$, and we can interpret $\alpha_{\min}(p)$ as the minimum conditional Type I error given the observed P value $= p$ (hence the suggestive notation).

If the lower bound now seems too harshly biased against the null—after all, it is least favorable to the null—we might consider the conditional Type I error probability, call it $\alpha(p)$, instead of its lower bound, $\alpha_{\min}(p)$. That is, $\alpha(p)$ is a conditional version of the ordinary frequentist probability of a Type I error, now conditioned on P value $= p$, as we alluded to above. Moreover, this conditional frequentist error rate is precisely a posterior probability of the null for some prior in a large class of priors. In other words, if we use the conditional Type I error probability, $\alpha(p)$, or its lower bound, $\alpha_{\min}(p)$, there is no risk of misinterpreting these as probabilities of the null; Bayesian and frequentist results coincide. However, while $\alpha_{\min}(p)$ is simple to compute, and, of course, if we have a prior, we can compute posterior probabilities, it is not otherwise clear which priors to use for $\alpha(p)$ in various testing scenarios; conditional testing requires more development (Berger 2003).

We have emphasized that the P value is not the probability of the null, as in Murtaugh's fifth criticism, and it is not an error probability. We should mention again that such disparities between the P value and posterior probability of the null are much smaller when considering one-sided alternatives, e.g., $H_0 : \theta \leq 0$, $H_1 : \theta > 0$, instead of the point null case (Sellke et al.

2001). Casella and Berger (1987) show that, for a large class of priors in the one-sided case, P values are *greater* than the smallest possible posterior probability of the null and, in some cases, equal to the smallest possible such probability over a large class of priors. (Some readers may recall, for example, that, for the typical one or two sample one–sided z or t test, with a certain improper prior, the P value and probability of the null coincide exactly.) In other words, for the one-sided case, P values are much more consistent with posterior probabilities of the null. Casella and Berger (1987) give expressions for the lower bound of the probability of the null for selected distributions in the one-sided case.

We end this section with a brief extension of our discussion to reproducibility of studies reported in the literature. Green and Elgersma (2010) randomly sampled P values and sample sizes reported in the journal *Ecology* in 2009 and used these to simulate probabilities of nulls and alternatives, assuming both the null and alternative were equally likely before observing the data, just as in our discussion above. A plot of their estimated posterior probabilities of the null vs. P values (not shown) reveals a lower bound similar to that in Fig. 1. Overall, they found that the null hypothesis was more probable than the alternative in about 10% of cases having P values less than 0.05. In addition, the average P value was 0.010 ± 0.017 (mean $\pm$ SD) while the average posterior probability was 0.117 ± 0.166, about 12 times higher. While Green and Elgersma (2010) focus on illustrating Lindley's Paradox and on offering a word of caution about the interpretation of P values, we can also look at their results, and our discussion above, from the perspective of the rate of reproducibility of studies in ecology and in science in general.

Inasmuch as we expect studies to be reproducible at a high (perhaps 95%) rate, we now see how use of the P value may contribute to actual rates that are lower than expected. (Noting possible misunderstandings of P values (as an error rate) and urging a lower expectation of reproducible results seems too simplistic a response in light of the seriousness of the issue and the aforementioned P value shortcomings.) Of course, P values are only one of several factors that may contribute to lower than expected rates of reproducibility. For example, publication bias, the tendency to publish only "positive" results, is another important factor. If a large proportion of results are never published because they are negative, then we can see how publication bias would also tend to lower the rate at which reported results can be reproduced. Some medical journals have responded by refusing to publish certain studies without registration at the outset: "If all the studies had been registered from the start, doctors would have learned that the positive data were only a fraction of the total" (Washington Post 2004). The old joke about the fictitious "Journal of Insignificant Results" containing 95% of all results seems to contain an element of seriousness. Incidentally, Fanelli (2012) estimates about 75% of articles in the discipline of "Environment/Ecology" report results that support the hypothesis tested, with an increasing trend over time; results are remarkably consistent across other disciplines and across locations.

Low reproducibility rates have led to relatively extreme criticism in the popular press by Matthews (1998), "The plain fact is that in 1925 Ronald Fisher gave scientists a mathematical machine for turning baloney into breakthroughs, and flukes into funding. It is time to pull the plug."

Closing Remarks

We have seen that the P value is incoherent, producing results that challenge reasonable expectations for a measure of evidence among hypotheses. We may view the P value as a function of the data given a single, fixed hypothesis and interpret the P values for different data sets as indicating different degrees to which each data set supports the single hypothesis. But, we typically deal with a single data set, and, as Sellke et al. (2001) remark, knowing that the data are rare under the null is of little use unless one determines whether or not they are also rare under the alternative, and we are led to consider posterior probabilities for comparing hypotheses. This remark by Sellke et al. (2001) seems to have an analog in the context of reproducibility: to the extent that we only know about the studies under which the null is found to be rare may not only be useless but potentially may be detrimental without also knowing the balance of studies under which the alternative is found to be rare. Thus, we seem to present a challenge to journal editors seeking high impact factors and to authors seeking recognition for "significant" work.

Our section, *Understanding P values*, suggests that we use probability of hypotheses to compare hypotheses. The most straightforward way to use probability is to adopt a Bayesian approach to compute posterior probabilities of hypotheses. The conditional frequentist Type I error, $\alpha(p)$, corresponds to the probability of a hypothesis for some prior, but conditional frequentist approaches remain largely under development. The lower bound, $\alpha_{\min}(p)$, offers a quick frequentist (and Bayesian) answer. Thus, in principle, those who are philosophically opposed to a Bayesian approach may retain their opposition by using a conditional frequentist approach while getting an answer that is equivalent (in value, not philosophy) to a Bayesian answer. Incidentally, Sellke et al. (2001) call $\alpha(p)$ and $\alpha_{\min}(p)$ "calibrated P values," perhaps due to being adjustments of the P value "toward the observed data" or at least to reflect the observed value of the P value, which is a function of the observed data.

There is a connection to Murtaugh's likelihood ratio test statistic. To see this, note that we may use posterior odds in an equivalent fashion to the posterior probability in the sense that each is a monotonic function of the other, i.e., hypotheses are ordered in the same manner using either probability or odds as evidence. Further, the

"impartial" assignment of equal prior probabilities to hypotheses, as was the focus above, results in prior odds of hypotheses of one, and the Bayes factor, which is a multiplicative adjustment of prior odds to obtain posterior odds, becomes equivalent to posterior odds. In the case of point hypotheses, the "impartial" Bayes factor is exactly the same as the likelihood ratio. In the case of composite hypotheses (e.g., intervals or half intervals), we must address the ambiguity in the parameter: it is no longer determined by point hypotheses. The likelihood ratio test statistic approach tells us to plug in the maximum likelihood estimate for the parameter in each of two likelihoods restricted to their respective hypotheses. The Bayes factor becomes the ratio of averaged likelihoods with respect to the prior restricted to their respective hypotheses. So, we might expect that using the likelihood ratio test statistic to perform similarly to posterior probabilities for ordering hypotheses.

However we interpret or use *P* values, we should realize that *P* values are functions of our data, and, as such, *P* values have distributions, a fact that we suspect is considered by few people, statisticians or otherwise; the *P* value is almost invariably reported only as a fixed value, and we almost never hear of it being considered as a realization from some distribution. Boos and Stefansky (2011) suggest that, often, the level of variability in the *P* value warrants that only its order of magnitude (e.g., 0.1, 0.01, 0.001, or *, **, ***) indicates any meaningful differences in significance; higher precision is often lost within the variability of the *P* value. This suggests that the word descriptions of *P* values in Murtaugh's Fig. 1 are about as precise as we should consider the *P* value to be.

Acknowledgment

Much of the motivation for this commentary was provided by a seminar, by Jim Berger, entitled *Reproducibility of Science: P-values and Multiplicity*, 4 October 2012, for the Section on Bayesian Statistical Science (SBSS) of the American Statistical Association (http://www.amstat.org/sections/sbss/webinarfiles/berger-webinar10-2-2012.pdf).

Literature Cited

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

Berger, J. O., and R. L. Wolpert. 1988. The likelihood principle. Second edition. Volume 6 of Lecture notes—monograph series. Institute of Mathematical Statistics, Hayward, California, USA.

Boos, D. D., and L. A. Stefansky. 2011. P-value precision and reproducibility. American Statistician 65(4):213–221.

Casella, G., and R. L. Berger. 1987. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. Journal of the American Statistical Association 82(397): 106–111.

Fanelli, D. 2012. Negative results are disappearing from most disciplines and countries. Scientometrics 90(3):891–904.

Gabriel, K. R. 1969. Simultaneous test procedures—some theory of multiple comparisons. Annals of Mathematical Statistics 40:224–250.

Green, E., and K. J. Elgersma. 2010. PS 48-179: How often are ecologists wrong? Lindley's paradox in ecology. Poster *in* 95th Ecological Society of America Annual Meeting, Pittsburgh, Pennsylvania, USA. http://eco.confex.com/eco/2010/techprogram/P25724.HTM

Jeffreys, H. 1961. Theory of probability. Oxford University Press, London, USA.

Lavine, M., and M. J. Schervish. 1999. Bayes factors: what they are and what they are not. American Statistician 53(2):119–122.

Lindley, D. V. 1957. A statistical paradox. Biometrika 44(1–2):187–192.

Matthews, R. 1998. The great health hoax. Sunday Telegraph, 13 September, 1998. Telegraph Media Group, London, UK.

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(3):203–206.

Sellke, T., M. J. Bayarri, and J. O. Berger. 2001. Calibration of p values for testing precise null hypotheses. American Statistician 55(1):62–71.

Washington Post. 2004. Medical journal editors take hard line on drug research. http://www.smh.com.au/articles/2004/09/09/1094530773888.html

Ecology, 95(3), 2014, pp. 627–630

P values are only an index to evidence: 20th- vs. 21st-century statistical science

K. P. Burnham[1] and D. R. Anderson

Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, Fort Collins, Colorado 80523 USA

Overview Comments

We were surprised to see a paper defending *P* values and significance testing at this time in history. We respectfully disagree with most of what Murtaugh (2014) states. The subject of *P* values and null hypothesis significance tests is an old one and criticisms by statisticians began in the late 1930s and have been relentless (see *Commentaries on Significance Testing* for a partial impression of the large literature on the subject [*available online*]).[2] Oakes (1986) summed it up over 25 years ago, "It is extraordinarily difficult to find a statistician who argues explicitly in favor of the retention of significance tests ..."

For the most part, we do not comment point by point, instead we briefly contrast several historical and contemporary aspects of statistical science. The emphasis is on the information-theoretic (IT) approaches that permit computing several post-data quantities that are evidential, avoid conditioning on the null hypothesis, avoid *P* values, provide model likelihoods and evidence ratios, and allow formal inferences to be made based on all the models in an a priori set (multimodel inference).

Historical Statistics

Murtaugh (2014) reviews several of the historical methods for data analysis in simple situations; these methods focus on "testing" a null hypothesis by computing a "test statistic," assuming its asymptotic distribution, setting an arbitrary α level, and computing a *P* value. The *P* value usually leads to an arbitrary simplistic binary decision as to whether the result is "statistically significant" or not. In other cases, the *P* value is stated and interpreted as if it were evidential. The *P* value is defined as the pre-data probability: Prob{*a test statistic as large as, or larger, than that observed, given the null*}. That is, the anticipated data are being thought of as random variables.

Manuscript received 6 June 2013; accepted 9 July 2013; final version received 17 July 2013. Corresponding Editor: A. M. Ellison. For reprints of this Forum, see footnote 1, p. 609.

[1] E-mail: kenb@colostate.edu

[2] http://www.indiana.edu/~stigtsts

Theory underlying these methods for statistical inference is thus based on pre-data probability statements, rather than on the exact achieved data, and reflects early approaches (e.g., Student's influential paper [Student 1908]). In general, these early methods are not useful for non-nested models, observational data, and large data sets involving dozens of models and unknown parameters. 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. There is a very large literature on problems and limitations of null hypothesis significance testing and it is not confined to ecology or biology.

At a deeper level, *P* values are not proper evidence as they violate the likelihood principle (Royall 1997). Another way to understand this is the "irrelevance of the sample space principle" where *P* values include probabilities of data never observed (Royall 1997). Royall (1997) gives a readable account of the logic and examples of why *P* values are flawed and not acceptable as properly quantifying evidence. *P* values are conditional on the null hypothesis being true when one would much prefer conditioning on the data. Virtually everyone uses *P* values as if they were evidential: they are not. *P* values are not an appropriate measure of strength of evidence (Royall 1997). Among other flaws, *P* values substantially exaggerate the "evidence" against the null hypothesis (Hubbard and Lindsay 2008); this can often be a serious problem. In controversial settings, such as many conservation biology issues, the null hypothesis testing paradigm, hence *P* values, put the "burden of proof" on the party holding the "null position" (e.g., state and federal agencies and conservation organizations).

In even fairly simple problems, one is faced with the "multiple testing problem" and corrections such as Bonferroni's are problematic when analyzing medium to large data sets. Anderson et al. (2000) provides a more detailed review of these and other technical issues. C. R. Rao, the well-known statistician and former Ph.D. student under R. A. Fisher (see Rao 1992), summarized the situation, "... in current practice of testing a null hypothesis, we are asking the wrong question and getting a confusing answer."

Statistical science has seen huge advances in the past 50–80 years, but the historical methods (e.g., *t* tests, ANOVA, step-wise regression, and chi-squared tests) are still being taught in applied statistics courses around the world. Nearly all applied statistics books cover only historical methods. There are perhaps two reasons for this: few rewards for updating course materials and lack of awareness of viable alternatives (e.g., IT and Bayesian). Students leave such classes thinking that "statistics" is no more than null hypotheses and *P* values and the arbitrary ruling of "statistical significance." Such courses are nearly always offered in a least squares setting, instead of the more general likelihood setting which would serve those wanting to understand generalized linear models and the Bayesian approaches. Murtaugh (2014) argues that *P* values and AIC differences are closely related (see his Fig. 2). However, the relationship holds only for the simplest case (i.e., comparison of two nested models differing by only one parameter). Thus, his "result" is *not* at all general. We believe that scientists require powerful modern methods to address the complex, real world issues facing us (e.g., global climate change, community dynamics, disease pandemics).

21st-Century Statistical Science

Methods based on Bayes theorem or Kullback-Leibler information (Kullback and Leibler 1951) theory allow advanced, modern approaches and, in this context, science is best served by moving forward from the historical methods (progress should not have to ride in a hearse). We will focus on the information-theoretic methods in the material to follow. Bayesian methods, and the many data resampling methods, are also useful and other approaches might also become important in the years ahead (e.g., machine learning, network theory). We will focus on the IT approaches as they are so compelling and easy to both compute and understand. We must assume the reader has a basic familiarity with IT methods (see Burnham and Anderson 2001, 2002, 2004, Anderson 2008).

Once data have been collected and are ready for analysis, the relevant interest changes to post-data probabilities, likelihood ratios, odds ratios, and likelihood intervals (Akaike 1973, 1974, 1983, Burnham and Anderson 2002, 2004, Burnham et al. 2009). An important point here is that the conditioning is on the data, not the null hypothesis, and the objective is inference about unknowns (parameters and models). Unlike significance testing, IT approaches are not "tests," are not about testing, and hence are free from arbitrary cutoff values (e.g., $\alpha = 0.05$).

Statisticians working in the early part of the 20th century understood likelihoods and likelihood ratios

$$\mathcal{L}(\theta_0)/\mathcal{L}(\hat{\theta}).$$

This is an evidence ratio about parameters, *given* the model and the data. It is the likelihood ratio that defines evidence (Royall 1997); however, Fisher and others, thinking of the data (to be collected) as random variables, then showed that the transformation

$$-2\log\left\{\mathcal{L}(\theta_0)/\mathcal{L}(\hat{\theta})\right\}$$

was distributed asymptotically as chi squared. Based on that result they could compute tail probabilities (i.e., *P* values) of that sampling distribution, given the null hypothesis. While useful for deriving and studying theoretical properties of "data" (as random variables) and planning studies, this transformation is unnecessary (and unfortunate) for data analysis. Inferential data analysis, given the data, should be based directly on the likelihood and evidence ratios, leaving *P* values as only an index to evidence. Such *P* values are flawed whereas likelihood ratios are evidential without the flaws of *P* values. Early statisticians (e.g., Fisher) had the correct approach to measuring formal evidence but then went too far by mapping the evidence into tail probabilities (*P* values). Likelihood ratios and *P* values are very different (see Burnham and Anderson 2002:337–339). Just because the two approaches can be applied to the same data should not, and does not, imply they are both useful, or somehow complementary. Inferentially they can behave quite differently.

The information-theoretic approaches allow a quantification of K-L information loss (Δ) and this leads to the likelihood of model *i*, given the data, $\mathcal{L}(g_i \mid \text{data})$, the probability of model *i*, given the data, $\text{Prob}\{g_i \mid \text{data}\}$, and evidence ratios about models. The probabilities of model *i* are critical in model averaging and unconditional estimates of precision that include model selection uncertainty. These fundamental quantities cannot be realized using the older approaches (e.g., *P* values).

Recent advances in statistical science are not always new concepts and methods, but sometimes an enlightened and extended understanding of methods with a long history of use (e.g., Fisher's likelihood theory). There is a close link between K-L information, Boltzmann's entropy ($H' = $ K-L), and the maximized log-likelihood. Akaike (1981, 1992) considered the information-theoretic methods to be extensions to Fisher's likelihood theory (Edwards 1992). In his later work, Akaike (1977, 1985) dealt more with maximizing entropy (H') rather than (the equivalent) minimizing K-L information. Entropy and information are negatives of each other (i.e., $-H' = $ information) and both are additive.

Twenty-first-century science is about making formal inference from all (or many of) the models in an a priori set (multimodel inference). Usually there is uncertainty about which model is actually "best." Information criteria allow an estimate of which model is best, based on an explicit, objective criterion of "best," and a quantitative measure of the uncertainty in this selection (termed "model selection uncertainty"). Estimates of precision, either for prediction or parameter estimation,

include a component for model selection uncertainty, conditional on the model set.

Information-theoretic approaches are very different from historical methods that focus on P values. There is no need for a formal null hypothesis, no concept of the asymptotic distribution of a test statistic, no α level, no P value, and no ruling of "statistical significance." Furthermore, the "burden of proof" is the same across hypotheses/models when using an IT approach. Chamberlain's famous (1890) paper advocated hard thinking leading to multiple hypotheses that were thought to be plausible (most null hypotheses are false on a priori grounds). He wanted post-data probabilities of these alternatives. He must have been disappointed to see the field of statistics lean toward testing null hypotheses with little attention to evidence for or against a single alternative hypothesis, much less multiple alternative hypotheses.

Simple P values conditioned on the null hypothesis prevent several important approaches useful in empirical science: ways to rank models and the science hypotheses they represent, ways to deal with non-nested models (most model sets contain non-nested models), ways to incorporate model selection uncertainty into estimates of precision, ways to model average estimates of parameters or predictions, ways to reduce model selection bias in high dimensional problems (Lukacs et al. 2007, 2010), ways to assess the relative importance of predictor variables, ways to deal with large systems and data sets (e.g., 50–100 models, each with 10–300 parameters, where sample size might be in the thousands), ways to analyze data from observational studies (where the distribution of the test statistic is unknown).

The limitations of P values, as above, are very serious in our current world of complexity.

Comments on the "Scientific Method" and Statistical Science

While the exact definition of the so-called "scientific method" might be controversial, nearly everyone agrees that the concept of "falsifiability" is a central tenant of empirical science (Popper 1959). It is critical to understand that historical statistical approaches (i.e., P values) leave no way to "test" the alternative hypothesis. The alternative hypothesis is never tested, hence cannot be rejected or falsified! The breakdown continues when there are several alternative hypotheses (as in most real-world problems). The older methods lack ways to reject or falsify any of these alternative hypotheses. This is surely not what Popper (1959) or Platt (1964) wanted. "Support" for or against the alternative hypothesis is only by default when using P values. Surely this fact alone makes the use of significance tests and P values bogus. Lacking a valid methodology to reject/falsify the alternative science hypotheses seems almost a scandal.

It seems that Chamberlin's (1890) notion concerning alternative science hypotheses that are considered plausible should also be considered an integral part of the scientific method. Perhaps it is best if the "scientific method" embraced the concepts of formal evidence and likelihood in judging the relative value of alternative hypotheses because they provide a formal "strength of evidence."

Another serious limitation relates to the common case where the P value is "not quite" statistically significant (e.g., $P = 0.07$ when $\alpha = 0.05$). The investigator then concludes "no difference" and the null hypothesis prevails. Even worse, they often also conclude there is no evidence against the null hypothesis. Evidence ratios provide actual evidence in terms of odds, for example, at $P = 0.07$ (under normal theory and 1 df) the evidence is 5.2 to 1 against the null hypothesis. At $P = 0.05$, the evidence is 6.8 to 1 against the null. At $P = 0.096$ the evidence is 4 to 1 against the null, or equivalently, 4 to 1 in favor of the alternative. Depending on the context, even 3 to 1 odds might be useful or impressive; this is very different from concluding "no evidence against the null hypothesis." If the odds are, say, 224 to 1 (this is for $P = 0.001$), then the result must be considered as very convincing and evidence presented this way is much more understandable than saying $P = 0.001$. No automatic "cut-off" (e.g., $P = 0.05$) is relevant in an evidential paradigm such as IT. The interpretation of the evidence, being usually context specific, is left to the investigator: science is about evidence, not about sharp dichotomies or decisions.

Summary

Early statistical methods focused on pre-data probability statements (i.e., data as random variables) such as P values; these are not really inferences nor are P values evidential. Statistical science clung to these principles throughout much of the 20th century as a wide variety of methods were developed for special cases. Looking back, it is clear that the underlying paradigm (i.e., testing and P values) was weak. As Kuhn (1970) suggests, new paradigms have taken the place of earlier ones: this is a goal of good science. New methods have been developed and older methods extended and these allow proper measures of strength of evidence and multimodel inference. It is time to move forward with sound theory and practice for the difficult practical problems that lie ahead.

Given data the useful foundation shifts to post-data probability statements such as model probabilities (Akaike weights) or related quantities such as odds ratios and likelihood intervals. These new methods allow formal inference from multiple models in the a prior set. These quantities are properly evidential. The past century was aimed at finding the "best" model and making inferences from it. The goal in the 21st century is to base inference on all the models weighted by their model probabilities (model averaging). Estimates of precision can include model selection uncertainty leading to variances conditional on the model set. The 21st century will be about the quantification of

information, proper measures of evidence, and multimodel inference. Nelder (1999:261) concludes, "The most important task before us in developing statistical science is to demolish the *P*-value culture, which has taken root to a frightening extent in many areas of both pure and applied science and technology."

Acknowledgments

The authors thank the Colorado Cooperative Fish and Wildlife Research Unit for support.

Literature Cited

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. Pages 267–281 *in* B. N. Petrov and F. Csaki, editors. Second International Symposium on Information Theory. Akademiai Kiado, Budapest, Hungary.

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control AC 19:716–723.

Akaike, H. 1977. On entropy maximization principle. Pages 27–41 *in* P. R. Krishnaiah, editor. Applications of statistics. North-Holland, Amsterdam, The Netherlands.

Akaike, H. 1981. Likelihood of a model and information criteria. Journal of Econometrics 16:3–14.

Akaike, H. 1983. Information measures and model selection. International Statistical Institute 44:277–291.

Akaike, H. 1985. Prediction and entropy. Pages 1–24 *in* A. C. Atkinson and S. E. Fienberg, editors. A celebration of statistics. Springer, New York, New York, USA.

Akaike, H. 1992. Information theory and an extension of the maximum likelihood principle. Pages 610–624 *in* S. Kotz and N. L. Johnson, editors. Breakthroughs in statistics. Volume 1. Springer-Verlag, London, UK.

Anderson, D. R. 2008. Model based inference in the life sciences: a primer on evidence. Springer, New York, New York, USA.

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.

Burnham, K. P., and D. R. Anderson. 2001. Kullback-Leibler information as a basis for strong inference in ecological studies. Wildlife Research 28:111–119.

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

Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference: understanding AIC and BIC in model selection. Sociological Methods and Research 33:261–304.

Burnham, K. P., D. R. Anderson, and K. P. Huyvaert. 2009. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behavioral Ecology and Sociobiology 65:223–235.

Chamberlin, T. C. 1890. The method of multiple working hypotheses. Science 15:92–96. (Reprinted 1965, Science 148:754–759.)

Edwards, A. W. F. 1992. Likelihood: expanded edition. Johns Hopkins University Press, Baltimore, Maryland, USA.

Hubbard, R., and R. M. Lindsay. 2008. Why *P* values are not a useful measure of evidence in statistical significance testing. Theory Psychology 18:69–88.

Kuhn, T. S. 1970. The structure of scientific revolutions. Second edition, University of Chicago Press, Chicago, Illinois, USA.

Kullback, S., and R. A. Leibler. 1951. On information and sufficiency. Annals of Mathematical Statistics 22:79–86.

Lukacs, P. M., K. P. Burnham, and D. R. Anderson. 2010. Model selection bias and Freedman's paradox. Annals of the Institute of Statistical Mathematics 62:117–125.

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

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

Nelder, J. A. 1999. Statistics for the millennium. Statistician 48:257–269.

Oakes, M. 1986. Statistical inference: a commentary for the social and behavioral sciences. Wiley, New York, New York, USA.

Platt, J. R. 1964. Strong inference. Science 146:347–353.

Popper, K. R. 1959. The logic of scientific discovery. Harper and Row, New York, New York, USA.

Rao, C. R. 1992. R. A. Fisher: The founder of modern statistics. Statistical Science 7:34–48.

Royall, R. M. 1997. Statistical evidence: a likelihood paradigm. Chapman and Hall, London, UK.

Student. 1908. The probable error of a mean. Biometrika 6:1–25.

Ecology, 95(3), 2014, pp. 631–636

Model selection for ecologists: the worldviews of AIC and BIC

Ken Aho,[1,4] DeWayne Derryberry,[2] and Teri Peterson[3]

[1]*Department of Biological Sciences, Idaho State University, Pocatello, Idaho 83209 USA*
[2]*Department of Mathematics, Idaho State University, Pocatello, Idaho 83209 USA*
[3]*Division of Health Sciences, Idaho State University, Pocatello, Idaho 83209 USA*

Introduction

Ecologists frequently ask questions that are best addressed with a model comparison approach. Under this system, the merit of several models is considered without necessarily requiring that (1) models are nested, (2) one of the models is true, and (3) only current data be used. This is in marked contrast to the pragmatic blend of Neyman-Pearson and Fisherian significance testing conventionally emphasized in biometric texts (Christensen 2005), in which (1) just two hypotheses are under consideration, representing a pairwise comparison of models, (2) one of the models, H_0, is assumed to be true, and (3) a single data set is used to quantify evidence concerning H_0.

As Murtaugh (2014) noted, null hypothesis testing can be extended to certain highly structured multi-model situations (nested with a clear sequence of tests), such as extra sums of squares approaches in general linear models, and drop in deviance tests in generalized linear models. This is especially true when there is the expectation that higher order interactions are not significant or nonexistent, and the testing of main effects does not depend on the order of the tests (as with completely balanced designs). There are, however, three scientific frameworks that are poorly handled by traditional hypothesis testing.

First, in questions requiring model comparison and selection, the null hypothesis testing paradigm becomes strained. Candidate models may be non-nested, a wide number of plausible models may exist, and all of the models may be approximations to reality. In this context, we are not assessing which model is correct (since none are correct), but which model has the best predictive accuracy, in particular, which model is expected to fit future observations well. Extensive ecological examples can be found in Johnson and Omland (2004), Burnham and Anderson (2002), and Anderson (2008).

Second, the null hypothesis testing paradigm is often inadequate for making inferences concerning the falsification or confirmation of scientific claims because it does not explicitly consider prior information. Scientists often do not consider a single data set to be adequate for research hypothesis rejection (Quinn and Keough 2002:35), particularly for complex hypotheses with a low degree of falsifiability (i.e., Popper 1959:266). Similarly, the support of hypotheses in the generation of scientific theories requires repeated corroboration (Ayala et al. 2008).

Third, ecologists and other scientists are frequently concerned with the plausibility of existing or default models, what statistician would consider null hypotheses (e.g., the ideal free distribution, classic insular biogeography, mathematic models for species interactions, archetypes for community succession and assembly, etc.). However, null hypothesis testing is structured in such a way that the null hypothesis cannot be directly *supported* by evidence. Introductory statistical and biometric textbooks go to great lengths to make this conceptual point (e.g., DeVeaux et al. 2013:511, 618, Moore 2010:376, Devore and Peck 1997:300–303).

Parsimony: Fit vs. Complexity

In deciding which model is the best, criteria are necessary that allow model comparisons. While some scientists feel that more complex models are always more desirable (cf. Gelman 2009), others prefer those that balance uncertainty, caused by excessively complex models, and bias, resulting from overly simplistic models. The latter approach emphasizes parsimony. A parsimonious model should (Aho 2013), "be based on (be subset from) a set of parameters identified by the investigator as ecologically important, including, if necessary, covariates, interactions, and higher order terms, and have as few parameters as possible (be as simple as possible, but no simpler)."

Consider the examination of species population descriptor (e.g., number of individuals) as a function of an environmental factor in which the true relationship between Y and X is $Y_i = e^{(X_i - 0.5)} - 1 + \varepsilon_i$, where $\varepsilon_i \sim N(0, 0.01)$ (black lines in Fig. 1). We randomly sample for the conditional values of Y_i 10 times and apply two models, a simple linear regression (Fig. 1a), and a fifth-order polynomial (Fig. 1b). The simpler model underfits the data and misses the nonlinear association of Y and X

Manuscript received 26 July 2013; revised 7 August 2013; accepted 12 August 2013. Corresponding Editor: A. M. Ellison. For reprints of this Forum, see footnote 1, p. 609.

[4] E-mail: ahoken@isu.edu

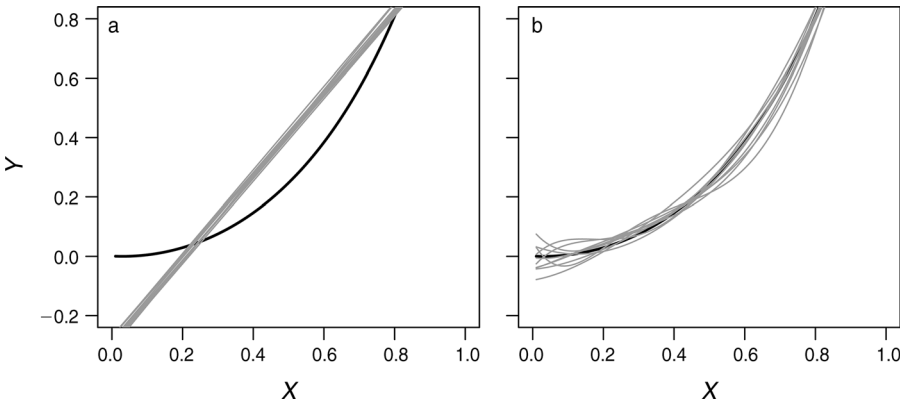

FIG. 1. Two sorts of models fit to a random process: (a) two parameter (simple linear regression) and (b) six parameter (fifth-order polynomial) (c.f., Sakamoto et al. 1986). The heavy black line indicates the true relationship between Y and X, while the gray lines are fits from linear models based on random data sets, each with 100 paired observations. Despite its complexity, the polynomial model is more parsimonious (average Akaike information criterion [AIC] = −167 vs. −42) because it captures the curvilinear nature of the association.

(Fig. 1a). The polynomial model, however, introduces erratic variability, and prevents statements of generality. Thus, both simplistic and overly complex models prevent valid inferences. The usefulness of a criterion that establishes the line between underfit and overfit models is obvious.

TWO PARSIMONY ESTIMATORS: AIC AND BIC

A Web of Science search conducted for this paper revealed that for ecological publications from 1993–2013, the two most popular measures of parsimony were the Akaike information criterion (AIC; Akaike 1973) and the Bayesian information criterion (BIC), also called the Schwarz or SIC criterion (Schwarz 1978). Specifically, for publications that implemented formal methods for multi-model inference, 84% used AIC, 14% used BIC, while only 2% used some other approach (Table 1). Murtaugh (2013) discusses AIC extensively in his defense of P values, but ignores BIC, prompting its consideration and comparison here. We posit that P values are at odds with BIC in the same way Bayesian hypothesis testing is at odds with P values (cf. Kass and Raftery 1995). Indeed, when substituting BIC for AIC in Murtaugh's derivation of P values from ΔAIC, fixed P values do not equate to fixed differences in BIC, unless n is fixed. This is consistent with the fact that P values must decrease (holding other factors constant) to favor the alternative hypothesis as sample size increases. AIC and BIC are defined as

$$\text{AIC} = -2 \ln \mathcal{L}(\hat{\theta}) + 2p$$

$$\text{BIC} = -2 \ln \mathcal{L}(\hat{\theta}) + p \ln n$$

where $\mathcal{L}(\hat{\theta})$ is the likelihood of the estimated model (in the context of general linear models, e.g., regression and ANOVA, this is the likelihood of the parameters in $N(0, \hat{\sigma}^2)$ given the model residuals, where $\hat{\sigma}^2$ is the maximum likelihood estimate for the variance of the error term distribution), p is the total number of parameters that are estimated in the model (including σ^2 for general linear models), and n is the sample size. For both indices, smaller values indicate better models.

AIC and BIC are generally introduced in textbooks (often together) as alternative measures for parsimony (cf. Kutner et al. 2005). Perhaps as a consequence, ecologists often use these measures interchangeably (or even simultaneously) without consideration of their differing qualities and merits. This view of exchangeability has, perhaps, been further entrenched by a recent ecological comparison of these methods that found no difference in efficacy among AIC, BIC, and several other criteria (Murtaugh 2009), and by articles that present these summaries side by side. However, we will show that this view is misplaced. In the remainder of this paper we explore and contrast BIC and AIC and make recommendations for their respective use in multi-model inference by ecologists.

TABLE 1. Results from a Web of Science search of publications on 13 August 2013 using the Science Citation Index Expanded (SCI-EXPANDED) database for the years 1993–2013.

Search terms for topic	No. citations	Proportion
Model selection AND (AIC* OR Akaike) AND ecol*	139	0.84
Model selection AND (BIC OR Bayes factor OR Schwarz) AND ecol*	23	0.14
Model selection AND (mallow OR FPE OR KIC OR Hannan-Quinn, Geweke-Meese) AND ecol*	4	0.02

AIC and BIC: Mathematical Motivations and Objectives

When simply examining the formula for AIC and BIC it is easy to misunderstand AIC and BIC as competing criteria intended to achieve the same goal. Both criteria balance simplicity (measured by, p, the dimension of the fitted model parameter space) and goodness of fit (measured by maximized likelihood). However the initial question, which curve "best" fits the data, can be paraphrased in a number of ways, and AIC and BIC are each answers to different questions, once the question is stated more precisely.

The most obvious reason likelihood alone cannot be used to pick between models is that models with more free parameters (when models are nested) will always have higher maximum likelihood. Akaike (1973) wanted to estimate the likelihood of a model while adjusting for the bias introduced by maximum likelihood. Using the Kullback-Liebler (KL) distance, he was able to formulate the log-likelihood maximization problem in such a way that the bias associated with likelihood maximization could be estimated and corrected for (see Burnham and Anderson 2002). Given discrete probability models, KL information is

$$I(f,g) = \sum_x f(x) \ln\left[\frac{f(x)}{g(x)}\right]$$

where $f(x)$, defines the probabilistic densities of the error distribution associated with the true model, while $g(x)$ defines the error density of an approximating model with known parameters. The term $I(f, g)$ represents the information lost when the candidate model is used to represent truth. Because the log of a quotient is the difference of logs, KL information can be separated into the difference of two summations. The first is equivalent to Shannon-Weiner diversity (information per individual) from community ecology (Pielou 1966). The second represents the log of the probability of the union of observed disjoint events.

Akaike's approach achieves an important objective: asymptotic efficiency (Shibata 1976). Asymptotic efficiency is essentially minimized prediction error. Criteria like AIC maximize predictive accuracy.

The approach taken by Schwarz (1978) is the asymptotic approximation, for the regular exponential family, of a Bayesian hypothesis testing procedure (Kass and Raftery 1995, Robert 2007). The BIC procedure derived by Schwarz is consistent. Thus, when the sample size increases, the correct model, from any group of models, is selected.

Schwarz and Akaike appear to have thought their approaches were in conflict. Schwarz (1978) wrote: "For large numbers of observations, the procedures differ markedly from each other. If the assumptions of Section 2 are accepted [for the formulation of the problem as a Bayesian hypothesis test see Kass and Raftery (1995)], Akaike's criterion cannot be asymptotically optimal." Akaike felt compelled to write a paper in response (Akaike 1978), which in our view does not clarify much, but does seem to indicate Akaike would like to address an apparent paradox. In fact the conflict is easily resolved once it is acknowledged that "asymptotically optimal" can have several meanings. Asymptotic efficiency and (asymptotic) consistency are different kinds of optimality.

McQuarrie and Tsai (1998) compare a large number of model selection procedures, and immediately divide them into two classes: consistent estimators, namely BIC, Hannan and Quinn information (Hannan and Quinn 1979), and GM (Geweke and Meese 1981), and efficient estimators, namely AIC, Mallows' C_p (Mallows 1973), predicted residual sum of squares (PRESS; Allen 1974), Akaike's FPE (Akaike 1969), and cross validation. A close link between leave-one-out cross validation and AIC can be found in Stone (1977).

It is now known that there is a class of model selection tools that provide the best predictive accuracy, and that class is headed by AIC. There is also a class of confirmation/falsification tools that are consistent, and that class is headed by BIC. So when would each be used?

Two World Views

Two different approaches to simulation

Consider two different simulations, A and B. In simulation A, a very complex model produces the data, and a number of models are candidates to fit the data. Because the process producing the data is very complex, we never expect the sample size of our data sets to approach d, the parameter space of the model (or process) producing the data (i.e., $d \gg n$), nor do we necessarily expect our candidate models to match the exact functional form of the true model. Thus, d, the number of parameters in the true model need not equal p, the number of parameters in a candidate statistical model, and the parameters for an optimal model may not include the complete pool of true parameters, and/or may include extraneous parameters.

In simulation B, a relatively simple process produces the data. The sample size of the data sets can be expected to greatly exceed d, the parameter space of the model generating the data (i.e., $d \ll n$). One of the candidate models being fitted to the data is actually equivalent to the actual model that produced the data

In these two contexts, the model that best fits the data must be interpreted differently. In simulation A, we can never find the true model, we can only find the model that maximizes predictive accuracy (model selection). In simulation B, we actually expect to find the correct model, as sample size increases (confirmation/falsification).

It will become clear that AIC is appropriate for real-world situations analogous to simulation A, and BIC is appropriate for real-world situations similar to simulation B. AIC will almost always outperform BIC in

simulations designed like simulation A, and BIC will almost always outperform AIC in simulations similar to simulation B.

The BIC world

In an effort to make Bayesian inference more objective and more closely tied to Jeffreys' (1935) notion of evidence for a hypothesis, a number of statisticians (e.g., Casella et al. 2009), biometrists (Goodman 1999, Suchard et al. 2005), and ecologists (Link and Barker 2006, Ellison 1996) have adopted the notion of the Bayes factor (or posterior P values, see Ramsey and Schafer 2012) for hypothesis or model comparison. Suppose that two hypotheses, H_1 and H_2, are to be compared, then $\Pr(H_1 \mid \text{data})/\Pr(H_2 \mid \text{data})$ = posterior odds = Bayes factor × prior odds.

Kass and Raftery (1995), in their definitive paper, provide a motivation for Bayes factors and a number of applications where Bayes factors seem especially useful. The BIC formulation is an asymptotic approximation to the Bayes factor (Schwarz 1978, Robert 2007). Kass and Raftery routinely treat BIC as an approximation to Bayes factors. Thus, applications in this paper provide excellent examples where BIC would also be appropriate.

Kass and Raftery provide five such examples, including two from biology/environmental management. Each of these has two characteristics: (1) only a few potential hypotheses are considered and (2) one of the hypotheses is (essentially) correct. Although the second characteristic is not always overtly stated, they often cite consistency as a desirable asymptotic property of Bayes factors and/or BIC.

It is clear from their discussion of "Bayes factors vs. the AIC" (which is primarily a comparison of AIC and BIC) that they value BIC over AIC because it is consistent. That is, when the sample size is sufficiently large, BIC picks the correct model, while AIC picks a model more complex than the true model. This reflects a "worldview" in which hypotheses are being compared, and one of the hypotheses is correct.

The AIC world

A few scientists have a very different "world view." Breiman (2001) writes: "There are two cultures in the use of statistical modeling to reach conclusions about data. One assumes the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown." Breiman does not have an opinion on the question "AIC or BIC?" but he nonetheless seems to live in the world of simulation type A: he emphasizes the importance of cross-validation predictive accuracy as the measure of success, and models that grow in complexity as sample size increases. Similarly, Hurvich and Tsai (1989), with reference to autoregressive moving average (ARMA) time series modeling, write: "If the true model is infinite dimensional, a case that seems most realistic in practice, AIC provides an asymptotically efficient selection of a finite dimensional approximating model."

The prevalence of type A thinking is obvious throughout the popular biometric text on model selection by Burnham and Anderson (2002) and in other works by these authors (e.g., Anderson and Burnham 2002). This is because this worldview corresponds more closely to the reality of many biological investigations, particularly in ecology: extremely complex systems with an unknown (and perhaps unknowable) underlying structure (cf. Johnson and Omland 2004, Burnham et al. 2011).

Fig. 1 is typical of a type A simulation. The correct model is not one of the candidate models, so consistency is irrelevant. For those who are interested in AIC (often in forecasting or open-ended model selection) the common characteristics are (1) numerous hypotheses and (2) the conviction that all of them are to differing degrees wrong.

In the type A world, efficiency (predictive accuracy) is important. Overfitting means a model that will have a lot of random noise if used for future prediction, while underfitting means a model that will have a bias when used for future prediction. In the type A world, as sample size increases, more small (tapering) effects are picked up, and the size of the selected model increases.

In the type B world, consistency is important. Overfitting is picking a model more complex than the true model, and underfitting is picking a model simpler than the true model. As sample size increases, the true model rises to the top (cf. Anderson 2008: Appendix E).

It is easy to confuse these worlds

The quest for a procedure that is both consistent and efficient seems impossible, when looked at in this way. Specifically, efficient methods must pick larger models with increased sample size, whereas consistent methods must settle on a fixed complexity with increased sample size. One approach to model selection cannot do both. This view is supported mathematically by Yang (2005) who showed that while BIC is consistent in optimal model selection, it cannot be optimal for regression function estimation in the sense of multi-model inference, and that while AIC represents minimax-rate optimal rules for estimating the regression function, it is not consistent for optimal model selection.

One of the paradoxes of model selection is that almost all research is based on type B simulations (Breiman 2001), but most statisticians, and even ecologists (e.g., Bolker 2008, Scheiner and Willig 2011) love to quote George Box: "All models are wrong, but some are useful." It should be noted that Box, and his most important work, time series forecasting, is fundamentally type A. At least one well-known statistics textbook suggests data splitting as the optimal way to find the best model, but if this is impossible one should use BIC, as opposed to AIC, because it is consistent—an odd

TABLE 2. The worlds of AIC and BIC contrasted.

Factor	AIC	BIC
Mathematical characteristics		
Derivation	Estimated information loss.	Approximate Bayes factor.
Optimality criterion	Asymptotic efficiency.	Consistency.
Close cousins	Data splitting, Mallows' C_p, PRESS.	Hannan-Quinn, Geweke and Meese, Bayes factors, and Bayesian hypothesis testing.
World View		
Problem statement	Multiple incompletely specified or infinite parameter models.	A small number of completely specified models/hypotheses.
Perspective	"All models are wrong, but some are useful."	"Which model is correct?"
Simulation structure	$d \gg n$	$d \ll n$
With increased n ...	Best model grows more complex.	Procedure focuses in on one best model.
Applications		
Context	Exploratory analysis; model selection to address which model will best predict the next sample; imprecise modeling; tapering effects.	Confirmatory analysis; hypothesis testing; model selection to address which model generated the data; Low dimension, precisely specified models.
Ecological examples	Complex model selection applications, e.g., predictive models for community, landscape, and ecosystem ecology; time series applications including forecasting.	Controlled experiments, for instance in physiology/enzymatics/genetics with a limited number of important, well-understood, biological predictors; models including expected or default (null) frameworks, e.g., enzyme kinetics models, Hardy-Weinberg equilibrium, or RAD curves, one of which is expected to be correct.

Notes: The number of parameters in the true model is d; sample size is n. Abbreviations are: PRESS, predicted residual sum of squares; and RAD, ranked abundance distribution.

mixture of type A and type B reasoning (Montgomery et al. 2008:60).

It is interesting to consider the performance of AIC and BIC in the context of increasingly large data sets. With respect to BIC, it is clear that, given type B simulation, the larger the sample size, the larger the probability BIC selects the true model. The relationship is less well defined with AIC (since n is not specified in its formula). However, one would expect that, in a type A simulation, as sample size increases (and consequently larger models are selected), that predictive power would also increase. Thus, as n grows larger both criteria will work better, but with different goals in mind.

There doesn't seem to be any basis for always preferring one world view over the other, both have a place in ecological model selection (cf. Murtaugh 2009). However, there are reasons to be aware that there are two world views, and to remain consistently within a given world view on a given modeling problem.

Model selection and confirmation/falsification contrasted

Table 2 can be seen as a framework for asking questions to pin down whether AIC (or related tools) or BIC (or related tools) are appropriate for a given application. Some questions, motivated by this table are: Is your analysis exploratory (AIC) or confirmatory (BIC)? Is the analysis open-ended (AIC), or are a few specific models representing a well understood process being compared (BIC)? As the data set gets larger, do you expect your model to grow in complexity (AIC), or stabilize (BIC)? Do you believe you have chosen the correct functional form of the relationship as well as the correct variables (yes, BIC; no, AIC)? Is your goal accurate prediction (AIC) or finding the correct model (BIC)?

CONCLUSION

Murtaugh (2014) revealed an important mathematical connection between ΔAIC and P values for a comparison of two models (one nested in the other). Such an application, however, constitutes a very narrow use of an information-theoretic criterion. We agree with Murtaugh that null hypothesis testing has an important role in ecology, and that conceptual problems with this paradigm are often due to misapplication and misunderstanding by users. Nonetheless, many ecological endeavors pose questions that are not easily answered by null hypothesis tests. For instance, models may not be nested, and the ecologist may want to treat the null and alternative hypothesis as having the same status with regard to support based on the evidence. There are tools for this situation, but the proper tool depends on a further distinction. What has often been designated as model selection has been here further parsed into complex (infinite) model selection, for which AIC and related tools are the appropriate; and confirmation/falsification, for which BIC and related tools are appropriate.

Literature Cited

Aho, K. 2013. Foundational and applied statistics for biologists using R. Chapman and Hall/CRC, Boca Raton, Florida, USA.

Akaike, H. 1969. Statistical predictor identification. Annals of the Institute of Statistical Mathematics 22:203–217.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pages 267–281 *in* B. N. Petrov and S. Caski, editors. Proceedings of the Second International Symposium on Information Theory. Akademiai Kaido, Budapest, Hungary.

Akaike, H. 1978. A Bayesian analysis of the minimum AIC procedure. Annals of the Institute of Statistical Mathematics 30:9–14.

Allen, D. M. 1974. The relationship between variable selection and data augmentation and a method of prediction. Technometrics 16:125–127.

Anderson, D. R. 2008. Model based inference in the life sciences: a primer on evidence. Springer, New York, New York, USA.

Anderson, D. R., and K. P. Burnham. 2002. Avoiding pitfalls when using information-theoretic methods. Journal of Wildlife Management 66(3):912–916.

Ayala, F. J., et al. 2008. Science, evolution, and creationism. National Academies Press, Washington, D.C., USA.

Bolker, B. 2008. Ecological models and data in R. Princeton University Press, Princeton, New Jersey, USA.

Breiman, L. 2001. Statistical modeling: the two cultures. Statistical Science 16(3):199–215.

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

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

Casella, G., F. J. Giron, M. L. Martinez, and E. Moreno. 2009. Consistency of Bayesian procedures for variable selection. Annals of Statistics 37(3):1207–1228.

Christensen, R. 2005. Testing Fisher, Neyman, Pearson, and Bayes. American Statistician 59(2):121–126.

DeVeaux, R. D., P. F. Velleman, and D. E. Bock. 2013. Intro stats. Fourth edition. Pearson, Upper Saddle River, New Jersey, USA.

Devore, J. L., and R. Peck. 1997. Statistics: the explorations and analysis of data. Duxbury, Pacific Grove, California, USA.

Ellison, A. M. 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. Ecological Applications 6:1036–1046.

Gelman, A. 2009. Bayes, Jeffreys' prior distributions and the philosophy of statistics. Statistical Science 24(2):176–178.

Geweke, J., and R. Meese. 1981. Estimating regression models of finite but unknown order. International Economic Review 22:55–70.

Goodman, S. N. 1999. Toward evidence based medical statistics 2: The Bayes factor. Annals of Internal Medicine 130(12):1005–1013.

Hannan, E. J., and B. G. Quinn. 1979. The determination of the order of an autoregression. Journal of the Royal Statistical Society B 41:190–195.

Hurvich, C. M., and C.-L. Tsai. 1989. Regression and time series model selection in small samples. Biometrika 76(2):297–307.

Jeffreys, H. 1935. Some tests of significance, treated by the theory of probability. Proceedings of the Cambridge Philosophical Society 31:203–222.

Johnson, J., and K. Omland. 2004. Model selection in ecology and evolution. Trends in Ecology and Evolution 19(2):101–108.

Kass, R. E., and E. Raftery. 1995. Bayes factors. Journal of the American Statistical Association 90(430):773–795.

Kutner, M. H., C. J. Nachtsheim, J. Neter, and W. Li. 2005. Applied linear statistical models. Fifth edition. McGraw-Hill, Boston, Massachusetts, USA.

Link, W. A., and R. J. Barker. 2006. Model weights and the foundations of multimodel Inference. Ecological Applications 87:2626–2635.

Mallows, C. L. 1973. Some comments on C_P. Technometrics 15 (4):661–675.

McQuarrie, A. D. R., and C.-L. Tsai. 1998. Regression and time series model selection. World Scientific, Singapore.

Montgomery, D. C., C. L. Jennings, and M. Kulahci. 2008. Introduction to time series analysis and forecasting. Wiley series in probability and statistics. Wiley, Hoboken, New Jersey, USA.

Moore, D. S. 2010. The basic practice of statistics. Fifth edition. Freeman, New York, New York, USA.

Murtaugh, P. A. 2009. Performance of several variable-selection methods applied to real ecological data. Ecology Letters 12(10):1061–1068.

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

Pielou, E. C. 1966. Shannon's formula as a measure of specific diversity: its use and misuse. American Naturalist 100(914):463–465.

Popper, K. 1959. The logic of scientific discovery. Routledge, London, UK.

Quinn, G. P., and M. J. Keough. 2002. Experimental design and data analysis for biologists. Cambridge University Press, Cambridge, UK.

Ramsey, F. L., and D. W. Schafer. 2012. The statistical sleuth: a course in the methods of data analysis. Third edition. Brooks/Cole, Belmont, California, USA.

Robert, C. P. 2007. The Bayesian choice: from decision-theoretic foundations to computational implementation. Second edition. Springer, New York, New York, USA.

Sakamoto, Y., M. Ishiguro, and G. Kitagawa. 1986. Akaike information criterion statistics. KTK Scientific Publishers, Tokyo, Japan.

Scheiner, S. M., and M. R. Willig. 2011. The theory of ecology. University of Chicago Press, Chicago, Illinois, USA.

Schwarz, G. 1978. Estimating the dimension of a model. Annals of Statistics 6:461–464.

Shibata, R. 1976. Selection of the order of an autoregressive model by Akaike's information criterion. Biometrika 63:117–126.

Stone, M. 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society B 39(1):44–47.

Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2005. Models for estimating Bayes factors with applications to phylogeny and tests of monophylogeny. Biometrics 61:665–673.

Yang, Y. 2005. Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. Biometrika 92(4):937–950.

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

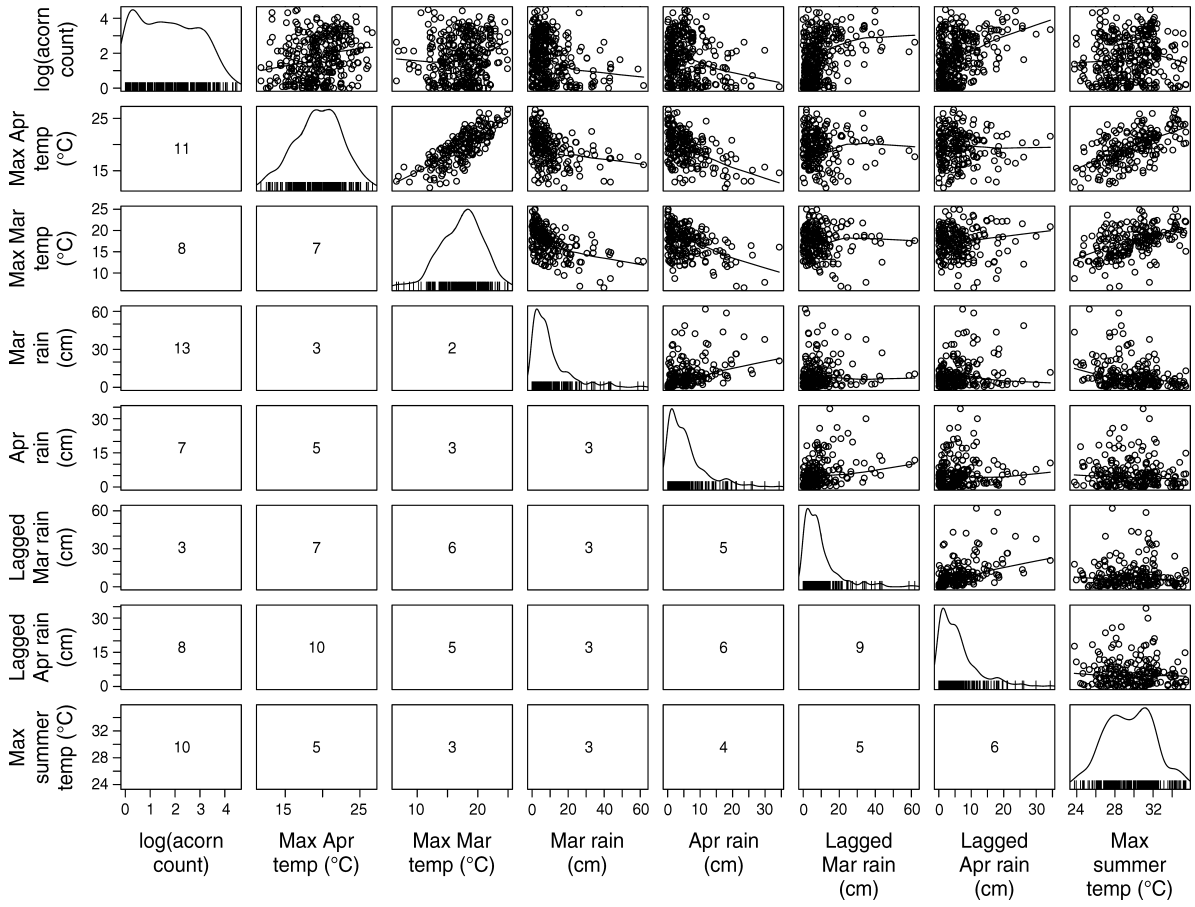

FIG. 1. Plot used for exploratory data analysis of the example data. The diagonal shows density plots representing the distribution of each variable. The upper right triangle of panels shows bivariate scatterplots of all variables, with a lowess smooth function. The lower left triangle of panels is a 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/

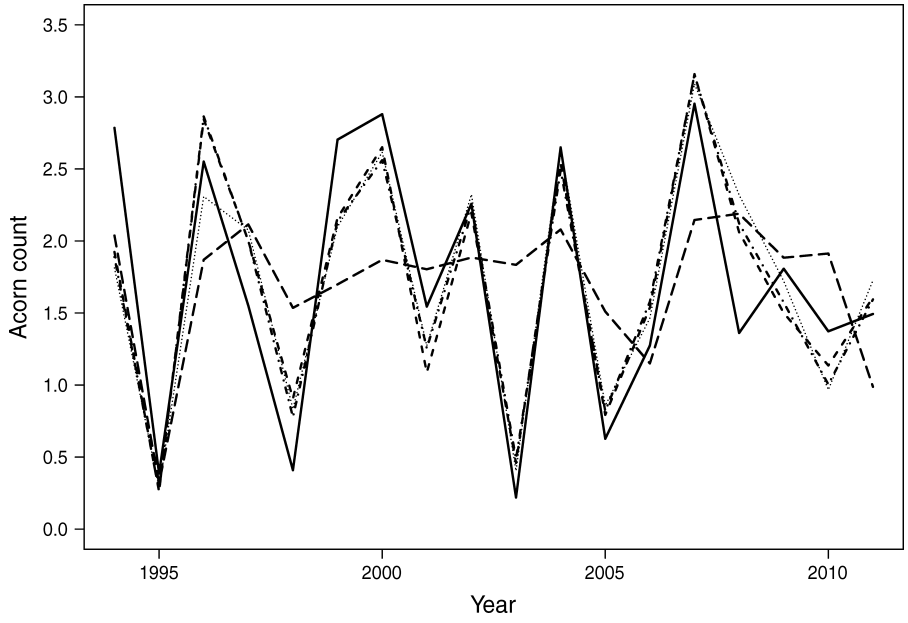

FIG. 2. Time series (solid line) of acorn count, averaged across species and sites. Predictions from models including a single parameter (long-dashed line), four parameters selected by null hypothesis significance testing (NHST; dotted line), five parameters selected by Akaike's information criterion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

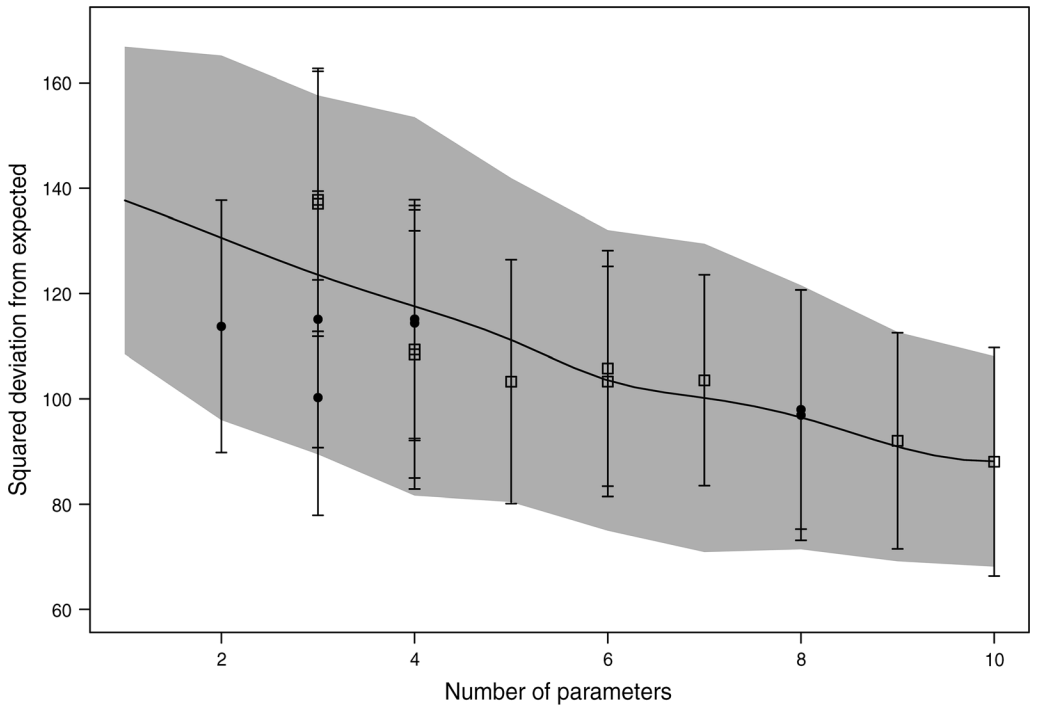

FIG. 3. Predictive ability of the final models selected by survey participants shown as the squared deviation from expected with 30% of observations as test data for 400 observations. Solid circles represent final models selected by AIC, while open squares are final models selected by NHST. Error bars 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

P *values, confidence intervals, ΔAIC, and evidence*

In a statistical model with p parameters where H_0 sets k of them to 0, the mathematical relationship between P values and ΔAIC is given by Murtaugh's Eq. 5:

$$P = \Pr(\chi^2_k > \Delta\text{AIC} + 2k) \text{ and } \Delta\text{AIC} = F^{-1}_{\chi^2_k}(1-P) - 2k$$

For example, if $k = 1$, then $P = 0.05$ corresponds to $\Delta\text{AIC} = 1.84$ and $\Delta\text{AIC} = 6$ corresponds to $P = 0.0047$. Or if $k = 2$, then $P = 0.05$ corresponds to $\Delta\text{AIC} = 1.99$ and $\Delta\text{AIC} = 6$ corresponds to $P = 0.0067$.

Murtaugh adds, "[t]he P value is a continuous measure of the strength of evidence against the null hypothesis" and presumably believes that ΔAIC is also a continuous measure of the strength of evidence. But both can't be precisely true, at least not in a formal sense, because, when k changes, the same P corresponds to different values of ΔAIC and the same ΔAIC corresponds to different values of P. Either the translation from P to evidence must change in different problems, or the translation of ΔAIC to evidence must change, or both. (Note that the P value goes backward, P goes down as the evidence against H_0 goes up, while ΔAIC goes forward. Murtaugh's Fig. 2 shows that P and ΔAIC go in opposite directions.)

Perhaps Murtaugh means that P values measure evidence informally. But when he says, "[t]he smaller the P value, the more evidence we have against the null hypothesis," he suggests a monotonic relationship between P and evidence. Schervish (1996) showed that such a suggestion cannot be true. According to Schervish, "[*P*-values have] often been suggested as a measure of the support that the observed data $X = x$ lend to [the hypothesis] H, or the amount of evidence in favor of H. This suggestion is always informal, and no theory is ever put forward for what properties a measure of support or evidence should have . . . . We [state] a simple logical condition . . . and show why a measure of support should satisfy this condition. We then demonstrate that *P*-values do not satisfy the condition."

If Schervish is right, we cannot interpret P values as measures of support or evidence, so it is worth understanding his logic. Let Θ denote the parameter space. Divide Θ into a null and an alternative hypothesis twice—(H_0, H_a) and (H_0', H_a')—so that $H_0 \subset H_0'$ In an example from Schervish's paper, θ is a one-dimensional parameter, H_0 is the point-null $\theta = 0$, and H_0' is the one-sided null $\theta \le 0$. In set notation, $H_0 \equiv 0 \subset (-\infty, 0] \equiv H_0'$. The alternative hypotheses are the complements: $H_a : \theta \neq 0$ and H_a': $\theta > 0$ and satisfy $H_a' \equiv (0, \infty) \subset (-\infty, 0) \cup (0, \infty) \equiv H_a$. Schervish's argument rests on the following four points:

1) Because $H_0 \subset H_0'$, any evidence against H_0' is also evidence against H_0. In the example, any evidence that θ lies outside H_0' (evidence that $\theta > 0$) is also evidence that θ lies outside H_0 (evidence that $\theta \neq 0$).

2) Therefore, any measure of evidence M against null hypotheses must satisfy $M(H_0) \ge M(H_0')$. Because P values go backward, we should have $P_{H_0} \le P_{H_0'}$.

3) P values do not satisfy this condition. In the example, if $x > 0$ then $P_{H_0} = 2P_{H_0'}$. We have $P_{H_0} > P_{H_0'}$ when we should have $P_{H_0} \le P_{H_0'}$.

4) Therefore, P values are not, and cannot be translated into, measures of evidence against null hypotheses.

According to Schervish, the problem is more general than just one-sided and two-sided alternatives: "one can try to think of the P values for different values of x as the different degrees to which different data values would support a single hypothesis H. This might work as long as we do not acknowledge the possibility of other hypotheses. [But a] serious drawback to this approach is that the scale on which support is measured is not absolute, but rather depends on the hypothesis."

By "scale . . . depends on the hypothesis," he means that $P = 0.05$ in one problem (e.g., the one-sided null) is a different amount of evidence than $P = 0.05$ in a different problem (e.g., the point null). See Schervish (1996) for details and for how point nulls and one-sided nulls are two ends of a continuum that includes interval nulls. See Lavine and Schervish (1999) for how Bayes factors exhibit the same flaw.

A similar reversal can occur in Murtaugh's setting of nested linear models. To keep things simple, adopt the model

$$X_1 \sim \text{N}(\theta_1, 1) \quad \text{and} \quad X_2 \sim \text{N}(\theta_2, 1)$$

and the hypotheses

$$H_0 : \theta_1 = \theta_2 = 0 \quad \text{and} \quad H_0' : \theta_1 = 0.$$

The model is admittedly unrealistic. Its value is that $\hat\theta_1 = x_1$ is independent of $\hat\theta_2 = x_2$, which makes the math much easier to follow. As in Schervish, $H_0 \subset H_0'$, so any evidence against H_0' is also evidence against H_0. Suppose the data are $x_1 = 2$ and $x_2 = 0$. Then

$$P_{H_0} = \Pr[X_1^2 + X_2^2 \ge 4] = \Pr[\chi^2_2 \ge 4] = 0.135$$

$$P_{H_0'} = \Pr[X_1^2 \ge 4] = \Pr[\chi^2_1 \ge 4] = 0.0455.$$

We have $P_{H_0} > P_{H_0'}$, so our P values go the same way as Schervish's and, like his, contradict his logical condition for evidence. More generally, χ^2_2 stochastically dominates χ^2_1, so the contradiction would occur for any value of x_1 as long as $x_2 \approx 0$. An examination of more typical models in which $\hat\theta_1$ is not independent of $\hat\theta_2$, would require consideration of the covariance matrix of $\hat\theta$ and the resulting two-dimensional confidence ellipses and is beyond the scope of this comment.

The contradiction occurs also for ΔAIC. Consider a linear model with $p = 2$ parameters and two null hypotheses $H_0 : \theta_1 = \theta_2 = 0$ and $H_0' : \theta_1 = 0$. In Murtaugh's notation, k is the number of parameters set to 0 in the null hypothesis, so $k = 2$ and $k' = 1$. Since H_0 is a

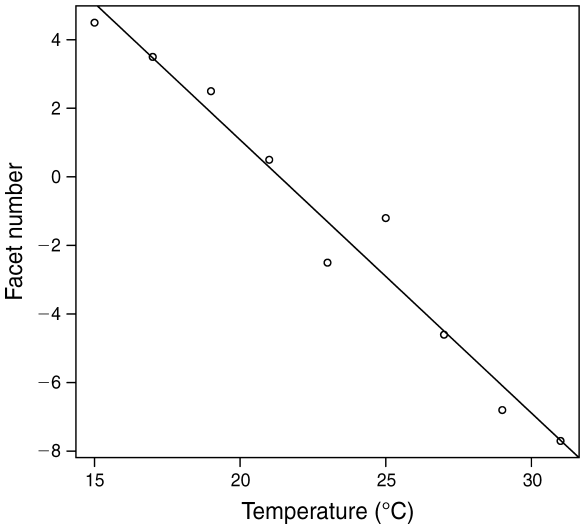

FIG. 1. Mean number of eye facets of *Drosophila melanogaster* raised at different temperatures. Based on data from Berkson (1942).

subset of H_0', we should have $\Delta\text{AIC} \geq \Delta\text{AIC}'$. By Murtaugh's Eq. 4

$$\Delta\text{AIC} = -2\log\left\{\mathcal{L}(\hat{\theta}_0)/\mathcal{L}(\hat{\theta})\right\} - 2k$$

and

$$\Delta\text{AIC}' = -2\log\left\{\mathcal{L}(\hat{\theta}_0')/\mathcal{L}(\hat{\theta})\right\} - 2k'$$

so

$$\Delta\text{AIC} - \Delta\text{AIC}' = -2\left\{\log(\mathcal{L}(\hat{\theta}_0)) - \log(\mathcal{L}(\hat{\theta}_0')) + (k - k')\right\}.$$

$H_0 \subset H_0'$ implies $k - k' \geq 0$ and $\log(\mathcal{L}(\hat{\theta}_0)) - \log(\mathcal{L}(\hat{\theta}_0')) \leq 0$, so the $\mathcal{L}$ and k terms work in opposite directions. The difference $\Delta\text{AIC} - \Delta\text{AIC}'$ will be either positive or negative according to whether the difference in loglikelihoods is larger than $k - k'$. To create a Schervish-style contradiction we need only create a dataset in which $\hat{\theta}_0 \approx \hat{\theta}_0'$, so that $\log(\mathcal{L}(\hat{\theta}_0)) - \log(\mathcal{L}(\hat{\theta}_0')) \approx 0$ and, consequently, $\Delta\text{AIC} - \Delta\text{AIC}' \approx -2(k - k')$ is negative. That will happen when the mle of the second coordinate is near 0, or $\hat{\theta}_2 \approx 0$, just as for the P value example.

We saw earlier that the translation from either P, ΔAIC, or both, to evidence must differ in different problems. Schervish showed that the translation from P differs in problems with point and one-sided null hypotheses. I have just shown that the translations from both P and ΔAIC must also differ in nested linear models. Neither P nor ΔAIC can be consistently interpreted as a measure of evidence without regard to the problem at hand.

What use is P *and what more is there?*

Murtaugh's interpretation is that "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." I agree with the first half of that interpretation: a small P indicates that H_0 does not explain the data well. But I also agree with Berkson (1942) about the second half: "As an illustration of a test of linearity under the caption, '*Test of straightness of regression line*,' R. A. Fisher utilizes data relating the temperature to the number of eye facets of *Drosophila melanogaster* .... Fisher says, 'The deviations from linear regression are evidently larger than would be expected, if the regression were really linear .... There can therefore be no question of the statistical significance of the deviations from the straight line.' I have plotted the data ...."

I, too, have plotted the data, in Fig. 1. Fisher finds a small P value and rejects linearity, even though the plot shows a strong linear relationship. This might seem to be a case of statistical vs. practical significance, but Berkson continues, "If the regression were curvilinear, a small P is to be expected relatively frequently .... But also a small P is to be expected relatively frequently if the regression is linear and the variability heteroscedastic ... [o]r if the regression is linear and ... [temperature, the x variable] is not constant but subject to fluctuation. And there may be other conditions which, with linearity, would produce a small P relatively frequently. The small P is favorable evidence for any or several of these."

How are we to tell which of these alternatives, or any, provide a better explanation of the data than H_0? The answer is, in a word, graphics. Plot data and plot residuals. Do not automatically adopt the obvious alternative hypothesis. Do not rely solely on P values or any other numerical summary. Plots can at once show nonlinearity, heteroscedasticity, and many other possible departures from H_0. For example, Fig. 1 suggests to me the possibility that the facet numbers for temperatures 23 and 25 have been swapped. I doubt that would have occurred to me had I looked only at numerical summaries.

Plots are descriptive statistics, to be used informally. So are P values, confidence intervals, and ΔAIC. In fact, the P value is just a one-number summary of the familiar plot of the null-density of a test statistic along with a mark for its observed location. That plot and its P value summary are sometimes useful, as are confidence intervals and ΔAIC. But other plots are typically just as useful, or even more.

Summary

(1) Murtaugh and I agree on an important point: abandon accept/reject declarations. That alone will go a long way to improving statistical practice. (2) Don't

confuse P values or ΔAIC with binary declarations. An argument against one is not necessarily an argument against the other. (3) Be careful interpreting a P value or ΔAIC as strength of evidence. That interpretation cannot be made formal and the connection between P, ΔAIC, and evidence must be recalibrated for each new problem. (4) Plot. Check models. Plot. Check assumptions. Plot.

Literature Cited

Berkson, J. 1942. Tests of significance considered as evidence. Journal of the American Statistical Association 37:325–335.

Lavine, M., and M. J. Schervish. 1999. Bayes factors: What they are and what they are not. American Statistician 53:119–122.

Murtaugh, P. A. 2014. In defence 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.

Ecology, 95(3), 2014, pp. 645–651

Recurring controversies about P values and confidence intervals revisited

Aris Spanos[1]

Department of Economics, Virginia Tech, Blacksburg, Virginia 24061 USA

Introduction

The use, abuse, interpretations and reinterpretations of the notion of a P value has been a hot topic of controversy since the 1950s in statistics and several applied fields, including psychology, sociology, ecology, medicine, and economics.

The initial controversy between Fisher's significance testing and the Neyman and Pearson (N-P; 1933) hypothesis testing concerned the extent to which the pre-data Type I error probability α can address the arbitrariness and potential abuse of Fisher's *post-data threshold* for the P value. Fisher adopted a falsificationist stance and viewed the P value as an indicator of disagreement (inconsistency, contradiction) between data $\mathbf{x}_0$ and the null hypothesis (H_0). Indeed, Fisher (1925:80) went as far as to claim that "The actual value of p ... indicates the strength of evidence against the hypothesis." Neyman's behavioristic interpretation of the pre-data Type I and II error probabilities precluded any evidential interpretation for the accept/reject the null (H_0) rules, insisting that accept (reject) H_0 does not connote the truth (falsity) of H_0. The last exchange between these protagonists (Fisher 1955, Pearson 1955, Neyman 1956) did nothing to shed light on these issues. By the early 1960s, it was clear that neither account of frequentist testing provided an adequate answer to the question (Mayo 1996): When do data $\mathbf{x}_0$ provide evidence for or against a hypothesis H?

Manuscript received 4 July 2013; revised 15 August 2013; accepted 16 August 2013. Corresponding Editor: A. M. Ellison. For reprints of this Forum, see footnote 1, p. 609.

[1] E-mail: aris@vt.edu

The primary aim of this paper is to revisit several charges, interpretations, and comparisons of the P value with other procedures as they relate to their primary aims and objectives, the nature of the questions posed to the data, and the nature of their underlying reasoning and the ensuing inferences. The idea is to shed light on some of these issues using the *error-statistical* perspective; see Mayo and Spanos (2011).

Frequentist Testing and Error Probabilities

In an attempt to minimize technicalities but be precise about the concepts needed, the discussion will focus on the hypotheses

$$H_0\colon \mu = \mu_0 \text{ vs. } H_1\colon \mu > \mu_0 \tag{1}$$

in the context of *the simple Normal model* $X_t \sim \text{NIID}(\mu, \sigma^2)$, $t = 1, 2, \ldots, n, \ldots$, where NIID stands for normal, independent, and identically distributed.

Fisher vs. Neyman-Pearson (N-P) approaches

In the case of the above null hypothesis, Fisher's significance and the Neyman-Pearson (N-P) hypothesis testing revolve around the test statistic

$$\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s} \overset{H_0}{\sim} \text{St}(n-1) \tag{2}$$

where $\text{St}(n-1)$ denotes a Student's t distribution with $(n-1)$ degrees of freedom, and

$$\bar{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t, \qquad s^2 = \frac{1}{n-1}\sum_{t=1}^{n}(X_t - \bar{X}_n)^2.$$

Fisher's significance testing ignores the alternative hypothesis in Eq. 1 and uses Eq. 2 to evaluate the P

value: $\mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); H_0) = p(\mathbf{x}_0)$, which is traditionally defined as the probability of obtaining a value of a test statistic $\tau(\mathbf{x})$ at least as extreme as the one observed $\tau(\mathbf{x}_0)$, assuming that H_0 is true. A P value lower then a designated threshold, say 0.05, is viewed as evidence against H_0. For historical accuracy, this needs to be viewed in conjunction with Fisher's *falsificationist stance* concerning testing in the sense that significance tests can *falsify* but never *verify* hypotheses (Fisher 1955). The subsequent literature, however, did extend the interpretation of P values to allow for large enough values to be viewed as *moderate* to *no evidence against* H_0; see Murtaugh (2013).

The same sampling distribution (Eq. 2) is used to define the Neyman-Pearson (N-P) Type I error probability: $\mathbb{P}(\tau(\mathbf{X}) > c_\alpha; H_0) = \alpha$, where c_α is the critical value for significance level α. This defines the t test

$$T_\alpha^{>} = \left\{ \tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s},\ C_1(\alpha) = \{\mathbf{x} : \tau(\mathbf{x}) > c_\alpha\} \right\} \tag{3}$$

where $C_1(\alpha)$ denotes the rejection region and the superscripted $>$ denotes a one-sided test in the positive direction. The N-P approach differs from that of Fisher by justifying the choice of both $\tau(\mathbf{X})$ and $C_1(\alpha)$ on optimality grounds, i.e., the choices in Eq. 3 maximize the power: $\mathbb{P}(\tau(\mathbf{X}) > c_\alpha; \mu = \mu_1) = \pi(\mu_1)$, for $\mu_1 > \mu_0$ Note that the Type II error probability is $\beta(\mu_1) = 1 - \pi(\mu_1)$, for all $\mu_1 > \mu_0$ To evaluate the power, one needs the sampling distribution of $\tau(\mathbf{X})$ under H_1

$$\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s} \overset{\mu=\mu_1}{\sim} \mathrm{St}(\delta_1, n-1), \qquad \text{for } \mu_1 > \mu_0$$

where

$$\delta_1 = \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}$$

denotes the non-centrality parameter. It can be shown that test $T_\alpha^{>}$, as defined in Eq. 3, is optimal, uniformly most powerful (UMP); see Lehmann (1986). The power of a N-P test provides a measure of its generic [for any $\mathbf{x} \in \mathbb{R}^n$] capacity to detect different discrepancies from the null, given α.

A crucial difference between the P value and the Type I and II error probabilities is that the former is defined *post-data,* since it requires $\tau(\mathbf{x}_0)$, but the latter are defined *pre-data* since they only require n and the choice of α. Despite that, the P value is often viewed by practitioners as the observed significance level and recast the accept/reject rules into (Lehmann 1986): reject H_0 if $p(\mathbf{x}_0) \leq \alpha$, accept H_0 if $p(\mathbf{x}_0) > \alpha$, because the data specificity of $p(\mathbf{x}_0)$ seems more informative than the dichotomous accept/reject decisions.

P *value and the large* n *problem*

A crucial weakness of both the P value and the N-P error probabilities is the so-called large n problem: there is always a large enough sample size n for which any simple null hypothesis. H_0: $\mu = \mu_0$ will be rejected by a frequentist α-significance level test; see Lindley (1957). As argued in Spanos (2013), there is nothing paradoxical about a small P value, or a rejection of H_0, when n is large enough.

It is an inherent feature of a good (consistent) frequentist test, as $n \to \infty$ the power of the test $\pi(\mu_1)$, for any discrepancy $\gamma \neq 0$ from the null goes to one, i.e., $\pi(\mu_1) \underset{n\to\infty}{\to} 1$. What is fallacious is to interpret a rejection of H_0 as providing the same weight of evidence *for* a particular alternative H_1, irrespective of whether n is large or small. This is an example of a more general fallacious interpretation that stems from the fact that all rejections of H_0 are viewed as providing the same weight of evidence for a particular alternative H_1, regardless of the generic capacity (the power) of the test in question. The large n problem arises because, in light of the fact that

$$\delta_1 = \frac{\sqrt{n}(\mu_1 - \mu_0)}{\sigma}$$

the power depends crucially on n; it increases with $\sqrt{n}$. This renders a rejection of H_0 with a small n (low power) very different—in evidential terms—than one with a large n (high power). Hence, the claim that "the smaller the P value the more the evidence we have against the null hypothesis" (Murtaugh 2013) needs to be qualified. Indeed, the real problem does not lie with the P value or the accept/reject rules as such, but with how such results are transformed into *evidence for* or *against* a hypothesis H_0 or H_1.

The large n constitutes an example of a broader problem known as the *fallacy of rejection*: (mis)interpreting reject H_0 (evidence *against* H_0) as evidence *for* a particular H_1; this can arise when a test has very high power, e.g., large n. A number of attempts have been made to alleviate the large n problem, including rules of thumb for decreasing α as n increases; see Lehmann (1986). Due to the trade-off between the Type I and II error probabilities, however, any attempt to ameliorate the problem renders the inference susceptible to the reverse fallacy known as the *fallacy of acceptance*: (mis)interpreting accept H_0 (*no* evidence *against* H_0) as evidence *for* H_0; this can easily arise when a test has very low power; e.g., α is tiny or n is too small.

These fallacies are routinely committed by practitioners in many applied fields. After numerous unsuccessful attempts, Mayo (1996) provided a reasoned answers to these fallacies in the form of a post-data severity assessment.

Severity and the Fallacies of Acceptance/Rejection

Whether data $\mathbf{x}_0$ provide evidence for or against a particular hypothesis H depends crucially on the generic capacity (power) of the test to detect discrepancies from the null. This stems from the intuition that a small P value or a rejection of H_0 based on a test with low power (e.g., a small n) for detecting a particular discrepancy γ provides stronger evidence for γ than using a test with

much higher power (e.g., a large n). This intuition is harnessed by a post-data severity evaluation of accept/reject based on custom-tailoring the generic capacity of the test to establish the discrepancy γ warranted by data $\mathbf{x}_0$; see Mayo (1996).

Post-data severity evaluation

The severity evaluation is a *post-data* appraisal of the accept/reject and P value results with a view to provide an *evidential interpretation*; see Mayo and Spanos (2011). A hypothesis H (H_0 or H_1) "passes" a severe test T_α with data $\mathbf{x}_0$ if (i) $\mathbf{x}_0$ accords with H and (ii) with very high probability, test T_α would have produced a result that accords less well with H than $\mathbf{x}_0$ does, if H were false (Mayo and Spanos 2006).

The notion of severity can be used to bridge the gap between accept/reject rules and P values and an evidential interpretation in so far as the result that H passes test T_α provides good evidence for inferring H (is correct) to the extent that T_α severely passes H with data $\mathbf{x}_0$. The severity assessment allows one to determine whether there is evidence for (or against) inferential claims of the form $\mu_1 = \mu_0 + \gamma$, for $\gamma \geq 0$, in terms of a discrepancy γ from μ_0, which includes H_0 as well as any hypothesis belonging to the alternative parameter space $\mu_1 > \mu_0$.

For the case of *reject* H_0, the relevant claim is $\mu > \mu_1 = \mu_0 + \gamma$, $\gamma \geq 0$, with a view to establish the *largest discrepancy* γ from H_0 warranted by data $\mathbf{x}_0$. In this case, $\mathbf{x}_0$ in condition (i) accords with H_1, and condition (ii) concerns "results $\mathbf{x} \in \mathbb{R}^n$ that accord less well with H_1 than $\mathbf{x}_0$ does." Hence, the severity evaluation is

$$\begin{aligned}\text{SEV}(T_\alpha; \mathbf{x}_0; \mu > \mu_1) &= \mathbb{P}(\tau(\mathbf{X}) \leq \tau(\mathbf{x}_0); \mu > \mu_1 \text{ false}) \\ &= \mathbb{P}(\tau(\mathbf{X}) \leq \tau(\mathbf{x}_0); \mu \leq \mu_1) \qquad (4)\end{aligned}$$

where $\mathbb{P}(\tau(\mathbf{X}) \leq \tau(\mathbf{x}_0); \mu \leq \mu_1)$ is evaluated at $\mu = \mu_1$ since the $\text{SEV}(\mu < \mu_1)$ increases for $\mu < \mu_1$. Analogously, for accept H_0 (Mayo and Spanos 2006)

$$\text{SEV}(T_\alpha; \mathbf{x}_0; \mu \leq \mu_1) = \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu = \mu_1). \qquad (5)$$

It should be emphasized that what is important for interpretation purposes is not the numerics of the tail areas, but the coherence of the underlying reasoning.

Revisiting the P *value: a severity perspective*

To bring out a key weakness of the P value as a measure of evidence, let us relate it to the severity evaluation for reject H_0 by restricting the latter at $\gamma = 0$:

$$\begin{aligned}\text{SEV}(T_\alpha; \mathbf{x}_0; \mu > \mu_0) &= \mathbb{P}(\tau(\mathbf{X}) \leq \tau(\mathbf{x}_0); \mu \leq \mu_0) \\ &= 1 - \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu \leq \mu_0) \\ &\geq 1 - P(\mathbf{x}_0).\end{aligned}$$

This suggests that, for a small P value ($P = 0.01$), $1 - P(\mathbf{x}_0) = 0.99$, provides a lower bound for the severity assessment of $\mu > \mu_0$. Viewed from this vantage point, a small P value establishes the existence of *some* discrepancy $\gamma \geq 0$, but provides no information concerning *its magnitude*.

The severity evaluation remedies this weakness of the P value by taking into account the generic capacity of the test to output the magnitude of the discrepancy γ warranted by data $\mathbf{x}_0$ and test T_α. This, however, necessitates considering alternative values of μ within the same statistical model. This is because N-P testing is inherently testing within the boundaries of a statistical model, as opposed to mis-specification (M-S) testing which probes outside those boundaries, with the prespecified model representing the null; see Mayo and Spanos (2004).

Statistical vs. substantive significance

The post-data severity evaluation in the case of *reject* H_0 outputs which inferential claims of the form $\mu > \mu_1$ are warranted (high severity) or unwarranted (low severity) on the basis of test T_α and data $\mathbf{x}_0$. This provides the basis for addressing the *statistical* vs. *substantive* significance problem that has bedeviled practitioners in several fields since the 1950s. Once the warranted discrepancy γ^* is established, one needs to confer with substantive subject matter information to decide whether this discrepancy is *substantively significant* or not. Hence, not only statistical significance does not imply substantive significance, but the reverse is also true. A *statistically insignificant* result can implicate a substantively significant discrepancy; see Spanos (2010*a*) for an empirical example.

The severity perspective calls into question the use of *effect size measures*, based on "distance functions" using point estimators, as flawed attempts to evaluate the warranted discrepancy by attempting to eliminate the influence of the sample size n in an ad hoc way. Indeed, classifying effect sizes as "small," "medium," and "large" (Cumming 2011), without invoking subject matter information, seems highly questionable. In contrast, the post-severity evaluation accounts for the effect of the sample size n by taking into consideration the generic capacity of the test to output the warranted discrepancy γ in a principled manner, and then lets the subject matter information make the call about substantive significance.

More generally, in addition to circumventing the fallacies of acceptance and rejection, severity can be used to address other charges like the "arbitrariness" of the significance level, the one-sided vs. two-sided framing of hypotheses, the reversing of the null and alternative hypotheses, the effect size problem, etc.; see Mayo and Spanos (2011). In particular, the post-data severity evaluation addresses the initial arbitrariness of any threshold relating to the significance level or the P value by relying on the *sign* of $\tau(\mathbf{x}_0)$, and not on c_α, to indicate the *direction* of the inferential claim that "passed." Indeed, this addresses the concerns for the dichotomy created by any threshold; see Spanos (2011).

P Values and CIs

For the simple Normal model, the $(1-\alpha)$ CI for μ

$$\mathbb{P}\left(\bar{X}_n - c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \bar{X}_n + c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)\right) = 1-\alpha \quad (6)$$

is optimal in the sense that it has the shortest expected length. Its optimality can be demonstrated using the mathematical duality between Eq. 6 and the UMP unbiased test (Lehmann 1986)

$$T_\alpha = \left\{\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s},\ \ C_1(\alpha) = \{\mathbf{x} : |\tau(\mathbf{x})| > c_\alpha\}\right\}$$

associated with the hypotheses $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$. The mathematical duality between hypothesis testing and CIs, however, has beclouded the crucial differences between the two types of inference procedures and led to several misleading claims, like (a) CIs are more informative than tests and (b) CIs avoid most of the weaknesses of tests. As argued by Murtaugh (2013): "*P* values and confidence intervals are just different ways of summarizing the same information." The truth is that these two procedures pose very different questions to the data and they elicit distinct answers.

CIs vs. hypothesis testing: the underlying reasoning

The key difference between hypothesis testing and CIs is that the sampling distribution underlying Eq. 6 does not coincide with Eq. 2, but instead takes the form

$$\tau(\mathbf{X};\mu) = \frac{\sqrt{n}(\bar{X}_n - \mu)}{s} \overset{\mu=\mu^*}{\sim} \mathrm{St}(n-1) \quad (7)$$

where $\tau(\mathbf{X};\ \mu)$ is a pivot (not a test statistic) and the evaluation does not invoke *hypothetical reasoning* ($\mu = \mu_0$), but *factual* $\mu = \mu^*$ (μ^* being the true value of μ, whatever that happens to be). Hence, a more pertinent way to write Eq. 6 is

$$\mathbb{P}\left(\bar{X}_n - c_{\frac{\alpha}{2}}(\frac{s}{\sqrt{n}}) \leq \mu \leq \bar{X}_n + c_{\frac{\alpha}{2}}(\frac{s}{\sqrt{n}});\ \mu = \mu^*\right) = 1-\alpha \quad (8)$$

which makes explicit the underlying reasoning. This crucial difference is often obscured by blurring the distinction between the null value μ_0 and the true value μ^* when deriving a CI by solving the *acceptance region*

$$C_0(\alpha) = \left\{\mathbf{x} : \left|\frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s}\right| \leq c_\alpha\right\}$$

for μ_0, and then pretending that μ_0 stands, not for all its *unknown* values μ within that interval. What makes the blurring between μ_0 and the true value μ^* particularly elusive is that the mathematical duality ensures that under both modes of reasoning, hypothetical and factual, one is evaluating the same tail areas of $\mathrm{St}(n-1)$ for hypothesis testing and CIs. What is important for interpretation purposes, however, is not the numerics of the tail areas, but the coherence of the underlying reasoning and the nature of the ensuing inferences.

An important upshot of factual reasoning is that, *post-data*, one cannot attach a probability to the observed CI

$$\mathrm{OCI} = (\bar{x}_n - c_{\frac{\alpha}{2}}(s/\sqrt{n}) \leq \mu \leq \bar{x}_n + c_{\frac{\alpha}{2}}(s/\sqrt{n})) \quad (9)$$

because the post-data coverage probability is either zero or one; the factual scenario $\mu = \mu^*$ has played out and OCI either includes or excludes μ^*. Hence, one has no way to distinguish between more likely and less likely values of μ within an OCI using factual reasoning. Note that in hypothesis testing, post-data error probabilities, like the *P* value, are definable since the reasoning is hypothetical, and thus it applies equally post-data as well as pre-data. However, the mathematical duality enables one to use OCI as a surrogate test for two-sided hypotheses, by (illicitly) switching between the two different modes of reasoning.

Ironically, practitioners in several applied fields are happy to use this mathematical duality, but ignore the fact that some of the charges leveled at the *P* value apply equally to CIs. For instance, the CI in Eq. 8 is equally vulnerable to the large *n* problem because its expected length

$$E\left(\left[\bar{X}_n + \frac{s}{\sqrt{n}}c_{\frac{\alpha}{2}}\right] - \left[\bar{X}_n - \frac{s}{\sqrt{n}}c_{\frac{\alpha}{2}}\right]\right) = 2c_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

shrinks down to zero as $n \to \infty$; see also Murtaugh (2013). This calls into question various claims that OCIs provide more reliable information than *P* values when it comes to the relevant "effect size" (whatever that might mean).

Observed CIs and severity

The *post-data severity evaluation* can be used to bring out this confusion and shed light on the issues of distinguishing between different values of μ within an OCI. Hence, one cannot attach probabilities to inferential claims of the form

$$\mu > \bar{x}_n - c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right), \quad \text{and} \quad \mu \leq \bar{x}_n + c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) \quad (12)$$

because the coverage probability is rendered degenerate post-data. On the other hand, severity can be used to evaluate inferential claims of the form

$$\mu > \mu_1 = \mu_0 + \gamma, \mu \leq \mu_1 = \mu_0 + \gamma, \quad \text{for some } \gamma \geq 0. \quad (13)$$

Thus, in principle one can relate the observed bounds

$$\bar{x}_n \pm c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)$$

to these inferential claims

$$\mu_1 = \bar{x}_n - c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)$$

and evaluating (Mayo and Spanos 2006)

$$\text{SEV}\left(\mu > \bar{x}_n - c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)\right) \quad \text{or SEV}\left(\mu \leq \bar{x}_n + c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)\right). \tag{14}$$

A moment's reflection, however, suggests that the connection between severity and the OCI is more apparent than real. This is because the reasoning underlying the severity evaluations in Eqs. 4 and 5 is *hypothetical,* evaluated under different values $\mu = \mu_1$, and *not* factual $\mu = \mu^*$. Indeed, the inferential claims and the relevant probabilities associated with SEV(.) in Eq. 4.7 have nothing to do with the coverage probability for μ^*; they pertain to the relevant inferential claims as they relate to particular discrepancies

$$\gamma = \left(\tau(\mathbf{x}_0) \pm c_{\frac{\alpha}{2}}\right)\left(\frac{s}{\sqrt{n}}\right)$$

in light of data $\mathbf{x}_0$.

CIs vs. hypothesis testing: questions posed

Inference procedures associated with hypothesis testing and CIs share a common objective: learn from data about the "true" ($\mu = \mu^*$) statistical model $M^*(\mathbf{x}) = \{f(\mathbf{x};\theta^*)\}$, $\mathbf{x} \in \mathbb{R}^n$ yielding data $\mathbf{x}_0$. What about the questions posed?

The question posed by a CI is: How often will a random interval $[L(\mathbf{X}), U(\mathbf{X})]$ cover the true value μ^* of μ, whatever that *unknown* value μ^* happens to be? The answer comes in the form of a $(1 - \alpha)$ CI using *factual* reasoning.

The question posed by a test is: how close is the prespecified value μ_0 to μ^*?

The answer comes in the form of an optimal test whose capacity is calibrated using the pre-data error probabilities. A closer look at the test statistic

$$\tau(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{s}$$

reveals that it is effectively a standardized distance between μ^* and μ_0, since $\bar{X}_n$ is an excellent estimator of μ^* and $\bar{x}_n$ is assumed to have been generated by $M^*(\mathbf{x})$.

Revisiting Akaike-type Model Selection Procedures

Akaike (1973) introduced *model selection* within a prespecified family

$$M(m) := \{M_{\theta_i}(\mathbf{z}) = \{f(\mathbf{z};\theta_i), \theta_i \in \Theta\}, \quad \mathbf{z} \in \mathbb{R}_Z^n, \ i = 1, 2, \cdots m\} \tag{15}$$

vwhich relies on minimizing a distance function based on the estimated log-likelihood (viewed as a goodness-of-fit measure) and a penalty function relating to the number of unknown parameters $\boldsymbol{\theta}_i$ associated with each model $M_{\theta_i}(\mathbf{z})$.

The objective function is

$$\text{AIC}(i) = -2\ln\mathcal{L}(\mathbf{z};\hat{\theta}_i) + 2K_i, \quad i = 1, 2, \cdots, m \tag{16}$$

where $\mathcal{L}(\mathbf{z};\theta_i) \propto f(\mathbf{z};\theta_i)$, $\theta_i \in \Theta$ is the likelihood function and K_i is the number of unknown parameters in $\boldsymbol{\theta}_i$. It can be viewed as trading goodness-of-fit/prediction against parsimony (simplicity). The primary aim is to *rank* all the models in $M(m)$ in terms of the estimated distance function, which is often interpreted as a metric of support; see Burnham and Anderson (2002).

In the particular case of nested regression models

$$M_{\theta_i}(\mathbf{z}) : y_t = \beta_0 + \sum_{j=1}^{i}\beta_j x_t^j + u_t, u_t \sim \text{NIID}(0, \sigma^2), \quad i = 1, 2, \cdots, m \tag{17}$$

the AIC takes the specific form $\text{AIC}(i) = n\ln(\hat{\sigma}_i^2) + 2K_i$, $i = 1, 2, \ldots, m$, where

$$\hat{\sigma}_i^2 = \frac{1}{n}\sum_{t=1}^{n}\left(y_k - \hat{\beta}_0 - \sum_{j=1}^{i}\hat{\beta}_j x_t^j\right)^2.$$

Evaluating the AIC(i) for all $i = 1, 2, \ldots, m$, yields a *ranking* of the models in $M(m)$, and the smallest is chosen.

Using goodness-of-fit/prediction as the primary criterion for "ranking the different models," however, can potentially undermine the reliability of any inference in two ways. First, goodness-of-fit/prediction is neither necessary nor sufficient for *statistical adequacy*: the model assumptions like NIID are valid for data $\mathbf{Z}_0$. The latter ensures that the actual error probabilities approximate closely the nominal error probabilities. Applying a 0.05 significance level test when the actual Type I error is closer to 0.60 can easily lead an inference astray! Indeed, the appropriateness of particular goodness-of-fit/prediction measures, such as $\ln\mathcal{L}(\mathbf{z};\hat{\boldsymbol{\theta}}_i)$, is questionable when $M_{\theta_i}(\mathbf{z})$ is statistically misspecified; see Spanos (2007).

One might object to this argument on the grounds that all inference procedures are vulnerable to statistical misspecification. Why single out Akaike-type model selection? The reason is that model validation based on thorough M-S testing to secure statistical adequacy (Mayo and Spanos 2004) is in direct conflict with such model selection procedures. This is because model validation will give rise to a choice of a particular model within Eq. 17 on statistical adequacy grounds, assuming Eq. 15 includes such an adequate model. This choice would render model selection procedures redundant and often misleading because the highest ranked model will rarely coincide with the statistically adequate one, largely due to the *second* way model selection procedures could undermine the reliability of inference. As shown below, the *ranking* of the different models is inferentially equivalent to N-P testing comparisons with a serious weakness: model selection procedures ignore the relevant error probabilities. If the implicit error probabilities are too low/high, that could give rise to unreliable inferences. In addition, if no statistically adequate model exists within Eq. 17, M-S testing would confirm that and no choice will be made, but model

selection procedures would nevertheless indicate a highest ranked model; see Spanos (2010*b*) for empirical examples.

AIC vs. N-P testing

At first sight, the Akaike model selection procedure's reliance on minimizing a distance function, combining the log-likelihood and the number of unknown parameters, seems to circumvent hypothesis testing and the controversies surrounding P values and accept/reject rules. Indeed, its simplicity and apparent objectivity made it a popular procedure among practitioners.

Murtaugh (2013) brings out the connections between P values, CIs, and the AIC, and argues that: "Since P values, confidence intervals, and ΔAIC [difference of AIC] 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."

This argument is misleading because on closer examination, minimizing the AIC does not circumvent these problems and controversies. Although proponents of AIC generally discourage comparisons of only two models, the *ranking* of the different models by the AIC is inferentially equivalent to *pairwise comparisons* among the different models in $\{M_{\theta_i}(\mathbf{z}), i=1, 2, \ldots, m\}$, using N-P testing with a serious flaw: it ignores the relevant error probabilities; see Spanos (2010*b*).

To illustrate the connection between the AIC ranking and N-P testing consider a particular pairwise comparison between the following two models within Eq. 15:

$$M_0 : y_t = \beta_0 + \beta_1 x_t + u_t;$$
$$M_1 : y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \beta_3 x_t^3 + u_t. \qquad (18)$$

Let us assume that the AIC procedure selects model M_1, i.e.,

$$[n \ln(\hat{\sigma}_0^2) + 2K_0] > [n \ln(\hat{\sigma}_1^2) + 2K_1] \Rightarrow$$
$$(\hat{\sigma}_0^2/\hat{\sigma}_1^2) > \exp([2(K_1 - K_0)]/n). \qquad (19)$$

One can relate this AIC decision in favor of M_1 to the rejection of H_0

$$H_0\text{: } \beta_2 = \beta_3 = 0, \text{ vs. } H_1\text{: } \beta_2 \neq 0, \quad \text{or } \beta_3 \neq 0 \qquad (20)$$

by the F test

$$F(\mathbf{Z}) = \left([\hat{\sigma}_0^2 - \hat{\sigma}_1^2]/\hat{\sigma}_1^2\right)\left(\frac{n - K_1}{K_1 - K_0}\right),$$
$$C_1 = \{\mathbf{z} : F(\mathbf{z}) > c_\alpha\} \qquad (21)$$

where c_α denotes the critical value for significance level α. This suggests that the AIC procedure amounts to rejecting H_0 when $F(\mathbf{z}) > c_{\text{AIC}}$, for

$$c_{\text{AIC}} = \left(\frac{n - K_1}{K_1 - K_0}\right)\left[\exp\left(\frac{2(K_1 - K_0)}{n}\right) - 1\right]$$

e.g., when $n = 100$, $c_{\text{AIC}} = 1.94$, implying that the actual Type I error is $\alpha_{\text{AIC}} = 0.15$; using α_{AIC}, one can derive the implicit power function for the above F test. This indicates that the ranking of M_1 higher than M_0 by AIC involves a much higher significance level than the traditional ones. In this sense, the AIC implicitly allows for a much higher probability of rejecting the null when true. More generally, the implicit error probabilities associated with the AIC procedure are at best unknown, calling into question the reliability of any inferences. These results can be easily related to those in Murtaugh (2013) between ΔAIC and the relevant P value: $\mathbb{P}(F(\mathbf{Z}) > F(\mathbf{z}_0); H_0)$.

Summary and Conclusions

The paper focused primarily on certain charges, claims, and interpretations of the P value as they relate to CIs and the AIC. It as argued that some of these comparisons and claims are misleading because they ignore key differences in the procedures being compared, such as (1) their primary aims and objectives, (2) the nature of the questions posed to the data, as well as (3) the nature of their underlying reasoning and the ensuing inferences.

In the case of the P value, the crucial issue is whether Fisher's evidential interpretation of the P value as "indicating the strength of evidence against H_0" is appropriate. It is argued that, despite Fisher's maligning of the Type II error, a principled way to provide an adequate evidential account, in the form of post-data severity evaluation, calls for taking into account the power of the test.

The error-statistical perspective brings out a key weakness of the P value and addresses several foundational issues raised in frequentist testing, including the fallacies of acceptance and rejection as well as misinterpretations of observed CIs; see Mayo and Spanos (2011). The paper also uncovers the connection between model selection procedures and hypothesis testing, revealing the inherent unreliability of the former. Hence, the choice between different procedures should not be "stylistic" (Murtaugh 2013), but should depend on the questions of interest, the answers sought, and the reliability of the procedures.

Acknowledgements

I would like to thank D. G. Mayo for numerous discussions on issues discussed in this paper.

Literature Cited

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pages 267–281 *in* B. N. Petrov and F. Csaki, editors. Second International Symposium on Information Theory. Akademia Kiado, Budapest, Hungary.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference. Second edition., Springer, New York, New York, USA.

Cumming, G. 2011. Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis. Routledge, London, UK.

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

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

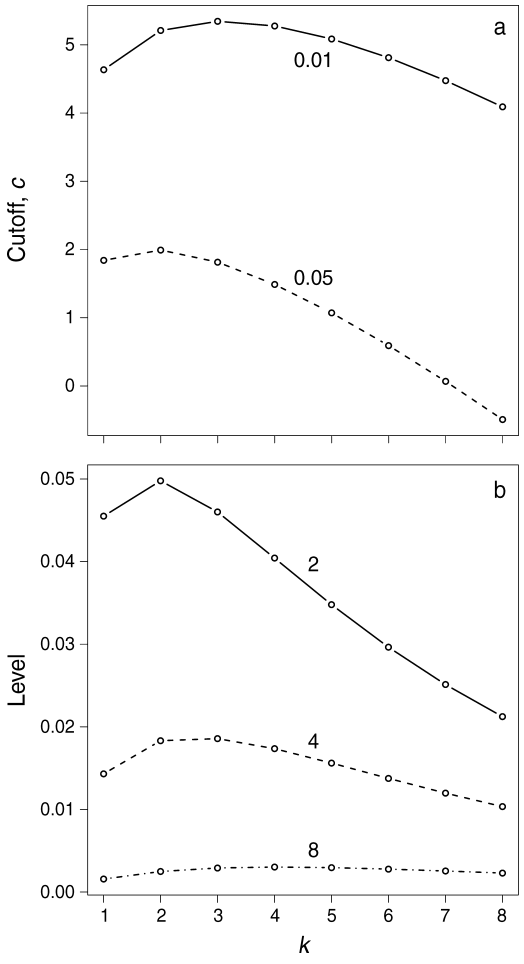

FIG. 1. For a test in which a reduced model is rejected in favor of a full model when $\Delta AIC = AIC_R - AIC_F > c$ (where AIC_R is AIC for the reduced model, AIC_F is AIC for the full model, and c is a cutoff value) (a) for a given level (i.e., probability of rejecting the reduced model when it is 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.