FORUM

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*(data | model) or *P*(model | 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.*