

# Alternative Statistical Approaches to the Use of Data as Evidence for Hypotheses in Human Behavioral Ecology

MARY C. TOWNER AND BARNEY LUTTBEG

In their ambitious *Evolutionary Anthropology* paper, Winterhalder and Smith<sup>1</sup> review the history, theory, and methods of human behavioral ecology (HBE). In establishing how HBE differs from traditional approaches within sociocultural anthropology, they and others laud its hypothetical-deductive research method.<sup>1–3</sup> Our aim is to critically examine how human behavioral ecologists conduct their research, specifically how they analyze and interpret data as evidence for scientific hypotheses. Through computer simulations and a review of empirical studies of human sex ratios, we consider some limitations of the status quo and present alternatives that could strengthen the field. In particular, we suggest that because human behavioral ecologists often consider multiple hypotheses, they should use statistical approaches that can quantify the evidence in empirical data for competing hypotheses. Although we focus on HBE, the principles of this paper apply broadly within biological anthropology.

HBE encompasses studies on mating, parenting, and subsistence strategies.<sup>1</sup> Human behavioral ecologists

ask, for example, why hunters target certain food sources more than others, why people choose the mates that they do, and why parents invest more in boys in some circumstances and girls in others. They ask these questions in an attempt to understand contemporary variation in human behavior according to the socioecological conditions faced by different individuals.<sup>4</sup> Some questions also bear on the likely social systems of our hominid ancestors and the selective pressures that led to the evolution of distinctive human traits such as a long postmenopausal life span. Human behavioral ecologists seek the answers to such questions in evolutionary theory by assuming that natural selection has led to physiological traits, decision-making strategies, and consequent behavior that, given current constraints, optimizes an individual's expected inclusive fitness.

How precisely do human behavioral ecologists frame their research questions? Complete HBE explanations specify both the costs and ben-

efits to individuals of alternative behaviors (“models of circumstance”) and how natural selection has shaped the traits involved (“models of mechanism”).<sup>1</sup> Explicitly or implicitly, such models include a goal, currency, constraints, and decision or alternative set.<sup>1</sup> Because of inherent difficulties in measuring and isolating influences on inclusive fitness, human behavioral ecologists usually focus on a more immediate goal and currency. An HBE parental investment study, for example, might assume that the mother's goal is to maximize the number of her children surviving to the age of 5 years, analyze early morbidity and mortality rates as currencies, consider constraints like maternal age and wealth, and look at the decisions about whether and for how long to breastfeed newborns of either sex.

Given formalized research questions and pertinent studies, human behavioral ecologists must then analyze their data and interpret the results as evidence for hypotheses. For example, how might data on breastfeeding duration support or refute the Trivers-Willard (TW) hypothesis<sup>5</sup> for differential parental investment? Evidence can be defined as the support data provide for a judgment or conclusion, with inference being the process through which such conclusions are drawn. Human behavioral ecologists use a variety of methods to summarize the evidence in their data and draw inferences from such data in evaluating hypotheses. Their analytical methods usually derive from a classical statistical tradition (see next section). They credit

Mary C. Towner is a human behavioral ecologist and postdoctoral scholar in the Department of Anthropology at the University of California, Davis. She has used archival records and computer modeling to investigate human dispersal in historical New England and is currently developing new methods for studying cultural evolution and trait transmission in cross-cultural datasets. E-mail: mctowner@ucdavis.edu

Barney Luttbeg is a behavioral ecologist in the Department of Environmental Sciences and Policy at the University of California, Davis. His research focuses on how behavioral flexibility shapes ecological and evolutionary dynamics. E-mail: btluttbeg@ucdavis.edu

Key words: human behavioral ecology; statistical inference; null hypothesis testing; significance testing; Bayesian statistics; likelihood statistics; *p*-values; model comparison; Akaike Information Criterion; parental investment; sex ratios; Trivers-Willard hypothesis

their inferential approach to a hypothetico-deductive (H-D) philosophy of science.<sup>1,2</sup>

The H-D philosophy originated in the nineteenth century with a shift away from the inductive and largely descriptive character of previous scientific methodologies.<sup>6,7</sup> Deduction is the process through which hypotheses are used to generate specific predictions that, by logic, must be correct if the generating hypotheses are correct. Scientific knowledge can therefore advance by evaluating the predictions in light of empirical evidence and subsequently identifying and modifying incorrect hypotheses. Twentieth-century practitioners of the H-D approach were additionally influenced by the philosophy of science championed by Karl Popper,<sup>8</sup> who argued that the hallmark of a scientific hypothesis is falsifiability. In contrast, a scientific hypothesis can never be proven, only supported for the time being.<sup>8</sup> And although concepts such as strong inference<sup>9</sup> encouraged scientists to work with multiple hypotheses, the task was to design crucial experiments, the outcomes of which would cleanly reject the incorrect hypotheses. As we will see, this focus on testability, specifically through falsification or rejection, fed into and was no doubt influenced by a traditional statistical approach centered on null hypothesis testing.

Unfortunately, in fields such as HBE a meaningful hypothesis can rarely be formulated so that a single test can reject it. If a prediction from a model is not supported by the data, it is always possible that the investigator used the wrong currency or overlooked key constraints. Even with a reliance on statistical hypothesis testing, it is usually the null hypothesis (typically, no difference between individuals of different categories) as opposed to the scientific hypothesis of interest that is falsifiable. One response to the criticism that HBE hypotheses are unfalsifiable is to draw strength from philosophies of science that better fit what human behavioral ecologists actually do. Keteleer and Ellis<sup>10</sup> argue that the research approach of evolutionary psychologists is more consistent with a Lakatosian philosophy of science. Lakatos suggested that scientific research programs can be viewed as consisting of a “hard

core” surrounded by layers of protective “belts.” Most basic research is not intended to falsify the hard core, but rather the more testable subsidiary hypotheses surrounding the core.<sup>10,11</sup> Even here, we face the problem of how best to “test” specific predictions.

In a recent overview, Gray<sup>12</sup> compliments the “trend [in HBE] toward testing alternative selectionist hypotheses” rather than testing a single evolutionary-minded hypothesis against a null hypothesis of random (presumably nonadaptive) behavior. We go a step farther by arguing that rather than focusing on testing hypotheses, human behavioral ecologists could better understand their data by comparing and revising mul-

---

**We go a step farther by arguing that rather than focusing on testing hypotheses, human behavioral ecologists could better understand their data by comparing and revising multiple alternative models or working hypotheses, including those that are not necessarily mutually exclusive in the strong inference sense.**

---

tiple alternative models or working hypotheses, including those that are not necessarily mutually exclusive in the strong inference sense.

Hilborn and Mangel<sup>11</sup> use the phrase “ecological detection” to refer to a broad approach to scientific research that in large part eschews the focus on testing single hypotheses, null or otherwise. Instead, ecological detection combines a variety of techniques for “confronting models with data.” To be effective, ecological detection usually requires analytical approaches that are different from the

classical significance testing approaches with which most human behavioral ecologists are familiar. Such a shift is already well under way in fields outside of anthropology.<sup>11,13–15</sup>

## DIFFERENT APPROACHES TO STATISTICAL INFERENCE

Statistical methods provide tools to examine data in an effort to make inferences about the world. Without such methods, human reasoning alone can lead us astray, finding patterns that do not exist and missing subtle patterns that do.<sup>16,17</sup> A fundamental concept in statistics is that data are a sample drawn from an existing or hypothetical larger population. In evaluating scientific hypotheses, we usually want to learn the truth about this larger population. For example, we might want to estimate from a sample of birth records the underlying probability of a mother in the population giving birth to a boy. We might also want to compare two or more samples of mothers to evaluate whether they represent different populations.

What is the best way to extrapolate from sample data to the populations from which they are derived? There are several alternative modes of statistical inference, with fundamental differences in their mathematical underpinnings, as well as corresponding acrimonious debate among their proponents. Here, we categorize these modes of inference according to whether a classical significance testing approach, a likelihood approach, or a Bayesian approach is adopted,<sup>18</sup> recognizing that these approaches did not develop and are not currently applied in isolation from one another.<sup>15</sup> We introduce the concepts underlying each mode of inference and identify some benefits and pitfalls of each (see Box 1 for quantitative illustrations).

### The Classical Significance Testing Approach

Significance testing, or null hypothesis testing, is the approach to statistical inference traditionally used by biological anthropologists. The early development of significance

testing as a mode of inference can be traced to the work of Fisher and of Neyman and Pearson.<sup>15</sup> (Note that Fisher was also instrumental in the origins of the likelihood approach to inference, as well as sex ratio theory.) Given a set of data, Fisher's method focuses on a single hypothesis and determining the probability of finding the data or data more extreme if the hypothesis were true. The smaller this probability, presumably the less likely it is that the hypothesis is true. In contrast, Neyman and Pearson's method emphasizes deciding between two hypotheses, a null hypothesis and an alternative hypothesis. After fixing  $\alpha$  (an acceptable probability of making a Type I error, falsely rejecting the null hypothesis) and subsequently minimizing  $\beta$  (the probability of making a Type II error, failing to reject a false null hypothesis), acceptance and rejection regions are calculated for possible data outcomes. The null hypothesis is then accepted or rejected based on whether the data fall into the acceptance or rejection regions.

The Fisher and Neyman-Pearson methods have merged into what is the typical practice today, despite fundamental inconsistencies in the underlying statistical theories.<sup>18-22</sup> A null and an alternative hypothesis are pitted against each other. If the data are found to be sufficiently extreme (usually a  $P$ -value less than 0.05) *assuming the null hypothesis were true*, then the null hypothesis is rejected in favor of the alternative hypothesis. We now highlight two well-known problems in how people use the merged theory.<sup>13,23</sup>

First, a  $P$ -value is often misinterpreted as the probability of the null hypothesis, when it is actually the probability of the data or more extreme data assuming the null hypothesis were true.  $P$ -values are also misinterpreted as a measurement of the strength of an effect, for instance when a result with  $p = 0.001$  is referred to as "highly significant" or "more significant" than a result with  $p = 0.05$ . Large sample sizes and precisely measured variables can lead to small  $P$ -values despite negligible effect sizes. (Effect size refers to the magnitude of an association or difference between groups, one measure of

which is Cohen's difference statistic,  $d$ , a staple of meta-analyses.<sup>24</sup>)

Also worrisome is the arbitrariness of a  $p = 0.05$  cut-off. Is it scientific to exalt a result that leads to a  $p = 0.04$  but ignore one with a  $p = 0.06$ ? Sometimes researchers describe results with  $P$ -values between 0.05 and 0.10 as being "marginally significant" and potentially important; at other times, similar outcomes are argued to show "no effect" or ignored altogether.

A more fundamental problem with significance testing is the inherent asymmetry between the null and alternative hypotheses. A Type I error is generally assumed to be worse than a Type II error, so that decision crite-

---

**Is it scientific to exalt a result that leads to a  $p = 0.04$  but ignore one with a  $p = 0.06$ ? Sometimes researchers describe results with  $P$ -values between 0.05 and 0.10 as being "marginally significant" and potentially important; at other times, similar outcomes are argued to show "no effect" or ignored altogether.**

---

ria are set to reduce the probability of Type I errors at the expense of increasing Type II errors. Subsequently, going by  $P$ -values alone, only the null hypothesis, not the alternative, can be rejected; in other words, only the alternative hypothesis, not the null, can be asserted. Some researchers operate as if the two are symmetric: a significant  $P$ -value is interpreted as evidence for the alternative hypothesis and a nonsignificant value is interpreted as evidence for the null. Unfortunately, this is just not appropriate (see Box 2). It may be impossible to reliably reject the null hypothesis with insufficient data

(small samples or effect sizes), even if the null hypothesis is in fact false as, at some level, almost all null hypotheses are. Power analysis provides some measure of a test's ability to correctly reject a false null hypothesis,<sup>24-26</sup> but it is rarely used in HBE.

### The Likelihood Approach

A second mode of statistical inference is based on likelihoods and the principle that the ratio of likelihoods for two models quantifies the statistical evidence in the data for one hypothesis *vis-à-vis* another hypothesis.<sup>22</sup> The likelihood of a hypothesis, given the data, is proportional to the probability of observing the data under the hypothesis. Fisher was the first to explicitly develop the concept of likelihood and suggest its potential use in statistical inference, notably as an alternative to "inverse probability," or what we would now term Bayesian approaches.<sup>18</sup> Although likelihood inference does not carry with it the name recognition of significance testing or Bayesian approaches, it has a long history in twentieth century statistics.<sup>18,27</sup>

To avoid potential confusion, note the important distinction between the mere presence of likelihood calculations, which play a central role in classical significance testing as well as Bayesian approaches, and the more broadly conceived "likelihood approach" to statistical inference, which is a paradigm for evaluating the evidence in data for alternative hypotheses.<sup>22</sup> Thus, a regression analysis that compares alternative models by interpreting a  $P$ -value associated with a Chi-square statistic is still making inferences through classical significance testing, even if the regression coefficients were determined through maximum likelihood estimation. In 1972, Edwards<sup>18</sup> wrote "even today, thirty-five years after Fisher drew attention to the importance of the *whole* likelihood function in estimation, it is difficult to convey to a statistical audience the vital distinction between likelihood regarded as a basis for a theory of inference, and likelihood regarded as a commodity to be maximized in a method of point estimation" (p. 101).

### Box 1. Illustration of Alternative Statistical Traditions with Sex Ratio Data

We illustrate the alternative statistical traditions with births to high-status (above median wealth) parents from the Kipsigis of Kenya (unpublished data courtesy of Monique Borgerhoff Mulder). The births consisted of 193 boys and 158 girls, yielding a secondary sex ratio (SR) of 122 (122 boys/100 girls) or equivalently a proportion of boys of 0.55 ( $\hat{\theta}$ ). We contrast this sample parameter to hypothesized population values for the probability of producing a boy, namely  $\theta = 0.50$  (SR 100, an equal sex ratio) and  $\theta = 0.53$  (SR 113, a moderate male bias).

#### The Classical Significance Testing Approach

A classical Fisherian approach focuses on a single hypothesis and the probability of observing the data or data more extreme were the hypothesis true. If  $X$  is a random variable for the number of boy births, the probability of observing  $x$  boys in a sample of  $n$  births can be modeled by the binomial distribution such that the probability of observing  $x$  or more boys is

$$P(X \geq x) = \sum_{i=x}^n \binom{n}{i} \theta^i (1-\theta)^{n-i}.$$

Given the hypothesis  $H$  that  $\theta = 0.50$ , the probability ( $p$ ) of observing 193 or more boys in a sample of 351 births is

$$p = \sum_{i=193}^{351} \binom{351}{i} (0.5)^i (1-0.5)^{351-i} = 0.035.$$

In contrast, the Neyman-Pearson approach constructs a rule for deciding between two alternative hypotheses (for example,  $H_1$  that  $\theta = 0.50$  and  $H_2$  that  $\theta \neq 0.5$ ). We divide the sample space for the random variable  $X$  into two regions: values of  $x$  for which we would decide

in favor of  $H_1$  versus those favoring  $H_2$ . (Note that our earlier Fisherian calculation was one-tailed; now  $H_2$  considers values of  $x$  either higher or lower than those predicted by  $H_1$ .) There are various methods for arriving at the critical values given  $\alpha$  (usually 0.05), the probability of a Type I error. Here we simply solve for the upper and lower of values of  $x$  for which  $p < 0.025$  ( $\alpha/2$ ). The resulting decision rule is to choose  $H_1$  if  $157 < x < 195$ , or choose  $H_2$  if  $x \leq 157$  or  $x \geq 195$ . Given  $x = 193$ , we should choose  $H_1$ ,  $\theta = 0.50$ .

In a standard merging of the Fisherian and Neyman-Pearson approaches,  $H_1$  would be labeled the null hypothesis and  $H_2$  the alternative; a two-tailed statistical test would produce  $p = 0.07$ . Although this value might be considered “marginally” significant, we would be unable to reject the null hypothesis that the population parameter  $\theta$  diverges from 0.50. The power ( $\beta$ ) of such a test to detect moderate differences (e.g.,  $\theta \pm 0.05$ ) is only 0.44 (see Cohen<sup>24</sup> for how to calculate power), so even if the actual population  $\theta = 0.55$ , we would fail to reject the incorrect null hypothesis roughly 56% of the time.

#### The Likelihood Approach

The likelihood approach calculates the statistical evidence in the data for one hypothesis relative to another. Consider the two hypotheses about the population parameter  $\theta$ ,  $H_1: \theta = 0.50$  and  $H_2: \theta = 0.53$ . Note that for illustrative purposes we are introducing these as a *priori* point hypotheses; in other words, they are not parameters estimated from the data. The likelihood of each hypothesis given the data is proportional to the probability of the data under the hypothesis.<sup>22</sup> (Because the proportionality constant is the

same for both models, this term drops out of the equation.) With the binomial distribution, the likelihood ratio for the two hypotheses is

$$\begin{aligned} \frac{L(H_2|X=x)}{L(H_1|X=x)} &= \frac{\binom{n}{x} \theta_2^x (1-\theta_2)^{n-x}}{\binom{n}{x} \theta_1^x (1-\theta_1)^{n-x}} \\ &= \frac{(0.53)^{193} (1-0.53)^{351-193}}{(0.50)^{193} (1-0.50)^{351-193}} = 4.35. \end{aligned}$$

In other words, the data indicate over four times the statistical evidence for  $H_2$  *vis-à-vis*  $H_1$ .

#### The Bayesian Approach

A Bayesian analysis uses new data to update one's prior beliefs in competing hypotheses. In the Kipsigis sex ratio example, consider two competing hypotheses about the population parameter  $\theta$ ,  $H_1: \theta = 0.50$  and  $H_2: \theta = 0.53$ . Based on knowledge of chromosomal sex determination and previous findings,<sup>46</sup> we will skew our prior beliefs in the hypotheses toward  $H_1$ . We might thus assign the priors as follows:  $P(H_1 \text{ true}) = 0.75$  and  $P(H_2 \text{ true}) = 0.25$ . (More detailed justifications of the prior estimates should accompany a complete Bayesian analysis.)

Bayes's theorem<sup>11</sup> states that

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Let  $A$  represent our hypotheses and  $B$  our data, such that  $P(A)$  represents our prior estimates and  $P(A|B)$  our posterior estimates. The binomial distribution provides the probabilities of observing the data under each prior hypothesis,  $P(B|A)$ , and the joint probability of observing the data under either hypothesis,  $P(B)$ . Substituting, the posterior probabilities for  $H_1$  is

$$P(H_1|x = 193) = \frac{P(x = 193|H_1)P(H_1)}{P(x = 193|H_1)P(H_1) + P(x = 193|H_2)P(H_2)}$$

$$P(H_1|x = 193) = \frac{\binom{n}{x} \theta_1^x (1 - \theta_1)^{n-x} P(H_1)}{\binom{n}{x} \theta_1^x (1 - \theta_1)^{n-x} P(H_1) + \binom{n}{x} \theta_2^x (1 - \theta_2)^{n-x} P(H_2)}$$

$$P(H_1|x = 193) = \frac{(0.50)^{193} (1 - 0.50)^{351-193} (0.75)}{(0.50)^{193} (1 - 0.50)^{351-193} (0.75) + (0.53)^{193} (1 - 0.53)^{351-193} (0.25)} = 0.41.$$

Similarly,

$$P(H_2|x = 193) = \frac{P(x = 193|H_2)P(H_2)}{P(x = 193|H_1)P(H_1) + P(x = 193|H_2)P(H_2)} = 0.59.$$

Thus, given the data, our belief in H1 has decreased from 0.75 to 0.41, and our belief in H2 has increased from 0.25 to 0.59.

A key philosophical component of the likelihood approach is that data are never considered as evidence for or against a single hypothesis in isolation, but only relative to one or more other hypotheses. The magnitude of the likelihood ratio (say, for  $H_1/H_2$ ) measures the strength of evidence for  $H_1$  over  $H_2$ . With likelihood statistics, it is also possible to calculate the probabilities of the data providing misleading evidence ( $m$ ) for one hypothesis over another or weak evidence ( $w$ ) that is unable to distinguish two hypotheses.<sup>22</sup>

The likelihood approach is advantageous because it places the focus on two or more hypotheses. Also, unlike  $P$ -values, likelihood ratios measure the strength of evidence in the data for the hypotheses relative to one another. Additional advantages include the property that the evidence does not change with the behavior of the person analyzing the data.<sup>22</sup> In the traditional approach,  $P$ -values should ideally be adjusted to take into account the amount of data snooping and multiple tests a researcher has done. With the likelihood approach, the data are what they are irrespective of the behavior of the researcher.

The likelihood approach does have disadvantages. Likelihood ratios compare what are known as point hypotheses, in which parameters take specific values, for example, the probability of an event is 0.6 versus

0.5. Tests based solely on likelihood ratios are not suited to composite hypotheses, which actually comprise many point hypotheses, for example, the probability of an event is greater than 0.6. Another complication arises when candidate models differ in the number of parameters and those parameters are being estimated from the data. Such cases call for using Akaike Information Criterion for model selection (see Box 3).

Another potential drawback of the likelihood approach is that it does not have strict cut-off points for interpreting data. In some areas, such as phylogenetic analyses, it is common practice to cross over to a significance testing approach and “test” likelihood ratios for significance using a Chi-squared distribution. Alternatively, various guidelines have been proposed for translating likelihood ratios into levels of strength of evidence or plausibility of one hypothesis over another.<sup>22,27</sup> For example, under one such guideline, likelihood ratios greater than ten would indicate strong evidence, ratios between three and ten would indicate moderate evidence, and ratios between one and three would indicate only weak evidence for the more likely hypothesis. Although convenient, a risk in applying such guidelines is that of falling into the same “reject” or “accept” mentality that has dominated classical significance testing.<sup>28,29</sup>

## The Bayesian Approach

A very different approach from either classical significance testing or likelihood inference is based on Bayes’s theorem.<sup>11,15</sup> The Bayesian approach emphasizes using data to update one’s prior belief in two or more competing hypotheses. Before data are gathered, one’s prior belief in each hypothesis is quantified as an estimated unconditional probability that the hypothesis is true. These hypotheses must be both mutually exclusive and all-inclusive, such that their probabilities sum to one. Using Bayes’s formula, the data and prior estimates are then combined to calculate posterior estimates of the probability that the hypotheses are true. Multiple hypotheses can then be evaluated without the inherent asymmetry between a null hypothesis and an alternative hypothesis.

One circumstance in which Bayesian analyses might be advantageous for biological anthropologists is when information from previous studies, beliefs, or expert opinions can be used to form prior estimates for hypotheses.<sup>30-32</sup> Although this leads to the common criticism that the method is inherently subjective, in fact it may represent a quantitative approach to something that we tend to do anyway: evaluate our hypotheses not only in light of our current data but also with respect to what we know or believe

**Box 2. Contrasting Chi-Squared and Likelihood Ratio Analyses of Simulated Data**

A simple simulation study (implemented in R<sup>53</sup>) illustrates the differences between classical and likelihood approaches. We start with two groups of mothers, one high status the other low status, and use a binomial probability distribution to draw 500 births for each group, resulting in a total sample of 1,000 babies of either sex. For each of three generating models, we generate 1,000 such samples of 1,000 births, setting the parameter values as follows: for M1  $\theta_h = 0.5$  and  $\theta_l = 0.5$ , for M2  $\theta_h = 0.52$  and  $\theta_l = 0.52$ , and for M3  $\theta_h = 0.57$  and  $\theta_l = 0.47$ . With the sample births in hand, we assess how often the classical significance testing and likelihood approaches correctly detect the generating model that produced the data. It is important to note that for the purposes of our simulation, the parameter values are fixed from the outset and used in calculating the likelihoods, whereas an actual empirical analysis would typically estimate parameter values from the data and potentially need to account for differences in the number of fitted parameters (see Box 3).

For the classical approach, we use a standard Chi-square test for independence of maternal status and offspring sex (typical of studies trying to test the TW hypothesis; see Bereczkei and Dunbar<sup>36</sup>). The null hypothesis is that the two variables are independent; the alternative is simply that they are not, evidence of which might be construed as supporting the TW hypothesis. For data from each gen-

erating model, we use a Chi-square test to test the null hypothesis and record the *P*-value.

For the likelihood approach, we use the same data to calculate likelihood ratios for the observed data under pairwise comparisons of the three models. For example, of those populations we generated using M3 (TW-like parameter values), how many would show scientific evidence for M3 over M2? Over M1? What is the strength of this evidence?

At first glance, the results from the Chi-square analyses appear to be reassuring (Table 1). When M1 or M2 generated the data, the null hypothesis (no interaction between sex ratio and maternal condition) is true, and in about 95% of samples simulated the

*P*-value was nonsignificant. In the case of M3, when an interaction was built into the generating model, the null hypothesis is not true, and the vast majority (87%) of samples yielded *P*-values less than 0.05. Note, however, the asymmetry between the models based on whether the model resembles the null or alternative hypothesis. Because the traditional 0.05 significance level is weighted to reduce Type I errors over Type II errors, Type II errors (failing to reject a false null hypothesis) were roughly three times more likely with M3. This number would be even higher with smaller samples or if the M3 generating parameter values were closer together (e.g.,  $\theta_h = 0.54$  and  $\theta_l = 0.50$ ). In other words, with repeated sam-

**TABLE 2. Likelihood Ratios for the Point Hypotheses Representing the Three Generating Models, Using the Same Samples as in Table 1**

		0 to <1/10	1/10 to <1/3	1/3 to <3	3 to <10	10+
<i>M1 True</i>	LR M1/M2	13	66	509	284	128
	LR M1/M3	9	16	65	63	847
<i>M2 True</i>	LR M2/M1	8	64	513	286	129
	LR M2/M3	8	19	89	77	807
<i>M3 True</i>	LR M3/M1	8	14	62	77	839
	LR M3/M2	15	13	84	93	795

**TABLE 1. Pearson's Chi-Squared Tests (with Yates's Continuity Correction) for 1,000 Samples per Generating (True) Model, with Simulated Data Used to Test the Null Hypothesis of No Difference Between High- and Low-Status Sex Ratios**

True Model	<i>P</i> ≤ 0.05	n.s.
M1	45 <sup>a</sup>	955
M2	47 <sup>a</sup>	953
M3	872	128 <sup>b</sup>

<sup>a</sup> Type I errors would be made in these cases.

<sup>b</sup> Type II errors would be made in these cases.

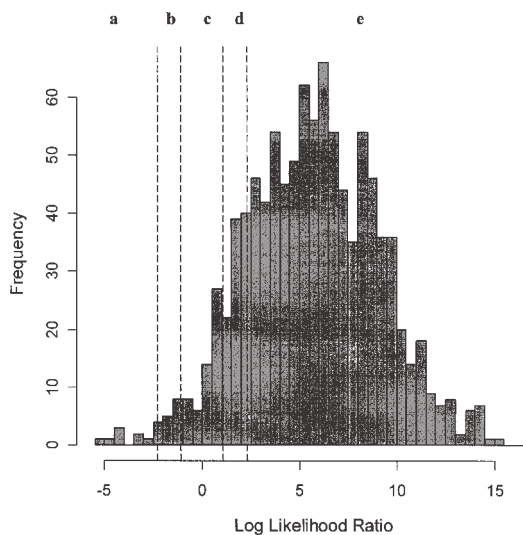


Figure 1. Log likelihood ratios of M3 to M1 when M3 was the generating model in *n* = 1,000 simulations. Dashed vertical lines correspond to the log likelihood ratio associated with likelihood ratios of 1/10, 1/3, 3, and 10, respectively. The simulations falling in areas a and e indicate strong evidence for M1 and M3, respectively; simulations in areas b and d indicate moderate evidence for M1 and M3, respectively; simulations in area c indicate weak evidence for either model.<sup>27</sup>

pling of a population, we would expect some nonsignificant results even if M3 were true, making it quite difficult to distinguish the hypotheses.

The likelihood ratios (LRs) consistently indicate evidence of the correct generating models (see Table 2 for a summary and Figure 1 for a closer look at the simulations when M3 generated the data). Looking at samples when M1 generated the data, and thus is the true model, 91% of the LRs comparing M1 to M3 are greater than 3, indicating at least moderate evidence for M1 over M3; 85% of the LRs are greater than 10, indicating strong evidence for M1 over M3. LRs comparing M1 to

M2 generally show more evidence for M1 than M2, but only 42% of LRs show at least moderate or strong evidence. The same pattern occurs when M2 is the true model. There is consistently (88%) at least moderate evidence for M2 over M3 and often (81%) that evidence is strong. The evidence for M2 versus M1 is again less definitive. When M3 is the true, there is consistent evidence for M3 over both M1 and M2. There is at least moderate evidence for M3 over M1 and M2 92% and 89% of the time, respectively; that evidence is strong 84% and 80% of the time, respectively.

Because the magnitudes of LRs give a measure of the strength of

the evidence, this type of analysis offers much more information than do Chi-square tests, which offered no straightforward way to interpret the nonsignificant results. The likelihood ratio analysis also avoids the asymmetry problem of the Chi-square analyses. When M3 is true, the LRs indicate the same strong evidence for it over M1 and M2 that is displayed when M1 or M2 are true and they are compared to M3. Because M1 and M2 are quite similar, the LRs do not distinguish between them very well, but they do provide a quantifiable way to compare the models. In contrast, a Chi-squared test is unable to distinguish them at all.

based on previous research. We will not develop our presentation of Bayesian inference to the same extent that we do the classical and likelihood approaches because our primary focus is on how the latter facilitate comparing multiple alternative hypotheses.

### SEX RATIO THEORY AND SPECIFYING COMPETING MODELS

Using sex ratio theory to illustrate the philosophical and statistical approaches in action, we now review several well-known hypotheses for sex ratio variation and show how they can be formalized as quantitative models. A fundamental problem faced by parents is how to schedule and allocate investment in offspring.<sup>33</sup> Offspring are expected to vary from one another in many respects, such as health, help to parents, and reproductive potential. Consequently, parents might benefit by investing differently in offspring according to such qualities. As a correlate of such qualities, offspring sex may have a major impact on the expected fitness returns on parental investment. Sex-biased variation has been reported in many societies for variables like birth sex ratios, infanticide, childhood disease and mortality risk, lactation and weaning, education, marriage payments, residence rules, and inheritance patterns.<sup>34</sup>

Many evolutionary hypotheses attempt to explain biases in birth sex ratios and subsequent parental investment. We focus on three fundamentals. The unifying core of these hypotheses is that biases stem from parental behavior to maximize their expected inclusive fitness. The hypotheses vary, however, in predicting how parents achieve this, given the costs, benefits, and constraints on investing in offspring of either sex. Fisher proposed that natural selection will lead to equal investment by parents in offspring of either sex.<sup>35</sup> Conditions where the offspring cost the same and have equal returns should lead to an equal sex ratio (Hypothesis 1: Fisher's Equal Production). If one sex costs less, however, parents should produce more of that sex (Hypothesis 2: Fisher's Equal Expenditure). Trivers and Willard<sup>5</sup> predict an interaction between parental condition and parental investment, with parents in better condition investing more in offspring of the sex with greater reproductive potential (Hypothesis 3: The TW Hypothesis).

Given empirical results bearing on these hypotheses, we might first simply describe the data; for example, what are the sample sex ratios? To evaluate the hypotheses, however, we want to make inferences about the underlying populations that produced the data. To do this, we need

to translate the verbal hypotheses into quantitative models. A well-known quantitative model for sex ratios is the binomial probability distribution (see Box 1). One way to contrast the three models is to divide the data into births from high-status mothers ( $h$ ) versus low-status mothers ( $l$ ), two subpopulations each with its own probability distribution for producing boy babies. The only parameter in the model that will vary between subpopulations is  $\theta$ , the probability of producing a boy.

The next step is to specify the models in more precise terms so they can be distinguished from one another. Hypothesis 1 predicts an equal sex ratio, so for Model 1 we can write:

$$M1: \theta_h = \theta_l = 0.5.$$

Note that this is a statistical point hypothesis, where the parameter takes on a single value.

For Model 2, the sex ratios in the two subpopulations are equal to each other, but differ from 0.5. At a minimum, we can write:

$$M2: \theta_h = \theta_l \neq 0.5.$$

Although this clearly distinguishes M2 from M1, left at this stage M2 is a composite hypothesis, encompassing a range of values for the para-

### Box 3. Model Selection Using Akaike Information Criterion

Various model selection approaches exist for concurrently comparing any number of models to a data set and measuring the relative support the data give each model.<sup>15,29,54,55</sup> Here we demonstrate model selection based on Akaike Information Criterion (AIC), a robust approach that is increasingly being used in other disciplines.<sup>54</sup> The AIC approach attempts to identify models that optimally describe the information in the data while minimizing the number of fitted parameters.<sup>54</sup> Adding fitted parameters always improves the fit of a model to the data, but overly complex models are harder to understand, often yield inferior predictions due to measurement errors in the parameters, and may overfit the data by explaining noise.<sup>56</sup> A model selection approach based on AIC can replace common approaches like stepwise regression that are used to choose model variables but have been criticized for inconsistent results and a focus on only the “best” model without any measure of model uncertainty.<sup>57</sup> In addition, unlike likelihood ratio tests, AIC can be used to compare non-nested models.

The basic steps in an AIC analysis are as follows: 1) identify the alternative models, including the number of parameters in each; 2) determine the best parameters for fitting the data to each model (often the maximum likelihood estimates); 3) given these parameter estimates, calculate the likelihood of the observed data associated with each model; 4) use the likelihoods and number of parameters to calculate AIC values for each model; and 5) rank the models in ascending order of these values. The best-supported model will have the lowest AIC value. We demonstrate these steps using childhood mortality data from the Kipsigis of Kenya.<sup>40</sup> Table 1 reports the number of deaths per group, classified by sex and parental wealth, for children aged 0–1 years. We use the data to arbitrate among four models that differ in whether mortality rates ( $q$ ) are inde-

**TABLE 1. Kipsigis Childhood Mortality (Ages 0-1 years)**

	Lower Wealth	Upper Wealth	Total
Male	19/167	15/193	34/360
Female	15/185	14/158	29/343
Total	34/352	29/351	63/703

pendent of sex and wealth (M1) or depend on sex (M2), wealth (M3), or both (M4) (Table 2). M4, in which mortality varies with both sex and wealth, is most representative of the TW hypothesis.

For each model, we calculate the values of  $q$  that are most likely to produce the data. In this case, the binomial distribution models the probability of observing a certain number of deaths ( $x$ ) out of the total number of children ( $n$ ), given the mortality rate ( $q$ ) in question. The likelihood ( $L$ ) of a model given the data is proportional to the probability of the data given the model (a proportionality constant across the models can be ignored). The AIC for model  $i$  is defined as

$$AIC_i = -2\log L_i + 2K_i,$$

where  $L_i$  is the likelihood of model  $i$  given the data, and  $K_i$  is the number of fitted parameters in the model. The models are ranked by their AIC values, with the lowest AIC being the optimal model among those considered.<sup>13</sup> The absolute magnitudes of AICs have no meaning since they depend on sample sizes and how

one has defined the problem, but the differences in AICs are informative. It is convenient to calculate the difference in AICs relative to the lowest AIC, defined as

$$\Delta_i = AIC_i - \min(AIC).$$

There is no critical value at which an AIC difference is “significant,” but as a general rule an AIC difference of 4 to 7 means a model has considerably less support than the best-supported model, while an AIC difference greater than 10 means the model has essentially no support.<sup>54</sup>

In this case M1, in which mortality is independent of both sex and wealth, appears to be the optimal model, and M4, the TW model, receives the least support (see Table 2). The AIC differences suggest that M4 has considerably less support than does M1, but the support for M1, M2, and M3 is comparable. We did similar analyses on Kipsigis childhood mortality from ages 1 to 5 years and found that M2, in which mortality is dependent only on sex, was the best-supported model. In this case, however, none of the models had considerably less support than any of the others, including M4. Taken as a whole, the AIC approach suggests that simple models may do just as well or better than a more complicated TW model in predicting Kipsigis childhood mortality.

To learn more about AIC, we strongly recommend Hilborn and Mangel<sup>11</sup> and Burnham and Anderson.<sup>54</sup> For a detailed anthropological

**TABLE 2. AIC comparison of Models of Kipsigis Mortality Ages 0-1**

Model	Mortality	K	AIC	$\Delta$
M1	$q_{all} = 63/703$	1	21.65	0
M2	$q_{male} = 34/360$ $q_{female} = 29/343$	2	23.43	1.78
M3	$q_{lower} = 34/352$ $q_{upper} = 29/351$	2	23.22	1.57
M4	$q_{ml} = 19/167$ $q_{mu} = 15/193$ $q_{fl} = 15/185$ $q_{fu} = 14/158$	4	26.02	4.37

illustration of a likelihood-based analysis using AIC, see Handcock and Jones's<sup>58</sup> study of disease transmission across sexual networks. The main impediment to using AIC is finding the maximum likelihood estimates of parameters and the associated likelihood of the observed data for the given model. However, it is quite simple to use

AIC analyses for linear models because the residual sum of squares from least squares regression or ANOVA analyses can be converted into AIC values.

$$AIC_i = n \log(\hat{\sigma}_i^2) + 2K_i \text{ with } \hat{\sigma}_i^2 = \frac{\text{Residual Sum of Squares}}{n}$$

The number of fitted parameters ( $K$ ) should include the intercept and  $\sigma_i^2$  which are being estimated. Excel tools are available for making this computation.<sup>59</sup> For nonlinear models, statistical packages such as R, S+, and Systat, can produce AIC values. For complex cases, likelihoods can be found through simulations.<sup>60,61</sup>

meter  $\theta$ , not all of which would support Hypothesis 2 of equal expenditure. In fact, M2 predicts that  $\theta$  would be proportional to the relative costs of producing boys and girls such that  $\theta = 1/(1 + c)$ , where  $c$  is the cost of producing boys relative to girls. Typically,  $c$  is left unmeasured and instead simply estimated as greater than or less than one, so that:

M2:  $\theta_h = \theta_l$ , greater than 0.5 if  $c < 1$  (boys cost less than girls), or

M2:  $\theta_h = \theta_l$ , less than 0.5 if  $c > 1$  (boys cost more than girls).

Model 3 is the only model that predicts that the sex ratios differ from equality but also that the subpopulations will differ from each other. At a minimum, we can write M3:  $\theta_h \neq \theta_l$ . Note, however, that as with M2, this composite hypothesis covers many conceivable values that would run counter to Hypothesis 3. We can therefore specify:

M3:  $\theta_h > \theta_l$  if higher parental status increases the reproductive potential of boys more than girls, or

M3:  $\theta_h < \theta_l$  if higher parental status increases the reproductive potential of girls more than boys.

With these general specifications of M1, M2, and M3 in hand, we are now better able to evaluate which model would give rise to data similar to those observed. In Box 2, we present results from computer simulations that create populations under alternative models for sex ratio variation. We analyze these population sex ratios using classical and likeli-

hood methods to compare the alternative approaches.

### HUMAN BEHAVIORAL ECOLOGY: PARENTAL INVESTMENT LITERATURE

We now consider specific examples of the statistical and philosophical approaches employed by human behavioral ecologists, focusing on a set of twelve papers bearing on the TW hypothesis.<sup>3,25,36-45</sup> Although the papers vary extensively in how heavily they use statistics in making arguments, from quite minimal use<sup>43</sup> to more detailed,<sup>25</sup> they all use classical significance testing approaches such as Chi-square tests, ANOVAs, and regression. An average of 48  $P$ -values appear in each paper (range 4–244), reported in a variety of forms: as an exact number, as greater or less than some cut-off, or simply as nonsignificant.

About half of the papers contain interpretations of  $P$ -values beyond simply reporting the exact value, as in a Fisherian approach, or the critical region in which the value falls, as in the NP approach. These interpretations include using adjectives such as “highly,” “marginally,” and “not quite” to describe the significance level, or using different numbers of asterisks to distinguish  $P$ -values of different quantities. The problem is not in the use of  $P$ -values, adjectives, or asterisks *per se*, but in the implied link between the magnitude of the  $P$ -value and the biological (substantive) significance of the result. No direct link exists. Very large sample sizes, for instance, will often lead to extremely small  $P$ -values despite small effect sizes.<sup>41</sup> The inconsistencies associated with the 0.05 cut-off

for statistical significance are also apparent in these papers. In Cronk<sup>3</sup> (Table 10.2), for instance, patterns associated with one-tailed  $P$ -values of 0.03 and 0.07 are given merit, while one with a  $P$ -value of 0.098 (going against the predicted pattern) is discounted.

Another problem with relying heavily on  $P$ -values is how to interpret nonsignificant results. Nearly all of the studies assert the null hypothesis for some nonsignificant results. For instance, Voland and coworkers<sup>39</sup> analyzed how the correlation between sex-biased infant mortality and population growth rates in historical Germany varied with time. They concluded that there is evidence of a “one-generation delay” between the two variables. This inference, however, rests on nonsignificant correlations for shorter (0 years) and longer time intervals (60 years) and a significant correlation for an intermediate time interval (30 years). Essentially, the authors had to assert the null hypothesis for two of the three tests to draw this conclusion. To make such arguments based on nonsignificant results, one first needs to show that the test had enough power to reliably detect an underlying population effect if one existed. Otherwise, such interpretations are unsound (see simulations in Box 2). Only one of the twelve HBE studies supports their interpretations of nonsignificant results with explicit power analyses.<sup>25</sup>

Closely linked to how we use statistical results is the question of how human behavioral ecologists use hypotheses in their research. All twelve studies are focused on one or more measures of parental investment but deal with hypotheses in a variety of ways. The majority of studies are not designed to test the “hard core” of

HBE, for instance the applicability of natural selection or parental investment theory to understanding human behavior. (In an exception, Freese and Powell<sup>41</sup> use the TW hypothesis as a representation of the entire HBE enterprise, with the hypothesis being the sole evolutionary explanation weighed against sociological hypotheses.) Some studies focus primarily on testing the TW hypothesis,<sup>43</sup> but others give considerable attention to alternative hypotheses.<sup>25,40</sup> The latter group includes studies that essentially assume that the TW hypothesis must be true based on evolutionary logic, then focus on subsidiary models about the contexts and directions in which TW “effects” are expressed.<sup>39</sup>

According to Winterhalder and Smith's<sup>1</sup> ideal of HBE research, the TW hypothesis must be formalized into “testable” models with specific goals, currencies, constraints, and decisions. However, when the TW hypothesis is expressed quantitatively, as with the binomial model, it becomes apparent that some studies fall short of evaluating the full hypothesis even when they claim to do so. Irons,<sup>43</sup> for example, uses a combination of ethnographic and demographic data from the Yomut Turkmen “to test the Trivers-Willard hypothesis.” The hypothesis predicts that Yomut parents, who are at the top of their local marriage system, should invest relatively more in sons than do the lower-status groups with whom they intermarry. The empirical data reflect only what the upper-status parents are doing, however, so no direct test of the TW hypothesis is possible. Similarly, Cronk<sup>3</sup> focuses primarily on the Mukogodo of Kenya, in this case a low-status group, and predicts greater investment in daughters than sons. In both examples, even if the data were to match the biased investment predicted by the TW hypothesis, the key prediction, an *interaction* between status and investment, is not subjected to a quantitative test; that is, only one of two parameters in the binomial model is considered.

Even when the data are available to test a more complete quantitative model for a hypothesis, the classical statistical approach often makes it

difficult to ascertain whether the evidence in the data support the model or not. Bereczkei and Dunbar's<sup>36</sup> Hungarian sample is divided by parental condition, in this case ethnicity (Gypsy versus non-Gypsy), allowing the estimation of both parameters in the TW binomial model. They report a female-biased sex ratio in the sample of Gypsy births ( $\theta = 0.47$ ) and a male-bias in the non-Gypsy sample ( $\theta = 0.53$ ). These biases seem to suggest a TW effect, but classical statistics are inconclusive. The paper reports  $p = 0.026$ , but we find, and the authors concur (personal communication) that the correct value is  $p = 0.095$ . Although this sometimes falls in the “marginally significant” category, it does not meet the more stringent  $p < 0.05$ .

Using a likelihood approach gives us greater insight into the statistical evidence in the data for one model versus another model. Let H1 represent equal investment in the two sexes by both ethnicities ( $\theta = 0.5$  for both Gypsies and non-Gypsies, estimated from the pooled sample data), and let H2 represent the TW hypothesis with the more affluent ethnic group investing more in males ( $\theta = 0.47$  for Gypsies and  $\theta = 0.53$  for non-Gypsies, estimated separately for each group). In this case, the likelihood ratio (LR; see box 1) for H2 over H1 is 4.5; the hypotheses, however, have a different number of fitted parameters. One way to compare nested models with different numbers of parameters is to cross over into a classical statistical approach and use a likelihood ratio test (see Box 3 for a different approach). The quantity  $-2 \log LR$  has a Chi-squared distribution.<sup>11</sup> When two nested models differ by one fitted parameter, this quantity must be greater than 3.84 to reach statistical significance at  $p < 0.05$ . For this example,  $-2 \log LR$  equals 3.0, so H2 is not supported over H1 despite a positive likelihood ratio.

We have already noted the prevalence of asserting the null hypothesis in the face of nonsignificant results. A likelihood analysis helps avoid this tendency by shifting our focus from two asymmetric hypotheses, a null and alternative, to multiple hypothe-

ses. Such a shift in focus could have a major impact on the direction our research takes, including interpretations of results and planning future analyses. Borgerhoff Mulder<sup>46</sup> considered parental investment by the Kipsigis of Kenya. The paper includes analyses that detected no interaction effect of wealth and sex on early childhood mortality and cites previous findings of “no Trivers-Willard bias” in secondary (birth) sex ratios. These null findings, reported in both the abstract and discussion of the paper, put Borgerhoff Mulder in the difficult position of making sense of the “inconsistent findings of the present study (no wealth/sex interactions with respect to mortality, but a marked effect with respect to education).”

We applied a likelihood approach to ask whether the Kipsigis data in fact supply scientific evidence for the *lack* of a TW effect, the conclusion being drawn from classical statistical analyses. Using the Kipsigis early childhood mortality data, graciously provided by Borgerhoff Mulder, we considered the basic models of variation presented previously for sex ratios. In line with Borgerhoff Mulder's conclusions, our analyses show that simpler models may be preferable to a TW model of childhood mortality, but not, however, that the TW model is untenable (see Box 3). In addition to reanalyzing the mortality data, we used both classical and likelihood statistics to examine secondary sex ratios in the 1998 data set and discovered that, in contrast to Borgerhoff Mulder's earlier results,<sup>46</sup> there was sex ratio variation consistent with a TW effect. It is possible that Borgerhoff Mulder's previous tests had modest power for detecting such effects and that the null hypothesis, although not rejected, was also not supported by the earlier data.

## DISCUSSION

Statistics, when used appropriately, offer us some objective means for identifying the scientific evidence in our data. Even if statistical significance testing is misused,<sup>52</sup> such practices may be an improvement over our human tendency to focus on “anecdotes, intuitions, and sensa-

tional and unlikely events.”<sup>16</sup> Still, as the examples in this paper reveal, we need to do better. To this end, statistical methods that allow thoughtful and quantitative weighing of several alternative hypotheses for a given phenomenon may be a more productive route than the classical methods that focus so much on significance testing and rejecting null hypotheses, which are often trivial in the first place. A likelihood approach to statistical inference may not always provide as clear-cut answers as a classical approach seems to, but it can give us a more thorough understanding of the evidence in our data for alternative models and a more realistic picture of the uncertainty that remains in comparisons of hypotheses.

As we noted at the outset of this paper, one way human behavioral ecologists distinguish themselves from other anthropological fields is by their emphasis on quantitative methods and an H-D research approach. The classical significance testing statistical approach, however, limits how effectively alternative hypotheses can be compared. In contrast, a likelihood approach to statistical inference better facilitates the comparison of alternative models. Research would also benefit from the more subtle and more realistic conclusions that can be drawn from the likelihood statistical approach.

We focused on sex ratio hypotheses, but the described approaches could be applied to any number of problems in biological anthropology. For example, in introducing a special issue of the *American Journal of Primatology* on advances in primate behavioral endocrinology, Strier and Ziegler<sup>47</sup> write that they “anticipate the emergence of a dynamic new field in which comparative models of primate behavioral endocrinology contribute new perspectives on primates.” In this field and primatology in general, small sample sizes are often inevitable, and a likelihood approach to inference could give researchers a better handle on the actual evidence in their data than does trying to interpret nonsignificant results from tests with low power. Another field ripe for a likeli-

hood approach to inference is the study of cultural macroevolution,<sup>48</sup> where a key debate focuses on the relative roles of vertical transmission, horizontal transmission, and innovation in explaining the distribution of cultural traits across human societies. Previous studies have typically focused attention on just one transmission mechanism or have attempted to identify prominent mechanisms through the inappropriate comparisons of *P*-values across analyses and through the assertion of null hypotheses.<sup>49,50</sup>

---

**A likelihood approach to statistical inference may not always provide as clear-cut answers as a classical approach seems to, but it can give us a more thorough understanding of the evidence in our data for alternative models and a more realistic picture of the uncertainty that remains in comparisons of hypotheses.**

---

In reality, of course, selecting appropriate statistical analyses and modes of inference must take into account practical as well as theoretical concerns. Circumstances may often necessitate having the flexibility and knowledge to move between a variety of inferential approaches. Chatfield<sup>51</sup> outlines what he and others call “pragmatic statistical inference,” which explicitly recognizes the importance of context and data attributes in formulating and evaluating statistical models. This is not to say that all statistical philosophies are equally sound, but rather that holding on dogmatically to one approach to the exclusion of others may not be fruitful or necessary. Moreover, all modes of statistical inference have the potential for misuse and all are

limited from the outset by data quality. As a safeguard, Chatfield argues that pragmatic statistical inference should always follow a thorough initial data analysis (IDA), which includes graphing the data and looking at summary statistics, as well as screening for outliers and missing or incorrectly entered data. “IDA is vital to understand the data and is sometimes all that is needed if, for example, the results are very clear cut or reveal such poor data quality that a more sophisticated model-based analysis cannot be justified.”<sup>51</sup>

Our critique of classical statistical methods, focused on *P*-values and hypothesis testing, is not new. In fields such as psychology and ecology, the debate already has a long history, and many people have advocated the complete abandonment of *P*-values in data analysis and interpretation. Even if such a move is fully justified based on statistical theory and common misuses of classical methods, change is slow for a variety of reasons, such as people’s statistical training, journal publication standards, and reviewer expectations. Meanwhile, adding power analyses to classical approaches would improve the statistical inferences that researchers could make within the realm of null hypothesis testing. In addition, looking at the actual effect sizes of statistically significant results would tell us more about the biological significance of our data than trying to read meaning into the magnitudes of *P*-values.

#### ACKNOWLEDGMENTS

We thank Eric Alden Smith, Jeremy Brooks, Subhash Lele, Marc Mangel, Richard McElreath, Mark Taper, Bruce Winterhalder, and several anonymous reviewers for their contributions. Special recognition goes to Monique Borgerhoff Mulder for sharing her Kipsigis data with us, to Mark Grote for his statistical expertise, and to both for their encouragement. This work was initiated while we were postdoctoral associates for the Evidence Project Working Group supported by the National Center for Ecological Analysis and Synthesis, funded by NSF (Grant #DEB-0553768), the University of

California, Santa Barbara, and the State of California.

## REFERENCES

- 1 Winterhalder B, Smith EA. 2000. Analyzing adaptive strategies: human behavioral ecology at twenty-five. *Evol Anthropol* 9:51–72.
- 2 Blurton Jones NG. 1986. Towards hypothetico-deductive anthropology? *Rev Anthropol* 13: 151–160.
- 3 Cronk L. 2000. Female-biased parental investment and growth performance among the Mukogodo. In: Cronk L, Chagnon N, Irons W, editors. *Adaptation and human behavior: an anthropological perspective*. New York: Aldine de Gruyter. p 203–222.
- 4 Borgerhoff Mulder M. 1991. Human behavioral ecology. In: Krebs JR, Davies NB, editors. *Behavioral ecology: an evolutionary approach*. Oxford: Blackwell Scientific. p 69–98.
- 5 Trivers RL, Willard DE. 1973. Natural selection of parental ability to vary the sex ratio of offspring. *Science* 179:90–92.
- 6 Ghiselin M. 1969. *The triumph of the Darwinian method*. University of California Press.
- 7 Laudan L. 1981. *Science and hypothesis*. Dordrecht: Reidel.
- 8 Popper KR. 1959. *The logic of scientific discovery*. London: Hutchinson.
- 9 Platt JR. 1964. Strong inference. *Science* 146: 347–353.
- 10 Ketelaar T, Ellis BJ. 2000. Are evolutionary explanations unfalsifiable? Evolutionary psychology and the Lakatosian philosophy of science. *Psychol Inquiry* 1:1–21.
- 11 Hilborn R, Mangel M. 1997. *The ecological detective: confronting models with data*. Princeton: Princeton University Press.
- 12 Gray JP. 2000. Twenty years of evolutionary biology and human social behavior: where are we now? In: Cronk L, Chagnon N, Irons W, editors. *Adaptation and human behavior: an anthropological perspective*. New York: Aldine de Gruyter. p 475–495.
- 13 Anderson DR, Burnham KP, Thompson WL. 2000. Null hypothesis testing: problems, prevalence, and an alternative. *J Wildlife Man* 64: 912–923.
- 14 Pigliucci M. 2002. Are ecology and evolutionary biology “soft” sciences? *Ann Zool Fennici* 39:87–98.
- 15 Taper M, Lele S, editors. 2004. *The nature of scientific evidence: statistical, philosophical, and empirical considerations*. Chicago: The University of Chicago Press.
- 16 Scarr S. 1997. Rules of evidence: a larger context for the statistical debate. *Psychological Science* 8:16–17.
- 17 Utts J. 1999. *Seeing through statistics*. Pacific Grove, CA: Brooks-Cole/Duxbury Press.
- 18 Edwards AWF. 1972. *Likelihood*. Cambridge: Cambridge University Press.
- 19 Cox DR. 1958. Some problems connected with statistical inference. *Ann Math Statistics* 29:357–372.
- 20 Gigerenzer G. 1993. The superego, the ego, and the id in statistical reasoning. In: Keren G, Lewis C, editors. *A handbook for data analysis in the behavioral sciences: methodological issues*. Hillsdale, NJ: Lawrence Erlbaum Associates. p 311–339.
- 21 Oakes ML. 1986. *Statistical inference: a commentary for the social and behavioral sciences*. New York: John Wiley & Sons.
- 22 Royall R. 1997. *Statistical evidence: a likelihood paradigm*. London: Chapman & Hall.
- 23 Cohen J. 1994. The earth is round ( $p < .05$ ). *Am Psychol* 49:997–1003.
- 24 Cohen J. 1988. *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.
- 25 Keller MC, Neese RM, Hofferth S. 2001. The Trivers-Willard hypothesis of parental investment: no effect in the contemporary United States. *Evol Hum Behav* 22:343–360.
- 26 Byers SN. 2000. Testing Type II error rates in biological anthropology. *Am J Phys Anthropol* 111:283–289.
- 27 Reid N. 2000. Likelihood. *J Am Statistical Assoc* 95:1335–1340.
- 28 Pigliucci M. 2002. Hypothesis testing and the nature of skeptical investigations. *Skeptical Inquirer* 26:27–30.
- 29 Anderson DR, Burnham KR. 2002. Avoiding pitfalls when using information-theoretic methods. *J Wildlife Man* 66:912–918.
- 30 Lele S. 2004. Elicit data, not prior: on using expert opinion in ecological studies. In: Taper M, Lele S, editors. *The nature of scientific evidence: statistical, philosophical, and empirical considerations*. The University of Chicago Press. p 410–423.
- 31 Konigsberg LW, Hens SM, Meadows Jantz L, Jungers WL. 1998. Stature estimation and calibration: Bayesian and maximum likelihood perspectives in physical anthropology. *Yearbook Phys Anthropol* 41:65–92.
- 32 Gowland RL, Chamberlain AT. 2002. A Bayesian approach to ageing perinatal skeletal material from archaeological sites: implications for the evidence for infanticide in Roman-Britain. *J Archaeol Sci* 29:677–685.
- 33 Clutton-Brock TH. 1991. *The evolution of parental care*. Princeton: Princeton University Press.
- 34 Hrdy SB. 1999. *Mother nature: maternal instincts and how they shape the human species*. New York: Ballantine Books.
- 35 Fisher RA. 1930. *The genetical theory of natural selection*. Oxford: Clarendon Press.
- 36 Bereczkei T, Dunbar RIM. 1997. Female-biased reproductive strategies in a Hungarian Gypsy population. *Proc R Soc Lond B* 264:17–22.
- 37 Chacon-Puignau GC, Klaus J. 1996. Sex ratio at birth deviations in modern Venezuela: the Trivers-Willard effect. *Soc Biol* 43:257–270.
- 38 Strickland SS, Tuffrey VR. 1997. Parental investment theory and birth sex ratios in Nepal. *J Biosoc Sci* 29:283–295.
- 39 Voland E, Dunbar RIM, Engel C, Stephan P. 1997. Population increase and sex-biased parental investment in humans: evidence from 18<sup>th</sup>- and 19<sup>th</sup>-century Germany. *Curr Anthropol* 38: 129–135.
- 40 Borgerhoff Mulder M. 1998. Brothers and sisters: how sibling interactions affect optimal parental allocations. *Hum Nat* 9:119–162.
- 41 Freese J, Powell B. 1999. Sociobiology, status, and parental investment in sons and daughters: testing the Trivers-Willard hypothesis. *Am J Soc* 106:1704–1743.
- 42 Scott S, Duncan CJ. 1999. Reproductive strategies and sex-biased investment: suggested roles of breast-feeding and wet-nursing. *Hum Nat* 10:85–108.
- 43 Irons W. 2000. Why do the Yomut raise more sons than daughters? In: Cronk L, Chagnon N, Irons W, editors. *Adaptation and human behavior: an anthropological perspective*. New York: Aldine de Gruyter. p 223–236.
- 44 Hagen EH, Hames RB, Craig NM, Lauer MT, Price ME. 2001. Parental investment and child health in a Yanomamö village suffering short-term food stress. *J Biosoc Sci* 33:503–528.
- 45 Koziel S, Ulijaszek SJ. 2001. Waiting for Trivers and Willard: do the rich really favor sons? *Am J Phys Anthropol* 115:71–79.
- 46 Borgerhoff Mulder M. 1989. Reproductive consequences of sex-biased inheritance for the Kipsigis. In: Standen V, Foley RA, editors. *Comparative sociobiology*. Oxford: Blackwell Scientific. p 405–427.
- 47 Strier KB, Ziegler TE. 2005. Introduction: advances in field-based studies of primate behavioral endocrinology. *Am J Primatol* 67:1–4.
- 48 Borgerhoff Mulder M, Nunn CL, Towner MC. 2006. Macroevolutionary studies of cultural trait transmission. *Evol Anthropol* 15:52–64.
- 49 Guglielmino CR, Viganotti C, Hewlett B, Cavalli-Sforza LL. 1995. Cultural variation in Africa: role of mechanisms of transmission and adaptation. *Proc Natl Acad Sciences USA* 92: 7585–7589.
- 50 Hewlett BA, de Silvertri A, Guglielmino CR. 2002. Semes and genes in Africa. *Curr Anthropol* 43:313–321.
- 51 Chatfield C. 2002. Confessions of a pragmatic statistician. *Statistician* 51:1–20.
- 52 Yoccoz NG. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bull ESA*. 72:106–111.
- 53 R Development Core Team 2006. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- 54 Burnham KP, Anderson DR. 2002. *Model selection and multimodel inference: a practical information-theoretic approach*. New York: Springer-Verlag.
- 55 Pitt MA, Myung IJ. 2002. When a good fit can be bad. *Trends Cogn Sci* 6:421–425.
- 56 Mikkelsen GM. 2001. Complexity and verisimilitude: realism for ecology. *Biol Philos* 16:533–546.
- 57 Judd CM, McClelland GH. 1989. *Data analysis: a model comparison approach*. New York: Harcourt Brace Jovanovich.
- 58 Handcock MS, Jones JH. 2004. Likelihood-based inference for stochastic models of sexual network formation. *Theor Popul Biol* 65:413–422.
- 59 Doak D, Pollock J, Rose A, Knowlton J, Booth M, Parker I. 2006. <http://bio.research.ucsc.edu/people/doaklab/natconserv/index.html>. Tool J. Using regression and ANOVA outputs to address management decision-making: Information-theoretic (AIC) interpretations of data analyses. Santa Cruz: University of California.
- 60 Geyer CJ. 1991. Markov chain Monte Carlo maximum likelihood. In: *Computer science and statistics. Proc. 23<sup>rd</sup> Symposium Interface*, p 156–163.
- 61 Luttbeg B, Langen TA. 2004. Comparing alternative models to empirical data: cognitive models of Western scrub-jay foraging behavior. *Am Nat* 163:263–276.