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

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: $P(y \geq y_{obs} \mid \boldsymbol{\theta})$. 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: