Evidential Statistics as a statistical modern synthesis to support 21st century science.

Mark L. Taper[1, 3] and José Miguel Ponciano[2]

[1,3]*Ecology Department, Montana State University, Bozeman, MT, 59717-3460, USA*

[2]*Department of Biology, University of Florida, Gainesville, FL, 32611-8525, USA*

[3]Corresponding author, e-mail: markltaper@gmail.com

Abstract

During the 20th century, population ecology and science in general relied on two very different statistical paradigms to solve its inferential problems: error statistics (also referred to as classical statistics and frequentist statistics) and Bayesian statistics. A great deal of good science was done using these tools, but both schools suffer from technical and philosophical difficulties. At the turning of the 21st century (Royall, 1997, Lele 2004), evidential statistics emerged as a seriously contending paradigm. Drawing on and refining elements from error statistics, likelihoodism, Bayesian statistics, information criteria, and robust methods, evidential statistics is a statistical modern synthesis that smoothly incorporates model identification, model uncertainty, model comparison, parameter estimation, parameter uncertainty, pre-data control of error, and post-data strength of evidence into a single coherent framework. We argue that evidential statistics is currently the most effective statistical paradigm to support 21st century science. Despite the power of the evidential paradigm, we think that there is no substitute for learning how to clarify scientific arguments with statistical arguments. In this paper we sketch and relate the conceptual bases of error statistics, Bayesian statistics and evidential statistics. We also discuss a number of misconceptions about the paradigms that have hindered practitioners, as well as some real problems with the error and Bayesian statistical paradigms solved by evidential statistics.

Keywords: evidential statistics; error statistics: Bayesian statistics, information criteria; likelihoodism; statistical inference

Introduction

We were very pleased when we were invited to present at the "Statistics in Population Ecology" symposium. The use of statistics in science is a topic dear to both of our hearts and has been the focus of both of our research programs for years. We were humbled and frightened by the later request, that as the first presentation in the symposium we should give an overview introducing not only our field of Evidential Statistics, but also Error Statistics, and Bayesian Statistics. We are well aware of the hubris of trying to define essentially all of statistics in a single essay, but we ask the readers' indulgence because we are just following instructions.

These are our ideas that we have come to through decades of struggling to make sense of ecology through statistics. It will be clear from the other papers in this special issue of Population Ecology that there are other viewpoints on the use of statistics in ecology. Nevertheless, we offer these ideas up to the readers in the hope that they may help some with their own struggle to support their scientific endeavors through statistics.

Technological tools have historically expanded the horizons of science. The telescope gave us the skies. The microscope gave us the world's fine structure. The cyclotron gave us the structure of matter. In our opinion, perhaps the ultimate technological tool helping scientists see nature is statistics. As we will see, it is not an exaggeration to state that statistics gives us all of science. Although mathematics, and in particular, probability and statistics, have been recognized many times as a fundamental tool ecologists can use to learn from the natural world (Underwood 1997, Cohen 2004), our central tenet is that more than just technical facility, an effective use of this tool

requires learning to filter scientific arguments through the sieve of statistical argumentation.

Despite its enormous power, there is great confusion about statistics among ecologists, philosophers and even statisticians. This confusion is terminological, methodological, and philosophical. As the statistician Richard Royall (2004) has said: "Statistics today is in a conceptual and theoretical mess." That doesn't mean that statistics isn't helpful, nor does it mean that scientific progress isn't being made. Scientists have a phenomenal ability to "muddle through" (Lindblom, 1959) with whatever tools they have. Our goal in this paper is to help working scientists understand statistical science, and thereby help them muddle through more effectively.

More concretely the goals of this paper are: 1) To sketch the 3 major statistical paradigms that can be used by researchers, and in so doing introduce to many readers evidential statistics as a formal inferential paradigm that integrates control of error, model identification, model uncertainty, parameter estimation and parameter uncertainty. 2) To clarify some of the major confusions infesting arguments among paradigm adherents. 3) To discuss a few real problems arising in the error statistical and Bayesian approaches. And, 4) To raise some ideas about statistics and science which may help scientists use statistics well.

For more than a century a scientist wanting to make inference from experimental or observational data was stepping onto a battlefield strongly contested by two warring factions. These camps are generally referred to as frequentist and Bayesian statistics. In order to understand these factions, and given that statistics' foundation lies in probability theory, one must be aware that the two camps have their roots in two widely different

definitions of probability (Lindley 2000). Already then, confusion starts because, as we shall see in the sequel, the labels "frequentist" and "Bayesian" confound two related but distinct arguments: one on definitions of probability and another on styles of inference.

Here, we will characterize the inferential debate as between error statistics and Bayesian statistics. Evidential statistics has arisen as a natural response to this tension, and has been constructed, more or less consciously, from both paradigms by appropriating good features and jettisoning problematic features (Lele 2004b, Royall, 2004)). With three choices the debate can shift from a winner take all struggle to a discussion of what is most useful when dealing with particular problems. Given the scope of topics, the discussion we present will be largely conceptual, with indicators into the scientific, statistical, and philosophical literatures for more technical treatment.

Interpretations of Probability

The idea of probability, chance or randomness is very old and rooted in the analysis of gambling games. In mathematics, a *random experiment* is a process whose outcome is not known in advance. One of the most boring yet simple to understand examples of a random experiment consists of (you guessed it) flipping a coin once. From the coin flip, we could go onwards defining the sample space of an experiment as the set of all possible outcomes in the sample (which in the coin flipping experiment is the set $\{Head, Tail\}$ typically denoted as Ω), and we could give an example of an event (like getting a "Heads" after a single coin flip). These definitions would then set the stage for defining what models derived from probability theory are, and explaining how these are useful because they can be applied to any situation in which the events occur randomly.

However, we caution that even the most apparently simple of these definitions and concepts have subtle and hidden complexities. In 2010, for instance, professor Perci Diaconis, the well known probabilist, gave a lecture entitled "The search for randomness". In it, he took a close look at some of the most primitive examples of randomness, and yes, flipping a coin was one of them. He showed that what we are used to call "random", like a coin flip, can be quite non-random. What we call and model as randomness comes from at least 4 different sources (Guttorp 1995): 1) Uncertainty about initial conditions, 2) Sensitivity to initial conditions, 3) Incomplete process description, and 4) Fundamental physical randomness.

Kolmogorov's axioms and measure theory give the tools to work with many kinds of probabilities. These axioms state that a probability is a number between 0 and 1 associated with a particular event in the sample space of a random experiment. This number is in fact a (positive) measure of the chance that the event will occur. If A is an event, then $\Pr(A)$ measures the chance that the event will occur. Furthermore, if Ω is the sample space of our random experiment, $\Pr(\Omega)=1$. Finally, if two or more events are disjoint (*i.e.*, do not have any outcomes in common), the probability of either of these events occurring, or all of them, is equal to the sum of the individual probabilities of each of these events.

Any system that satisfies the requirements of the preceding paragraph is a probability and can be manipulated according to the rules of probability theory. However, what these manipulations mean will depend on how probability is interpreted. There are 5 major schools of interpretation of probability: classical (or Laplacian), logical, frequentist, subjective, and propensity. All of them can be and have been critiqued (see

Hajek 2012). When we think about science we use a combination of the frequentist, propensity, and subjective interpretations so for the purposes of this essay, we will give a brief introduction to only these three interpretations of probability. Laplacian probability is discussed by Yamamura (2015).

The frequency interpretation of probability itself has two flavors. The finite frequency interpretation of probability states that the probability of an event is just the proportion of times that event occurs in some finite number of trials (Venn 1876). The countable frequency interpretation of probability is as follows: if the random process is hypothetically repeated, then the long-run proportion of times an event occurs is the probability of the event (von Mises 1928).

Propensity probability (Peirce, 1878; Popper, 1959) is simply the innate or natural tendency of an event to occur in an experimental or observational setting. If you flip a coin, there is an innate tendency for it land showing heads. Similarly, a radioactive atom has an innate tendency to decay in a given time period.

In our opinion, combining these two frequency definitions of probability with the propensity definition of probability creates an effective framework for learning from nature. While we cannot know this propensity fully, we can approximate it using finite frequencies. On the other hand, if one has a model or models of the workings of nature, one can calculate the long run frequency probabilities of events under the model. It is the matching of finite frequency approximations of event propensities with model based long run frequency calculations of event probabilities that form the bases of inference.

The subjective interpretation of probability involves personal statements of belief regarding the chance of a given event, with beliefs being constrained to vary between 0

142 and 1. Subjective probabilities vary from individual to individual. A betting scenario is an
143 ideal representation for this interpretation: if you bet 'a' dollars to my 'b' dollars that
144 your favorite horse will win a race, then your probability that this horse wins the race is
145 $\Pr(win) = a/(a+b)$.

146 The different interpretations of probability described constitute fundamentally
147 different approaches to representing the world. Consequently they lead to intrinsically
148 different ways of carrying a statistical analysis in science. Explicating these differences
149 is the goal of this article.
150

151 **Fisher's foundational contribution to statistics using probability**

152 Fisher's likelihood function lies at the very foundation of statistics as we know it
153 today, and extensive book-length treatments and papers have been written about it (e.g.
154 Edwards, 1992; Pawitan 2001). To introduce likelihood here, we consider the example
155 about a simple experiment in which a series of success/failure trials are carried and their
156 results recorded. These types of experiments arise often in a wide array of scientific
157 disciplines, such as medical trials where a drug is tested or wildlife management in mark-
158 recapture studies. How do we go about writing a probability model for an experiment of
159 this type? Can we build a statistical model to explain how the data arose?

160 The data being the number of successes recorded in a given experiment, it is
161 natural to try to model these counts as the outcome of a binomial random variable *X*. By
162 so doing, the set of all possible outcomes, or sample space, is formally associated with a
163 set of probabilities. These sample space probabilities naturally add up to one. Let *n* be
164 the number of (independent) trials carried out (set *a priori)* and *x* the number of successes

actually observed in one realization of the experiment. Assume that the probability of success p in each trial remains unchanged. Hence, the probability of a particular sequence of x successes and n-x failures is $p^x(1-p)^{n-x}$ and it follows that

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

The probabilities depend critically on the parameter p. Thus this model is useless for prediction and understanding the nature of the trials in question if the value of p is not estimated from real data. Once estimation is achieved, we may seek to answer questions such as: can the success probability be assumed to be constant over a given array of experimental settings? Using the same example, Fisher (1922) argued that, given an outcome x, graphing $\binom{n}{x} p^x (1-p)^{n-x}$ as a function of the unknown p, would reveal how *likely* the different values of p are in the face of the evidence. This is a switch in focus from the descriptive inference about the data common at the time to inference about the process generating the data. Noting that the word 'probability' implies a ratio of frequencies of the values of p and that "about the frequencies of such values we can know nothing whatever", Fisher spoke instead of the likelihood of one value of the unknown parameter p being a number of times bigger than the likelihood of another value. He then decided to define the likelihood that any parameter should have any assigned value as being proportional to the probability of observing the data at hand if this was so. Thus, following Fisher, we refer to the function

$$\ell(p) = c.\binom{n}{x} p^x (1-p)^{n-x}$$

where 'c' is a constant that doesn't depend on the parameter of interest as the likelihood

function of p (see for instance Kalbfleisch 1985). This function uses the relative frequencies (probabilities) that the values of the hypothetical quantity p would yield the observed data as support for those hypothetical values (Fisher 1922). The distinction between likelihood and probability is paramount, because as a function of p, $\ell(p)$ is not a probability measure (*i.e.*, it does not integrate to 1).

The value $\hat{p}$ that maximizes this function is called the Maximum Likelihood (ML) estimate of the parameter p. The graphing of the likelihood function supplies a natural order of preference among the possibilities under consideration (Fisher 1922). Such order of preference agrees with the inferential optimality concept that prefers a given probability model if it renders the observed sample more probable that other tentative explanations (i.e. models) do. Thus, by maximizing the likelihood function derived from multiple probability models (in this case values of p) as hypotheses of how the data arises, one is in fact seeking to quantify the evidential support in favor of one probabilistic model (value of p) over the others (other values of p in our example. See introductions to the likelihood function by Fisher 1922, Kalbfleisch 1985, Pawitan 2001, Sprott 2000, Royall 2004).

Finally, because likelihood ratios are ratios of frequencies, they have an objective frequency interpretation that can be verified by computer simulations. Stating that the relative likelihood of one value p_1 of the unknown parameter over another value p_2, written as $\ell(p_1)/\ell(p_2)$, is equal to a constant k means that the observed data will occur k times more frequently in repeated samples from the population defined by the value p_1 than from the population defined by p_2 (Sprott 2000). Because of this meaningful

frequentist interpretation of likelihood ratios, authors like Barnard (1967), or Sprott (2000) stated that the best way to express the order of preference among the different values of the parameter of interest using Fisher's likelihood is by working with the relative likelihood function, given by

$$R(p;x) = \frac{\ell(p;x)}{\sup_{p} \ell(p;x)} = \frac{\ell(p;x)}{\ell(\hat{p};x)}.$$

As we will see later, this frequency interpretation of the likelihood ratios is the fundamental basis for likelihood inference and model selection.

At this point, it may be useful to expand on the understandings in this paper of the terms "model", "parameter", and "hypothesis". For us, a model is a conceptual device that explicitly specifies the distribution of data. To say for instance that the data are "gamma distributed" is a only a vague model, inasmuch the values of the shape and rate parameters of this mathematical formulation of a hypothesis are not specified. Here, we adhere to the formalism where biological hypotheses aren't fully specified as a mathematical model until the parameter values of the probabilistic model are themselves explicitly defined. This requisite is not a mere formalism because different parameter values, or sets of values, truly index different families of models. Hypotheses then, become posited statements about features of the mathematical models that best describe data.

Fisher's principles of experimentation and testing assertions in science

In R.A. Fisher's experimental design book (1971) there is a ten pages account of an experiment where he laid out some of the most important principles of experimentation.

The experiment is famously known as “Fisher’s lady tasting tea experiment”. This account tells the story of a lady that claimed to be able to distinguish between a tea cup which was prepared by pouring the tea first and then the milk and another tea cup where the milk was poured first. Fisher then wonders if there is there a good experiment that could be devised in order to formally test the lady’s claim using logical and mathematical argumentation. Although seemingly trivial, this setting where a scientist, and in particular, an ecologist claims to be able to distinguish between two types of experimental units is a daily reality.

Decades ago, in the late 80’s, one of us was faced with a similar experimental problem. While in Japan doing research on seed-beetles, MLT taught himself to visually distinguish the eggs of *Callosobruchus chinensis* and *C. maculatus* to the point where he asserted that he could indeed make such distinction. Doubting himself (as he should have), MLT recruited the help of prof. Toquenaga to set up tea-lady like blind trials to test his assertion (except there was no beverage involved and the subject certainly isn’t a lady, and perhaps not even a gentleman). In this case, testing the researcher’s claim involved giving the facts –the data– a chance of disproving a skeptic’s view (say, prof. Toquenaga’s position) that the researcher had no ability whatsoever to distinguish between the eggs of these two beetle species.

This tentative explanation of the data is what is generally called “the null hypothesis”. To Fisher, the opposite hypothesis that some discrimination was possible was too vague and ambiguous in nature to be subject to exact testing and stated that the only testable expectations were “those which flow from the null hypothesis” (Fisher 1956). For him it was only natural to seek to formalize the skeptic’s view with an exact

probabilistic model of how the data arose and then ponder how tenable such model would be in the face of the evidence. By so doing, he was adopting one of the logic tricks that mathematicians use while writing proofs: contradiction of an initial premise. Applied to this case, and given that MLT had correctly classified 44 out 48 eggs, the trick goes as follows: First we suppose that the skeptic is correct and that the researcher has no discrimination ability whatsoever, and that his choices are done purely at random, independently of each other. Then, because the seed-beetle experimental data is a series of classification trials with one of two outcomes (success or failure), we naturally model the skeptic's hypothesis using a binomial distribution X counting the number of successfully classified eggs, with a probability of success $p = 0.50$. Next we ask, under this model, what are the chances of the researcher being correct as often as 44 times out of 48 (the observed count) or even more? According to the binomial model, that probability is about $8*10^{-10}$. That is, if the skeptic is correct, a result as good or better than the one actually recorded would be observed only about 0.000008% of the time under the same circumstances. Hence, either the null hypothesis is false, or an extremely improbable event has occurred.

The proximity to 0 of the number 0.000008% (the P-value) is commonly taken as a measure of the strength of the evidence against the null hypothesis. Such an interpretation is fraught with difficulty, and we would advise against it. This account is important insofar as it illustrates how the enumeration of the sample space probabilities can be used to test via inductive inference the validity of an assertion. We also find the researcher vs. skeptic setting (Dennis 2004) valuable in and of itself to explain Fisher's P-value.

A Sketch of Error Statistics

Error Statistics (Mayo 1996) is the branch of statistics most familiar to ecologists, and certainly to beginning ecologists. All of the methods in this category share the organizing principle that control of error is a paramount inferential goal. These procedures are designed so that an analyst using them will make an error in inference no more often than a pre specified proportion of the time.

Instead of focusing on testing a single assertion like Fisher, Neyman-Pearson (NP) showed that it was possible to assess one statistical model (called the null hypothesis) against another statistical model (called the "alternative hypothesis"). A function of potential data, *T(X),* is devised as a test statistic to indicate parameter similarity to either the null hypothesis or the alternate. A critical value or threshold for T is calculated **so that** if the null is true, the alternate will be indicated by *T* no more than a pre-designated a proportion of the time α. The test is designed so that the null hypothesis will be incorrectly rejected no more than a proportion α of the time. The NP test was designed as a data-driven choice between two competing statistical hypotheses of how the data arose, and appears to be a straight ahead model comparison.

However, one can, as Fisher did, unravel its unexpected connections with the Fisherian P-value. NP's model-choice strategy could indeed deal with vague alternatives (or null hypotheses, for that matter), such as ``the researcher has indeed some discrimination ability". NP termed these "composite hypotheses", as opposed to fully defined "simple" statistical models.

NP's approach proceeds as follow: the researcher implicitly concedes that the null hypothesis could be true. If that is the case, then the probability distribution of the test statistic can readily be computed (either analytically or computationally). This computation is possible because the test statistic, by being a function of the potential outcomes, inherits randomness from sample space probabilities. The difference between NP and Fisher resides in what questions they would seek to answer with this distribution. Fisher would ask here: if the null hypothesis is true, what is the probability of observing a value of the test statistic as extreme or more extreme (in the direction of the research hypothesis) than the test statistic actually observed? Fisher maintained that if such probability (the P-value) is very small, then the null model should be deemed untenable.

NP recognized on the other hand that in order to make a decision one could simply assume that the skeptic has a fixed threshold for such probability. If, say, the probability of observing a value of the test statistic as large or larger than the one recorded is smaller than 1%, then that would be enough to convince the skeptic to decide against her/his model. Adopting such threshold comes with the recognition that whichever decision is made, two possible errors arise: first, the null hypothesis could be true, but it is rejected. The probability of such rejection is simply given by the value of the adopted threshold, since we reject the null hypothesis whenever we observe a P-value smaller than it (after having assumed that the null is true). That error, for lack of a better name, was called an "error of the first type", or "Type I error" and the probability of this kind of error is denoted as α. Second, it may be possible that we fail to reject the null, even if it is false. This type of error is called "Type II" error. The probability of this error is usually denoted by β and can be computed from the probabilistic definition of

the alternative hypothesis via its complement, $1-\beta$. This is the probability of rejecting the null when it is indeed false. Thus, by considering these two errors, NP tied the testing of the tenability of a null hypothesis to an alternative hypothesis.

Let us return to our seed-beetles eggs classification problem. The null hypothesis is that the counts *X*, are binomially distributed with an n=48 and p=0.5. Suppose that before starting the test, professor Toquenaga (our skeptic) would have stated that he would only have conceded if MLT correctly classified 85% or more of the eggs. That is, a number of successful classification events greater or equal to 41/48 would represent a rejection of the null. Under such null the skeptic threshold α is $\alpha = \Pr(X \geq 41) = \sum_{x=41}^{48} \binom{48}{x} 0.5^x (1-0.5)^{48-x} = 3.120204 * 10^{-07}$. If in fact, MLT's probability of success is, say, p=0.90, then the power of the test is computed by calculating the probability that the observed count will be greater than or equal to 41/48 under the true model is $1-\beta = \Pr(X \geq 41) = \sum_{x=41}^{48} \binom{48}{x} 0.9^x (1-0.9)^{48-x} \approx 0.89$. In closing this account, note that an ideal test would of course have a pre-defined $\alpha = \beta = 0$ but this can only be achieved for certain non-practical cases. Because of the way these error probability calculations are set up, to increase the value of one error means the value of the other one needs to decrease. In practice, before the experiment starts, the researcher fixes the value of α in advance and then changes the sampling space probabilities by increasing the sample size and thus adjusts β to a desired level. Although NP require setting the Type I error in advance, the magnitude of acceptable error is left to the researcher.

Thus, Neyman and Pearson took Fisher's logic to test assertions and formalized the scenario where a data-driven choice between two tentative explanations of the data needed to be made. Although their approach resulted in a well-defined rule of action with respect to such decision that quickly became the workhorse of scientific inquiry, Fisher quickly pointed out how such paradigm had unfortunately lost track of the strength of the evidence and also, that the possibility existed that such evidence would, with further experimentation, very well become stronger or even weaker.

The NP test requires a prespecification of hypotheses (i.e. parameter values). Often however, data are collected before knowledge of parameter values is in hand. The error statistical approach to inference is still feasible. Confidence intervals, do not pre-specify the hypotheses, data are collected, a parameter value estimated, and an interval constructed around the estimate to represent plausible values of the parameter in such a fashion that under repeated sampling, the true parameter will be outside of the interval no more than a pre-specified α proportion of the time. Nevertheless, the connection between hypothesis tests and confidence intervals is very close. Confidence intervals can be conceived of, and calculated as, inverted hypothesis tests.

Fisher's P-value wears many hats in statistics. But, one of its interpretations lands it squarely in the Error Statistics category. The Fisherian significance test does not compare multiple models as do the NP-test and confidence intervals. A single null hypothesis is assumed, and a test statistic is devised to be sensitive to deviations from the hypothesis. If data are observed and the calculated test statistic is more dissimilar to the null hypothesis than a prespecified P-value proportion of data randomly generated from the null, then the null hypothesis is rejected, otherwise one fails to reject it. If the P-value

366 is not pre-specified, but only observed post-sampling then it does not control error in the
367 same fashion the NP-test and confidence interval do, yet it is regarded by many as a
368 quantitative measure of the evidence or against the null hypothesis.

369 The mathematical statistics theory concerning the distribution of likelihood ratios
370 made possible connecting Fisher's maximum likelihood with hypotheses tests, and gave
371 rise to many of the tests that are nowadays the workhorse of statistical testing in science
372 (Rice 1995). The idea of evaluating the likelihood of one set of parameters vis-à-vis the
373 maximum likelihood gave rise not only to confidence intervals, but to relative profile
374 likelihoods where the likelihood of every value of the parameter of interest is divided by
375 the maximum of this curve. And this idea in turn motivated the use of likelihood ratios to
376 carry model selection via likelihood ratio tests. Sample space probabilities pass on
377 randomness not only to the test statistic, but also, to the likelihood profile and of course,
378 likelihood ratios.

379

380 **A Sketch of Bayesian Statistics**

381 A discussion of Bayesian statistics has to begin with a description of what probability is
382 to a Bayesian. Formally, Bayesian probabilities are measures of belief by an agent in a
383 model or parameter value. The agent learns by adjusting her beliefs. Personal beliefs are
384 adjusted by mixing belief in the model with the probability of the data under the model.
385 This is done with an application of a formula from conditional probability known as
386 Baye's rule: If A and C are two events and their joint probability is defined, then

387 $$\Pr(A \mid C) = \frac{\Pr(A \text{ and } C)}{\Pr(C)} = \frac{\Pr(C \mid A)\Pr(A)}{\Pr(C)}.$$

The application of Bayes rule in Bayesian statistics runs as follows. Given the conditional probability of observing the data x under the model M_i written as $f(x \mid M_i)$, and if our prior opinion about such model is quantified with a prior probability distribution, $f_{prior}(M_i)$, then the updated, conditional probability of a model given the observed data becomes:

$$f_{post}(M_i \mid x) = \frac{f(x \mid M_i) f_{prior}(M_i)}{\sum_j f(x \mid M_j) f_{prior}(M_j)}.$$

In English this equation reads that your belief in a model M_i after you have collected data x (that is your posterior probability) is a conditional probability, given by the product of the probability of the data under the model of interest and the prior probability of the model of interest, normalized so that the resulting ratios (posterior probabilities) of all of the models under consideration sum to one. This is a pretty important constraint. If they don't sum to one, then they are not probabilities and you cannot employ Baye's rule. If the models lie in a continuum, that is the models are indexed by a continuous parameter, then the sum in the denominator is replaced by an integral.

While the notation in Baye's rule treats all the probabilities as the same, they are not the same. The prior distribution, $f_{prior}(M_i)$, quantifies the degree of belief, a personal opinion, in model *i*. The model or parameter of interest is then seen as a random variable. By so doing, a key inferential change has been introduced: probability has been defined as a measure of beliefs. Let's call, for the time being, these probabilities "b-probabilities". Now the term $f(x \mid M_i)$ is taken as a conditional measure of the frequency with which data like the observed data x would be generated by the model. It is

409 taken to be equal to the likelihood function (aside from the constant 'c', which cancels
410 out with the same constant appearing in the denominator of the posterior probability).
411 This is not a belief based probability, it is the same probability used to define likelihood
412 ratios and carry frequentist inference. Let's call it an "f-probability" to distinguish it from
413 the beliefs-derived probabilities. In the application of Bayes formula above, probability
414 of the data appears as multiplying the prior beliefs in the numerator. The resulting
415 product, after proper normalization becomes the posterior probability of the model at
416 hand, given the observations. It is true that both, f-probabilities and b-probabilities are
417 true probabilities because they both satisfy Kolmorgorov's axioms (Kolmorogorov 1933),
418 but to think that they are the same is to think that cats and dogs are the same because they
419 are both mammals: one is a beliefs probability whereas the other one is a sample space
420 probability. It is important to note that when you mix an f-probability with a b-probability
421 using Bayes Theorem, one ends up with a b-probability, an updated beliefs probability.
422 To make these ideas concrete, we work out our binomial egg classification
423 problem using Bayesian statistics. Our general model of how the data arises for this
424 experiment is given by the binomial formula with *n* trials, *x* successes and a probability *p*
425 of success. Changing *p* in such formula changes the hypothesized model of how the data
426 arises. Because binomial formula accepts for *p* any value between 0 and 1, changing *p*
427 amounts to changing models along a continuum. Let our prior beliefs about this
428 parameters be quantified with the probability distribution $g(p)$. The beta distribution
429 with parameters *a* and *b* is a convenient distribution for $g(p)$. The posterior distribution
430 of *p* given the data *x* is proportional to:

$$f_{post}(p \mid x) \propto f(x \mid p) g(p) = \binom{n}{x} p^x (1-p)^{n-x} \, p^{a-1} (1-p)^{b-1} \propto p^{a+x-1} (1-p)^{n+b-x-1}.$$

Note that the resulting expression of the posterior distribution shown above is in fact, after proper normalization, another beta distribution with parameters $a+x$ and $b+n-x$. Note also that, the mean of our prior distribution is by definition $a/(a+b)$. By the same token, the mean of the posterior distribution is:

$$\frac{a+x}{(a+x)+(b+n-x)} = \frac{a+b}{a+b+n}\left(\frac{a}{a+b}\right) + \frac{n}{a+b+n}\bar{x},$$

where $\bar{x} = x/n$ is the sample mean. Therefore, the posterior mean is seen to be a weighted average of the prior mean and the sample mean. In a very real sense, the posterior mean is a mixture of the data and the prior beliefs. As the sample size gets large, however, the weight of the first term in this sum goes to 0 while the weight of the second one converges to 1. In that case, the influence of the prior beliefs gets "swamped" by the information in the data. Dorazio (2015) claims that the Bayesian posterior is valid at any sample size. That doesn't mean that anything useful has been learned from the data, as this author also later suggests (Dorazio 2015, this volume). We can see from the above expression that the Bayesian posterior may well be dominated by the prior at low sample sizes.

In any case, however, Bayesian learning occurs when this process is iterated upon collecting new data. The posterior distribution becomes the instrument for inference: if the parameter of interest is assumed to be a random variable, then the posterior distribution instantly gives the probability that such value lies between any two limits, say p_{low} and p_{high}. Hence, although for estimation purposes either the posterior mean or

mode are given as estimates of the unknown parameter, the entire distribution can be used for statistical inference.

The Bayes Factor (Kass and Raftery, 1995; Raftery, 1995) is used Bayesian statistics to measure the evidence in the data for one model over another. Written as $\Pr(D|M_1)/\Pr(D|M_2)$ where D denotes the data and M_i the i^{th} model, the Bayes factor looks very similar to the ratio of likelihoods evaluated under the two different models, and in fact serves a similar function. For models with specified parameter values, the two are the same. But, for the more common situation where the parameter values are yet to be determined by the analysis, the likelihood ratio and the Bayes factor are not the same. In this latter case, the Bayes Factor is computed as the ratio of two averaged likelihoods each averaged (integrated) over the prior b-probability of the parameters, whereas the likelihood ratio is calculated as the ratio of the two likelihood functions evaluated at the ML estimates (i.e., at the maximum, see the account by A. Raftery 1995, section 3.2). Consequently, the Bayes Factor is not a measure of evidence independent of prior belief.

The above description of Bayesianism perhaps gives the impression that it is a monolithic school. It is not. In the interests of brevity we will speak of only three different Bayesian schools that focus each on different interpretation of the prior distribution. In Subjective Bayesianism the prior is a quantitative representation of your personal beliefs. This makes sense as a statistics of personal learning. Although the subjectivity involved has made many scientists uncomfortable, subjective Bayesians posit that it is the prior distribution that conveys initial information and thus provides the starting point for the Bayesian learning process (Lindley 2000, Rannala 2002). Indeed, an

often repeated justification for using the Bayesian solution in intricate biological problems is the ability to bring into the analysis external, prior information concerning the parameters of interest (Rannala 2002).

Objective Bayesianism on the other hand was developed to respond to the discomfort introduced by subjective priors. Under that school of thought, the prior distribution is a quantitative representation of a declaration of ignorance about the parameters of interest. Prior probabilities are assigned to alternative models/parameter values so as favor one individual model over another as little as possible given mathematical constraints. These priors are called non-informative. Royle and Dorazio (2008) present an ecologically oriented introduction statistical analysis emphasizing objective priors.

Another kind of analysis often falling under the Bayesian rubric is empirical Bayesianism. Here the prior probabilities are estimated from external empirical information. Clearly this is a different beast from either forms of belief based Bayesianism described above, and extensive discussions about this approach and the other two Bayesian views presented here can be found in the statistical literature. The critiques of Bayesianism found below are not directed at empirical Bayes. An excellent introduction to empirical Bayes can be found in Efron (2010).

A Sketch of Evidential Statistics

Richard Royall begins his 1997 book Statistical Evidence: A likelihood paradigm with 3 questions:

1) What do I believe, now that I have this observation?

2) What should I do, now that I have this observation?

3) What does this observation tell me about model/hypothesis A versus B? (How should I interpret this observation as evidence regarding A versus B?).

This third question is not clearly addressed by error statistics. Nor is it addressed by Bayesian statistics, because belief and confirmation are actually quite distinct from evidence, as is argued forcefully in Bandyophadyay et al. (2015). Following Hacking (1965) and Edwards (1992), Royall axiomatically takes the likelihood ratio as his measure of evidence and proceeds to develop a very powerful inferential frame work.

Royall divides the result space of an experiment differently than the Neyman-Pearson paradigm. The NP test has two regions: one where you accept A, and another region where you accept B. For Royall, there are 3 regions: one where evidence is strong for A over B, another where evidence is strong for B over A, and a region between where evidence (whether leaning towards A or towards B) is weak. The advantage of this in the actual interpretation of scientific results is obvious. First, no decision is made only the strength of evidence is determined. Second, there is a region of indeterminacy, where the primary conclusion is that not enough data have been obtained.

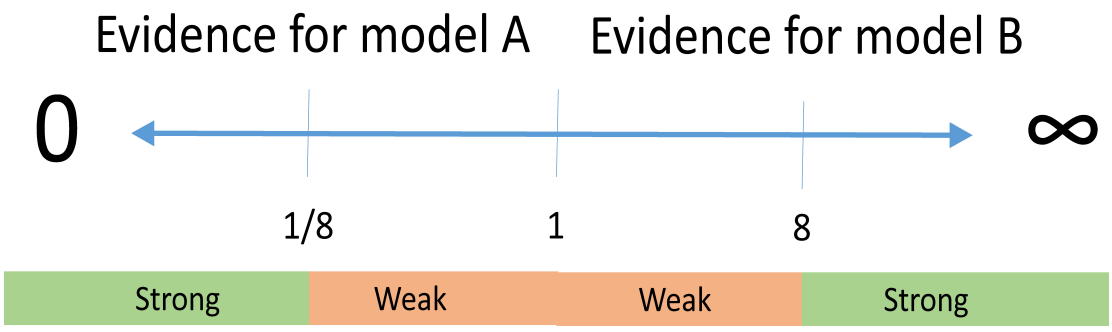


Figure 1: A graphical representation of evidence in the likelihood ratio for one model over another. The numbers reflect Royall's treatment of evidence as a ratio, while

517 the actual scale of the figure reflects our preference to representing evidence by a log of
518 the likelihood ratio.

519 Neyman-Pearson hypothesis tests have two important error rates, the probability
520 of type I error, α, and the probability of type II error, β. With evidential statistics you
521 never actually make an error, because you are not making a decision, only determining
522 the strength of evidence. Nevertheless, evidence even properly interpreted can be
523 misleading – one may find strong evidence for one model when in fact the data was
524 generated by the other. This allows for two interesting probabilities reminiscent (but
525 superior) to α and β. These are: the probability of misleading evidence, *M*, and the
526 probability of weak evidence, *W*. This distinction will be discussed further later.

527 This approach immediately combines strengths from the Neyman-Pearson
528 hypothesis tests, and from Fisherian pure significance tests. Requiring evidence to pass
529 an *a priori* threshold gives a control of error. Royall (1997) shows that if the threshold
530 for strong evidence is *k*, the probability of misleading evidence is $M \leq 1/k$. The basis for
531 such conclusion stems from the frequency interpretation of Royall's measure of evidence:
532 the likelihood ratio between any two models. As we mention above, writing that
533 $\frac{\ell(p_1)}{\ell(p_2)} = k$ means that the observed data will occur *k* times more frequently in repeated
534 samples from the population defined by the value p_1 than from the population defined by
535 p_2 . Hence, this ratio can be interpreted as a random variable, one which happens to be
536 on average (over hypothetical repeated sampling) equal to 1 if in fact, the two models
537 (parameter values in our example) explain the data equally well. If we deem as most

likely the first model only when the likelihood ratio exceeds a value k, then, a direct application of Markov's Theorem allows us to write that

$$\Pr\left(\frac{\ell(p_1)}{\ell(p_2)} \geq k\right) \leq \frac{1}{k}.$$

Therefore, the chance of observing a misleading likelihood ratio, one greater than the cut-off for strong evidence k, is in fact less than or equal to $1/k$.

The strong evidence threshold is a pre-data control of error, very much like NP's Type I error rate. Post data the actually observed evidence (likelihood ratio for Royall) is a fine grained measure. Thus, evidential statistics allows researchers to simultaneously make pre and post data inferences in a coherent framework, as so long craved by practitioners.

The mathematical treatment in Royall (1997) makes a true model assumption (i.e. one of the models in the evidential comparison is true). For the most honest and effective inference, the true model assumption needs to be relaxed. Lele (2004a) eliminates this assumption when he generalizes the likelihood ratio to evidence functions which are conceptualized as the relative generalized discrepancy between two models and reality. Relaxing the true model assumption creates a great philosophical advantage for the evidential approach, but because the focus of this essay is practical, we direct interested readers to Bandyophadyay et al. (2015) for a fuller discussion.

Rather than presenting a single monolithic evidence function, Lele sets out a structure for constructing evidence functions. Lele (2004a) and Taper and Lele (2011) discuss desirable features for evidence functions. These desiderata include:

D1) Evidence should be a data based estimate of the relative distance between two models and reality.

D2) Evidence should be a continuous function of data. This means that there is no threshold that must be passed before something is counted as evidence.

D3) The reliability of evidential statements should be quantifiable.

D4) Evidence should be public not private or personal.

D5) Evidence should be portable, that is it should be transferable from person to person.

D6) Evidence should be accumulable: If two data sets relate the same pair of models, then the evidence should be combinable in some fashion, and any evidence collected should bear on any future inferences regarding the models in question.

D7) Evidence should not depend on the personal idiosyncrasies of model formulation. By this we mean that evidence functions should be both scale and transformation invariant.

D8) Consistency, that is as $M+W \rightarrow 0$ as $n \rightarrow \infty$. Or stated verbally, evidence for the true model/parameter is maximized at the true value only if the true model is in the model set, or at the best projection into the model set if it is not.

Although the formal structure of evidence functions is relatively new, a number of evidence functions have long been proving their utility. Likelihood ratio and log likelihood ratios, for instance, are evidence functions. Other evidence functions include order consistent information criteria, such as Schwarz's (1978) information criterion, SIC also known as the BIC (Bayesian Information Criterion), the consistent AIC, CAIC, (see Bozdogan 1987), and the information critierion of Hannan and Quinn (1979), ICHQ. These information criteria are all functions of the log-likelihood maximized under the

583 model at hand plus a penalty term. As a result, the difference in the values of a given
584 information criteria between two models is always a function of the likelihood ratio.
585 Because the likelihood ratio is an evidence function, maximum likelihood
586 parameter estimation is an evidential procedure. Furthermore, likelihood ratio based
587 confidence intervals can also be interpreted as evidential support intervals.
588 Not all information criteria are *sensu stricto* evidence functions (Lele 2004).
589 There is a class of information criteria, strongly advocated by Burnham and Anderson
590 (2002) that are not. These forms can be designated Minimum Total Discrepancy (MTD)
591 forms (Taper, 2004). They meet desiderata D1)-D7), but not D8). The very commonly
592 employed Akaike (1974) information criterion, the biased corrected AIC (AICc, Hurvich
593 and Tsai, 1989) are MTD criteria. That these forms are not strict evidence functions is not
594 to say that these forms are wrong per se, or that they shouldn't be used evidentially, but
595 that these criteria are evaluating models with a slightly different goal than are evidence
596 functions. The design goal of these forms is to select models so as to minimize
597 prediction error, while the design goal for evidence functions is to understand underlying
598 causal structure (see discussion in Bozdogan, 1987, Taper 2004, and Aho et al., 2014).
599 The consequence of this is that asymptotically, all MTD forms will over fit the data by
600 tending to include variables with no real association with the response. But at smaller
601 sample sizes the differences between the classes is not clear cut. The AIC tends to over
602 fit at all sample sizes, while the AICc can actually have a stronger complexity penalty
603 than the order consistent forms.
604 A small conceptual leap that needs to be made to recognize information criteria as
605 evidence functions is the change of scale involved. Royall uses the likelihood ratio as his

evidence measure while the difference of information criterion values can be thought of as a log likelihood ratio with bias corrections. Take for instance the difference in the score given by an information criterion (IC) between a model deemed as best among a set of models and any other model *i* within that set, and denote it as $\Delta IC_i = IC_i - IC_{best}$. Note that because the IC of the best model is the smallest, by necessity this difference is positive. Because all information criteria can be written as twice the negative log-likelihood maximized under the model at hand plus a complexity penalty that can be a function of both, the sample size and the number of parameters in the model, we can write a general equation for the difference in any IC score. Denote the complexity penalty for model *i* as $cp(d_i, n)$, where d_i is the dimension (number of estimated parameters) under model *i* and *n* is the sample size. For example, in the case of AIC, $cp(d_i, n) = 2d_i$ whereas for SIC, $cp(d_i, n) = d_i \ln(n)$. Accordingly,

$$\begin{aligned} \Delta IC_i &= -2\ln\hat{\ell}_i + cp(d_i, n) - \left(-2\ln\hat{\ell}_{best} + cp(d_{best}, n)\right) \\ &= -2\ln\left(\frac{\hat{\ell}_i}{\hat{\ell}_{best}}\right) + \Delta cp, \end{aligned}$$

where $\frac{\hat{l}_i}{\hat{l}_{best}}$ is the ratio of maximized likelihoods under each model, and $\Delta cp = cp(d_i, n) - cp(d_{best}, n)$ denotes the difference in the complexity penalties from model *i* and the best model. For instance, in the case of the SIC, $\Delta cp = \ln(n)(d_i - d_{best})$, and in the case of the AIC, $\Delta cp = 2(d_i - d_{best})$. Writing the difference in this fashion makes it clear that a ΔIC is indeed a log-likelihood ratio plus a bias correction constant that depends on the sample size and the difference in the number of parameters between the two models. In the case of the AIC and

the SIC, depending on whether the best model is or not the most parameter rich, one would be either subtracting or adding a penalty to the log-likelihood ratio.

Finding the probability of misleading evidence given a strong evidence threshold k if in fact the two models explain the data equally well amounts to finding $\Pr\left(\Delta IC_i \geq k\right)$, which is equal to $1-\Pr\left(\Delta IC_i \leq k\right)$. This quantity is readily recognized as one minus the cumulative density function (cdf) of the ΔIC_i evaluated at k. And yes, talking about the difference in IC having an associated cdf implies that one should be able to say something about the long-run distribution of such difference. Indeed, because ΔIC_i is written as the log likelihood ratio plus a constant, we can use the frequency interpretation of likelihood ratios, find the probability distribution of $\Lambda = -2\ln\left(\frac{\hat{\ell}_{\text{i}}}{\hat{\ell}_{\text{best}}}\right)$ under hypothetical repeated sampling and then express the distribution of the ΔIC_i as that of Λ shifted by the constant Δcp. Operationally however, the pre-data control of error is achieved by fixing first the size of the probability of misleading evidence M, and then solving for the value of the threshold k that leads to $\Pr\left(\Delta IC_i \geq k\right) = M$. Upon substituting the expression for ΔIC_i in this equation we get that

$$\Pr\left(\Lambda + \Delta cp \geq k\right) = M \Leftrightarrow \Pr\left(\Lambda \geq k - \Delta cp\right) = M,$$

or $1-\Pr\left(\Lambda \leq k - \Delta cp\right) = M$. From this calculation, it is readily seen that the pre-data control of the probability of misleading evidence strongly depends on the form of the complexity penalty.

We now turn to an example, one where we give a closer look to the assumptions behind the now ubiquitous cut-off of two points in ΔIC_i. The cut-off of two points of difference in IC is readily derived from the calculations above, yet it implies that the user is facing a rather stringent model selection scenario. To see why, it is important to know first that the long-run distribution

of the log-likelihood ratio is in general very difficult to approximate analytically. Samuel Wilks (1938) provided for the first time the approximate distribution of Λ for various statistical models. If model *i* is true, as it is assumed when testing a null hypothesis vs. an alternative, and if the model deemed as best is the most parameter rich, then Wilks found that Λ has an approximate chi-square distribution with degrees of freedom equal to $d_{best} - d_i$. In this case, the expression $1 - \Pr(\Lambda \le k - \Delta cp)$ can be readily computed using any statistical software, like R. In the case of AIC, this expression becomes $1 - \Pr(\Lambda \le k - 2(d_i - d_{best}))$ and in the case of the SIC, it is $1 - \Pr(\Lambda \le k - \ln(n)(d_i - d_{best}))$. Using the now "classic" $k = 2$, $d_i - d_{best} = -1$ gives $1 - \Pr(\Lambda \le k - 2(d_i - d_{best})) = 0.0455$ for the AIC. In the case of the SIC, assuming a sample size of $n = 7$ we get $1 - \Pr(\Lambda \le k - \ln(n)(d_i - d_{best})) = 0.0470$. This example shows that under Wilks model setting (where the two models are nested and the simple model is the truth) a cut off of 2 does give an error control of about the conventional 0.05 size. Also, note that for the AIC and the SIC (unless sample size is tiny) an increase in difference in the number of parameters between the models results in an even stronger control of error. Finally, note that the strength of the error control does not vary when sample size is increased in the AIC but does so in the SIC. For the SIC, M decrease as sample size increases. This is what, in fact, makes the SIC an order consistent form.

Exact values of M will vary with criterion, sample size, structure of the models, nestedness of models, and the nearness of the best model to the generating process. If you are acting in a regulatory setting, or in an experimental design setting, then the precise value of M may matter. In these cases M should be explicitly calculated *a priori.* But, in the general prosecution of science, it really matters very little whether M is bounded at 0.07 or 0.03; both give moderately strong control of error. Adopting an *a*

669 *priori* cut off of say 2 for moderately strong control of error or of 4 for strong control of
670 error gives the scientist and the scientific community the protection from wishful thinking
671 that it needs without the fiction that control of error is known more precisely than it is.

672 Increasingly, towards the end of the 20th century, ecological statistics shifted its
673 focus from point and interval estimation for parameters in models that magically seemed
674 to appear from nowhere and whose connection to hypotheses of real scientific interest
675 were often somewhat tenuous, to trying to incorporate theories of ecological processes
676 directly in models to be statistically probed.

677 We strongly believe that the major source of error in all statistical analysis is due
678 to using the wrong model, and traditional statistics did not adequately address model
679 uncertainty. At least this was the state of affairs in 1995 (Chatfield, 1995). Since then,
680 Royall's (1997) reconstruction of traditional statistics, and Lele's (2004a) extension of
681 the likelihood ratio to evidence functions has allowed a statistical modern synthesis that
682 smoothly incorporates model identification, model uncertainty, parameter estimation,
683 parameter uncertainty, pre-data error control, and post-data strength of evidence into a
684 single coherent framework. We believe that that evidential statistics is currently the most
685 effective statistical paradigm for promoting progress in science.

686 For completeness, we need to draw attention to another recent statistical paradigm
687 called "severe testing" (e.g. Mayo and Cox, 2006; Mayo and Spanos, 2006). Similar to
688 evidential statistics, severe testing combines pre-data control of error with a post data
689 measure of the strength of inference. Despite very different surface presentations, there is
690 considerable similarity in their underlying mathematics between evidence and severe
691 testing. We find the evidential approach more useful for us for several reasons: First, in

evidence the primary object of inference is the model, while the primary object of inference in severe testing is the parameter value. Second, we find the direct comparison involved in evidence very intuitive and clear; we have always been confused by the counterfactual arguments required for testing (this of course is our shortcoming).--

An Example Evidential Application Using Information Criteria

To illustrate the evidential use of information criteria, we revisit an example from Lele and Taper (2012). That is the single-species population growth data from Gause's (1934) laboratory experiments with *Paramecium aurelia* with interest in the scientific questions of: 1) Does the population exhibit density dependent population growth? And, 2) If so what is the form of density dependence? The observed growth rate for a population is calculated as $r_t = \ln\left(N_{t+1}/N_t\right)$. By definition the growth rate of a population with density dependence is a function of population size, N_t (Figure 2). Consequently, we model the population's dynamics by $r_t = g\left(N_t, \underline{\theta}\right) + v_t\left(\sigma\right)$, where g is a deterministic growth function, $\underline{\theta}$ is a vector of parameters, $v_t\left(\sigma\right)$ is an independent random normally distributed environmental shock to the growth rate with mean 0 and standard deviation σ representing the effects of unpredictable fluctuations in the quality of the environment.

We use a suite of common population growth models: Ricker $\left(g\left(N_t, \underline{\theta}\right) = r_i\left(1 - N_t/K\right)\right)$, generalized Ricker $\left(g\left(N_t, \underline{\theta}\right) = r_i\left(1 - \left(N_t/K\right)^{\gamma}\right)\right)$, Beverton-Holt

$\left(g\left(N_t,\underline{\theta}\right)=r_i K\Big/\left(K+r_i N_t-N_t\right)\right)$, Gompertz $\left(g\left(N_t,\underline{\theta}\right)=a\left(1-\ln\left(N_t/K\right)\right)\right)$, and the density independent exponential growth model $\left(g\left(N_t,\underline{\theta}\right)=r_i\right)$. These models have been parameterized in as similar a fashion as possible. *K* represents the equilibrium population size, and r_i is the intrinsic growth rate, or limit to growth rate as N_t approaches 0. In the Gompertz model the parameter '*a*' also scales growth rate, but is not quite the same thing as r_i because in this model growth rate is mathematically undefined at 0.

The log-likelihood function for all of these models is

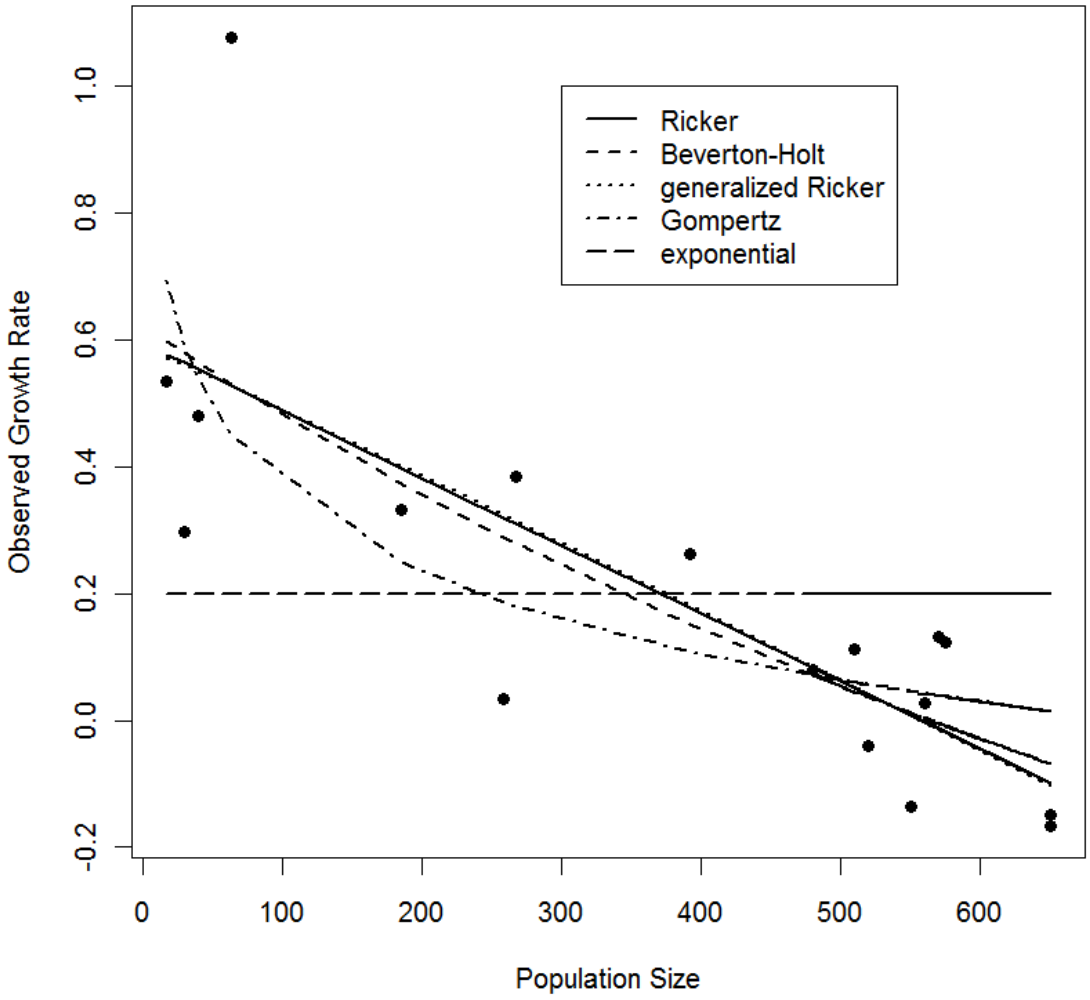
$$\log L\left(r_t,N_t;\underline{\theta},\sigma\right)=\frac{\sum_{t=0}^{T-2}\left(g\left(N_t,\underline{\theta}\right)-r_t\right)^2}{2\sigma^2}-\frac{\left(T-1\right)\log\left(2\pi\sigma^2\right)}{2}$$

, where *T* is the total number of population sizes observed. For the construction of information criteria, the number of parameters, p, is the length of the vector $\underline{\theta}$+1; the addition of 1 for the parameter σ.

Table 1 is typical of the tables produced in information criteria analysis. It contains the log-likelihoods, the number of parameters, and for several common criteria, the IC and ΔIC values. To have *a priori* control of error, we need to specify a threshold for strong evidence. As with α, the size of NP tests, this threshold depends on the researchers needs. To match with scientific conventions, we set this threshold at a ΔIC value of 2. As we have seen above, this translates roughly to a bound on misleading evidence (see above) a probability of misleading evidence of M<0.05. From table 1, one

732 can make a number of observations that are very useful in framing our thinking about our
733 driving questions. 1) The ΔIC values are all >14 for the exponential model, confirming
734 quantitatively what is visually obvious from the figure that it is essentially impossible
735 that *P. Aurelia* is growing in a density independent fashion under the conditions of
736 Gause's experiment. 2) All of the information criteria give strong evidence against the
737 Gompertz as a potential best model given our threshold for strong evidence. 3) The
738 Ricker model is nested within the generalized Ricker, and the exponential within the
739 Ricker. As dictated by theory, the generalized Ricker has the highest log-likelihood
740 among these three models, but it is not the best model according to the information
741 criteria. 4) Different information criteria favor different models with different degrees of
742 strength. Both the SIC and the AICc indicate moderately strong evidence that the
743 generalized Ricker is not the best model. The evidence from the AIC is more equivocal
744 than that registered by the other two criteria. This may be an example of the tendency of
745 the AIC to over fit. Although not the case in this example, the rank order for some
746 models can change between different criteria. 5) The Ricker model has the lowest IC
747 value, indicating that it is the "best model", but the difference with the Beverton-Holt
748 model is small, thus the evidence that the Ricker model is superior to the Beverton-Holt
749 is very weak, and both models should be considered for prediction and interpretation, as
750 should the generalized Ricker and Gompertz to considerably lesser degrees 6) There are
751 three classes of non-nestable models in this problem. Classical likelihood ratio tests do
752 not compare across model families, thus an information criterion based analysis allows a
753 richer probing of nature. In this case we see that the Beverton-Holt model is essentially

754 indistinguishable in merit from the Ricker, at least for this population on the basis of this
755 data. We also see that there is strong evidence that the Gompetz is not the best model.



756

757 Figure 2: Observed population growth rate plotted population size. The lines are
758 expected growth rates for five fitted growth models. The data are the first of 3 replicate
759 times series for *Paramecium Aurelia* given in *The Struggle for Existence.* (Figure after
760 Figure 1 Lele and Taper 2012)

761

Model	LogLikelihood	# Parameters	AIC	AICc	SIC	ΔAIC	ΔAICc	ΔSIC
Ricker	4.90	3	-3.80	-1.96	-1.30	0.00	0.00	0.00
Beverton-Holt	4.82	3	-3.63	-1.79	-1.13	0.17	0.17	0.17
Generalized Ricker	4.91	4	-1.81	1.52	1.52	1.99	3.48	2.83
Gompertz	2.66	3	0.68	2.53	3.18	4.48	4.48	4.48
Exponential	-3.72	2	11.40	12.30	13.10	15.20	14.20	14.40

762

763 Table 1: Population dynamic model identification for Gause's *P. aurelia* using

764 information criteria.

765

766 **Common confusions about the three paradigms**

767 *What Is the Frequency in Frequentism?*

768 Frequentism is an overloaded term within the field of statistics referring both to a

769 definition of probability and to a style of inference. *Sensu stricto*, a frequentist is

770 someone who adheres to a frequency definition of probability, under which an event's

771 probability is long run limit that the event's relative frequency in a series of trials.

772 Another common use of the term frequentist is to describe a person who uses the

773 frequency of error in a decision rule as their principle warrant for inference. Sometimes

774 this branch of statistics is called "Classical Statistics", but this itself is a bad term because

775 Bayesian-like statistics considerably predated this approach. We have followed Deborah

776 Mayo (e.g. Mayo 1996) in referring to this style of inference as "error statistics".

Do Hierarchical Models Require a Bayesian Analysis?

Hierarchical models are not Bayesian. Hierarchical models are probabilistic models aiming at including two or more layers of uncertainty in the statistical model of how the data arises. Which includes latent variable and missing data problems (Dennis et al. 2006, Dennis and Ponciano 2014). Inference on Hierarchical models (HM) can in principle be made under all three approaches. However, maximum likelihood estimation of HM can be very difficult. Generally accessible computer implementations of Monte Carlo Markov chain (MCMC) algorithms made Bayesian estimation and inference broadly accessible in the 1990s. Although biological models with deep roots in stochastic processes, and in particular, Markov Chains had long been used in ecology and evolution (see Cohen 2004) by the 90's, the ease with which the Bayesian solutions yielded inferential conclusions of value for managers and practitioners quickly triggered a "Bayesian revolution" (Beaumont and Rannala 2004). This revolution prompted heated discussions between the proponents of frequentist and Bayesian statistics resulting in the marked growth of various biological scientific communities, including Ecology. As a result, topics of inference once deemed too difficult or almost inaccessible for practitioners, such as stochastic population dynamics modeling, have found a well-defined niche in Ecology (Newman et al 2014).

The drive to improve inference using Bayesian statistics has generated a plethora of technical novelties to sample from posterior distributions (like Approximate Bayesian Computation, see https://approximatebayesiancomputational.wordpress.com/), and even motivated novel approaches to ML estimation. Data Cloning (Lele et al. 2007, 2010) for

instance, is a recent algorithmic device inspired by Bayesian statistics that allows likelihood estimation by a simple algorithmic trick. It has long been known (Walker 1969) that in Bayesian analysis as the amount of data increases the posterior distribution converges to a normal distribution with the same mean and variance as the sampling distribution of the maximum likelihood estimate. The Lele et al. papers show that this same effect can be achieved simply by creating large data sets from multiple (say k) copies of an original data set (preserving data dependencies). The mean of the resulting posterior distribution approximates the maximum likelihood estimate, but the variance is too low. An estimate of the asymptotic variance is recovered by multiplying the variance of the posterior by k. These estimates can be made arbitrarily accurate by increasing k and the MCMC run length.

As presented above, inference is available through t-tests and Wald intervals, Ponciano et al. (2009) extend the data cloning inference tools to include information criterion based model selection, likelihood ratio tests and profile likelihood computations for hierarchical models relevant in Ecology. Using data cloning a full likelihood solution can be achieved for any hierarchical model

The R package dclone (Solymos 2010) provides easy access to data cloning to anyone who can write a Bayesian model in WinBugs, OpenBugs, or JAGS. Gimenez et al. (2014) attribute the rise of Bayesian applications in Ecology to the ease of software applications, and wonder what will be the consequence of readily available data cloning software. We would like to point out that Yamamura (2015) in this symposium introduces “empirical Jeffreys’ priors”, another computational device for achieving maximum likelihood inference for complex HM.

Are likelihood and probability the same thing?

This is a point that often confuses students making their first foray into mathematical statistics. The difficulty arises from the equation defining likelihood as $L(M_i; x) = f(x; M_i)$ and not as we do here the first time we present the likelihood function. The likelihood function is in fact proportional to the probability of observing the data under a given model. The left hand side of this equality is the likelihood while the right hand side is the probability, so they must be the same thing. Not at all, the likelihood is supposed to be understood as a function of the model (parameter) given the data, while probability is a function of the data given the model. This probability can be thought of as the long run frequency with which a mechanism would generate all the possible observable events, while the likelihood, or rather the relative likelihood, is the support in the data for certain value(s) of the parameter(s) of interest vis-à-vis other values.

The examples shown in this paper deal mostly with discrete probability models (the binomial distribution). In the case of continuous probability models, writing the likelihood function as the joint probability density function of the data evaluated at the observations at hand is not the exact likelihood function (i.e., it is not the joint probability of the observations evaluated at the data at hand). The joint probability density function is only an approximation introduced for mathematical convenience (Barnard 1967, Sprott 2000, Montoya et al 2008, 2009), one that works most of the time and hence advocated as the true likelihood function of continuous models in standard mathematical statistics books (e.g. Rice 1995). This approximation however sometimes leads to strange behavior

and singularities. For that, the likelihood has been sometimes critiqued. However, the likelihood is proportional to probabilities and for that, cannot have singularities. When these issues arise, Montoya et al (2009) show how returning to the original definition of the likelihood function, not the approximation, solves the problems.

Are confidence intervals and credible intervals really the same thing?

The error statistical confidence interval is constructed so that under repeated sampling of data confidence intervals constructed with the same method will contain the true value a specified f-probability of the time. The Bayesian credible interval is constructed so that in this instance the true value is believed to be within the interval with a specified b-probability. Thus, confidence intervals are really about the method, while credible intervals are about the instance. However, a confidence interval do also inform about the instance. A measurement made by a reliable method should be reliable. The width of a confidence interval is a function of the variance of the ML estimator of the parameter of interest (Rice 1995). If the data-gathering process is reliable and generates observations with high information content, then repeated instances of this sampling process will result in very similar estimators of the parameter of interest. In other words, the variance of this estimator over hypothetical repeated sampling will be small and the confidence interval will be narrow. The "confidence" then would stem from the reliability and repeatability of the conclusions.

A confidence interval informs that there is evidence that the instance is within the confidence interval (see Bandyophadyay et al. 2015 appendix chapter 2) . Many flavors of confidence intervals exist, but one most relevant to scientists is the one derived from

profile likelihoods, or relative profile likelihoods (Royall, 2000; Sprott, 2004). Profile likelihoods allow one to evaluate the verisimilitude of a set of values of the parameter of interest vis-à-vis the likelihood of the ML estimate. Intuitively, there is no reason why parameter values to the left or right of the ML estimate that are say, 85% as likely as the ML estimate shouldn't be considered. The evidential support built in the profile likelihood interval gives a continuous measure of the likelihood of nearness to the central value, which is as close as you can get to a credible interval without crossing the philosophical divide between frequentist and Bayesian definitions of probability.

A common criticism of the confidence interval relative to the credible interval is that they can include impossible values such as population sizes below the number of observed values. But these problems only occur in approximate confidence intervals. It is important to realize that this criticism does not apply to confidence intervals based on relative likelihoods or relative profile likelihoods (see Sprott, 2000 page 16).

Is Bayesianism the only paradigm that can use expert opinion?

The ability to incorporate expert opinion into the statistical analysis of ecological problems is often cited as one of strengths of the Bayesian approach (Kuhnert et al. 2010). Lele (2004b) and Lele and Allen (2006) show how to elicit pseudo data not priors from experts and to treat these as measurements with observation error. This approach is easier for experts than supplying priors. Further, the reliability of the experts can be probed in ways not available with elicited priors.

Is Bayesianism the only paradigm that allows updating?

892 The ability to "update" on the basis of new data has been stated (e.g. Ellison 2004) as a
893 major advantage of Bayesian analysis. However, as pointed out by van der Tweel (2005)
894 all three paradigms allow updating. What is updated differs, but in each case relates to
895 the paradigms core inferential process. A sequential Bayesian analysis updates belief, a
896 sequential evidential analysis updates evidence, and a sequential error statistical analysis
897 updates both the test statistic and critical values. Desiderata 6) in the sketch of evidential
898 statistics given above indicates that updating is one of the defining characteristics of the
899 evidential approach.
900
901 *Confusions about the interpretations of classes of information criteria*
902 There has been a long entrenched confusion in the literature about the interpretation of
903 information criteria. Bozdogan (1987) insightfully addressed this confusion, but
904 insufficient attention has been paid to it in the subsequent decades. Bozdogan noted that
905 every estimated model has error in it, and following Akaike, Bozdogan characterized this
906 error in terms of Kulback-Liebler (K-L) divergences. He then decomposed the total
907 divergence for an estimated model into two parts: 1) a divergence between the true
908 distribution and the model parameterized in the best possible manner given the
909 constraints of model structure, and 2) a further divergence due to errors of estimation.

910 This decomposition yields two reasonable but distinct targets for model
911 identification. The "minimum total discrepancy" forms (e.g. AIC and AICc) seek to
912 identify the model in the model set that, when estimated, will on average have the lowest
913 K-L divergence. The "order consistent" forms (e.g. CAIC, SIC, and ICHQ) seek to
914 identify the model in the model set that will have the lowest K-L divergence under best

possible parameterization. Asymptotically, both types of criteria achieve their goals. As a generality (but with exceptions in particular data sets and classes of problems) MTD forms tend to select models with slightly lower prediction mean squared errors, while order consistent forms tend to select models with somewhat less spurious complexity (see Taper 2004). These two classes of information criteria are sometimes referred to as "consistent" and "non-consistent". We prefer our terminology because "non-consistent" implies that the MTD forms are doing something wrong as opposed to just different.

The failure to understand the distinction in the targets of identification has led to a sea of wasted ink (e.g. Burnham et al., 2011) regarding the assumption purportedly required by the order consistent forms that the "true model" is in the model set. Mathematically, this assumption doesn't exist. We speculate that the origin of the myth derives from loose language in Schwarz' (1978) paper. When deriving the SIC (which he calls the BIC) Schwarz declared α_j, the prior for model j, to be the probability that model j is the true model. He immediately states that the specification of the priors doesn't matter because they are eliminated in the derivation. In fact, he could have just as well declared α_j to be "*the probability that model j is the model of best possible approximation*".

Does model choice inherently make frequentist statistics subjective?

There is some truth, but little sting to this criticism to frequentist statistics often raised by Bayesian scientists. Certainly, if we understand the world through the use of models; the models we actually use limit our understanding. Thus model choice does add a subjective element to science, which can influence the rate of gain of knowledge.

However, what knowledge is gained is objective. For the evidential statistician, this is most clear. The evidential statistician makes no claim to the truth of any of the models that investigated. This statistician only claims that given the data in hand one model is estimated to be closer to truth than another. This claim is entirely objective. Further, the subjective choice of models act as a challenge to other scientists to subjectively choose other models that may themselves objectively prove closer to truth. We return to these important points in our conclusions.

Error statistics also maintains objectivity, although in a more cumbersome fashion. The carefully wrought and strict definitions of NP and significance testing make it clear both that the evidence is conditional on the models considered, and that the tests make no claims as to the truth of any hypotheses. NP (1933) "Without hoping to know whether each separate hypothesis is true or false" thought that operational and temporary decisions should be made between models based on the data and objective criteria. Similarly, Fisher's significance tests only indicate when a model is inadequate and make no exhortation to belief in the model when it is not rejected. However, the claim to objectivity for error statistics is slightly weaker than that of evidential statistics because error probabilities are the primary evidential measure, and error probabilities are calculated assuming one of the models is true.

Problems in the use of the paradigms:

Difficulties in the relationships among P-values, error probabilities and evidence

The bulk of science has been done using as statistical tools Neyman-Pearson hypothesis tests and Fisherian significance tests of P-values. Much of this science has been solid,

which is amazing because both methods are seldom used the way they were intended. The NP test does not present output which can be interpreted as evidence. Neyman and Pearson were clear on this in labeling it a decision procedure. The size of the test, α, which is an *a priori* error rate, could be taken as a crude measure of evidence under the rubric of realiablism, but it is almost never reported. What is reported as a "P-value" is the minimum α that would have been rejected with the observed data. This value is not the size of the test, it isn't really evidence, and it isn't a *post hoc* type I error rate. There is a vast number of papers over many decades discussing these points, but Blume and Peipert 2003 is a good introduction. The persistence of this treatment of the NP test in the face of all statistical education and literature is informative. Scientists very much want to be able to design experiments and studies with modest *a priori* control of error rates, ***and*** they want a *post hoc* interpretation of evidence which is something more than accept/reject. The NP test does not give them both but evidential statistics does.

Another problem with the dominant error statistical procedures is that the evidence for or against a single model, H_0, represented by a Fisherian significance test is not commensurate with the evidence for or against that hypothesis when it is contrasted with an alternative model, H_1. This is known as the Lindley paradox. Lindley (1957) originally contrasted a significance test with a Bayesian comparison of two models. Interestingly, as with all Bayesian inference, how often the contradiction occurs depends on the priors set on the two models.

The Lindley paradox is not restricted to Bayesian analysis. The problem can be reconstructed comparing a P-value with a Neyman-Pearson test. The problem is that the significance test may indicate a rejection of H_0 when a comparison of the two models

indicates that there is more evidence for H_0 than for H_1. The converse can also be true, a significance test can fail to reject H_0 whereas a model comparison indicates that there is more evidence for H_1 than there is for H_0. For the general prosecution of science, this is a flaw, although in certain contexts, such as drug trials, which require a conservative "first do no harm" attitude, it is a design feature.

Having discarded the "true model" assumption, an evidentialist statistician has trouble thinking in terms of evidence for a single model. For the evidentialist, these attempts are better described as model adequacy measures (Lindsey, 2004). Basu et al. (2011) have recently published a technical treatment on the development and use of generalized distance measures for statistical inference. As pointed out by Taper and Lele (2004) evidence functions are the difference (or possibly ratio) of 2 model adequacies. Thus, the Basu et al. book can provide rich material for the construction of future evidence functions. Further, the model adequacy of the best model in a model set represents a limit on how much better a perfect model could do in representing the data.

Problems with error statistical inference and & sample size

It is a long standing joke that a frequentist, (really an error statistician) is someone happy to be wrong 5% of the time. This is more than just a joke – it is a reality. The way the control of error is built into error statistical tests implies that while the type I error doesn't increase when sample increase, it also doesn't decrease. Under the evidential paradigm, both error probabilities, the probability of strong misleading evidence, *M*, and the probability of weak evidence, *W*, go to zero as sample size increases (see Royall, 1997; Royall,2000). To illustrate this fact, in Figure 3 we present Royall's example where the

setting was as follows: The null hypothesis (model 1) was that the data is normally distributed with mean θ_1 and variance σ^2. The alternative is that the data is normally distributed, with the same variance but with mean $\theta_2 > \theta_1$. If the null hypothesis is true, then the sample mean $\overline{X} \sim \mathrm{N}\left(\theta_1, \sigma^2 / n\right)$ and the critical threshold at a level $\alpha = 0.05$ for the observed mean above which we would reject the null is given by $\overline{x}_{crit} = \frac{\sigma}{\sqrt{n}} z_\alpha + \theta_1 = \frac{\sigma}{\sqrt{n}} 1.645 + \theta_1$, where z_α is the percentile of a standard normal distribution so that $(1-\alpha)100\%$ of the area under the Gaussian curve lies to the left of it. In that case, the Type II error, or probability of observing a sample mean that happens to fall within the "failing to reject" region given that the true probability model is $\overline{X} \sim \mathrm{N}\left(\theta_2, \sigma^2/\mathrm{n}\right)$ is computed as $\Pr\left(\overline{X} \le \overline{x}_{crit}\right)$. On the other hand, the probabilities of misleading evidence and of weak evidence as a function of n in this case are computed respectively as

$$M(n) = \Pr\left(\ell_2 / \ell_1 > k\right) = \Pr\left(\ell_1 / \ell_2 > k\right),$$
$$W(n) = \Pr\left(1/k < \ell_2 / \ell_1 < k\right).$$

Using standard mathematical statistic results pertaining transformation of variables, these probabilities can be readily computed for various sample sizes, and a given cut-off k for the strength of evidence (see Royall 2000 and Figure 3).

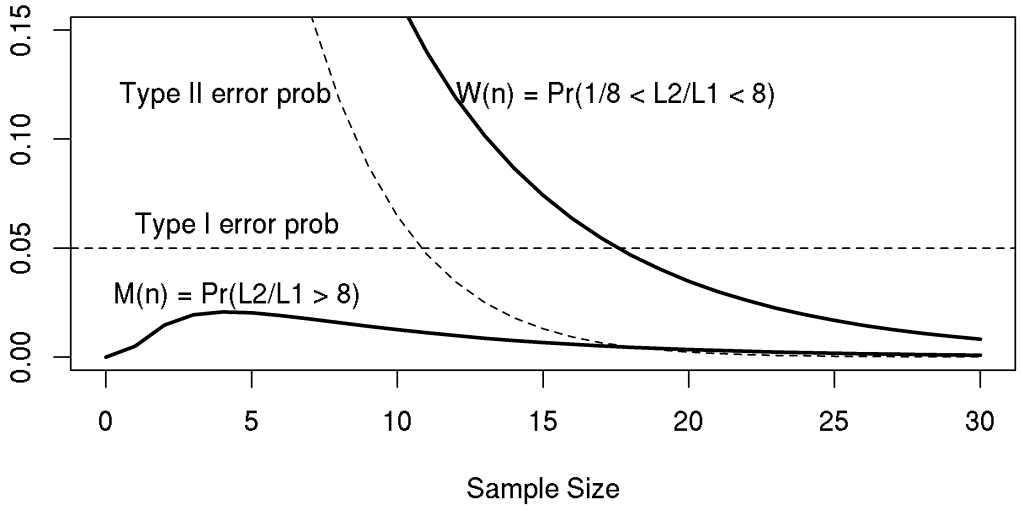


Figure 3. A comparison of the behavior with increasing sample size of Neyman-Pearson error rates (Type I and Type II) with evidential error rates (M and W). The critical distinction is that NP type I error remains constant regardless of sample size while both evidential error rates go to zero as sample size increases (Figure re-drawn after Royall (2000)'s Figure 2 using $|\theta_2 - \theta_1| = \sigma = 15; \theta_1 = 100; k = 8; \alpha = 0.05.$).

Fisherian significance also has sample size difficulties. In this case, it is with the interpretation of a P-value as the strength of evidence against a model. The common practice of science implicitly assumes that a P-value from one study implies more or less the same degree of evidence against the null hypothesis that the same P-value from another study would even if the two studies have different sample sizes. Unfortunately this isn't true. But, how the evidence varies with sample depends on subtleties of the scientist's interpretation of the procedure. If you impose a significance level and treat every P-value greater than the level simply as exceeding the level than there is greater

evidence against the null in small samples than in large. If on the other hand, the scientist is directly comparing P-values without an a priori cut off, then there is greater evidence in large samples than small samples for a given P-values. In either case the evidence depends on sample size making a hash of interpretation of published work (see Royall 1986 for further details).

Bayesian Difficulties with non-identifiability

A model is said to be non-estimable if the maximum value of the likelihood function evaluated at the data occurs for more than one different sets of parameters. That is to say that the data can't be used to distinguish between multiple possible estimates. If this failure is not due to a quirk of sampling, but is instead determined by the way the model is configured, then a model is said to non-identifiable if it is non-estimable for all possible data sets.

Non-estimability may cause programs that calculate maximum likelihood estimates through numerical optimization to return an error. This is generally annoying, but is an important indication that something is wrong with the way you are modeling your data.

A Bayesian estimation on the other hand will be completely oblivious to the non-estimability. Bayesian estimates are a combination of information from the data and information from the prior beliefs. The hope is that information from the data will swamp that in the prior:

> "Specification of the prior distribution can be viewed as the 'price' paid for the exactness of inferences computed using Bayes Theorem. When the sample size is low, the price of an exact inference may be high. As the size of a sample increases the price of an exact inference declines because the information in the data eventually exceeds the information in the prior" Royle & Dorazio. 2008. Hierarchical Modeling and Inference in Ecology. Page 55

1063 However, this is not always true. In the case of non-estimability/non-
1064 identifiability there is no information in the data to distinguish between alternative
1065 estimates, and the decision is made entirely on the basis of the prior. Often with complex
1066 hierarchical models where non-estimability/non-identifiability might occur is not
1067 obvious.
1068 As mentioned above, data-cloning is a method of transforming a Bayesian
1069 analysis into a likelihood analysis. In situations where non-estimability/non-
1070 identifiability is suspected, this is particularly useful. A data cloned estimation will
1071 return estimates of estimable parameters and diagnostics indicating that non-
1072 identifiability exists in the remainder (Lele et al. 2010; Ponciano et al. 2012).
1073
1074 *Informative, non-informative or mis-informative priors?*
1075 As our sketch of Bayesian inference indicates, a specified prior is mandatory for
1076 Bayesian calculations. To avoid "subjectivity" many Bayesian scientists prefer to
1077 employ "objective" or "non-informative" priors.

1078 "To compute the posterior distribution, the Bayesian has to prescribe a prior
1079 distribution for θ, and this is a model choice. Fortunately, in practice, this is
1080 usually not so difficult to do in a reasonably objective fashion. As such, we view
1081 this as a minor cost for being able to exploit probability calculus to yield a
1082 coherent framework for modeling and inference in any situation."

1083 Royle & Dorazio. 2008. Hierarchical Modeling and Inference in Ecology. Page
1084 21

1085 The problem, is that what constitutes a non-informative prior depends on how the model
1086 is parameterized (Fisher, 1922). Lele (2015) analyses 2 important ecological problems
1087 with simulated and real data sets. Each problem has multiple equivalent and commonly
1088 used parameterizations. Lele analyses population persistence projections for the San

Joaquin kit fox using a Ricker equation parameterized in terms of growth rate and density dependence (a, b) or in terms of growth rate and carrying capacity (a, K). The two forms are mathematically equivalent. However, Bayesian estimation using "non-informative" priors yield very different parameters estimates and very different predictions of population persistence. Similarly occupancy models for the American toad can be parameterized either in terms of probabilities of occupancy and detection, or in terms of the logits of those quantities. Both formulizations are commonly used in studying occupancy. Again parameter estimates and posterior distributions from Bayesian estimates using non-informative priors are substantially different. Lele (2015) further demonstrates that the maximum likelihood estimates for these problems achieved through data cloning are transformation invariant.

While many statistical ecologists (e.g. Clark, 2005) agree with Royle and Dorazio that non-informative priors are benign, other eminent statisticians are much more cautious. Bradley Efron, a major proponent of empirical Bayes, closes a 2013 article with the statement: "be cautious when invoking uninformative priors. In the last case, Bayesian calculations cannot be uncritically accepted and should be checked by other methods, which usually means frequentistically." Gelman and Shalizi (2013) also strongly argue for frequentist/falsificationist checking of Bayesian solutions, and go as far as saying that

> "the idea of Bayesian inference as inductive, culminating in the computation of the posterior probability…has had malign effects on statistical practice. At best, the inductivist view has encouraged researchers to fit and compare models without checking them; at worst, theorists have actively discouraged practitioners from performing model checking because it does not fit into their framework".
>
> Gelman and Shalizi, 2013.

We are of the opinion that, while doing Bayesian statistics, practitioners should run frequentist checks on the validity of the inferences, despite the computational cost of so doing. By frequentist checks here we mean running a large number of simulations under the model (i.e. a parametric bootstrap) or a more complex setting where truth is known (i.e. a model structure adequacy analysis sensu Taper et al 2008) so that the reliability of the inferences with the posterior distribution can be assessed.

The true model assumption and the difficulty of using probability as a measure of evidence

A cryptic but fundamental assumption of Bayesian analysis is that the true model is in the model set. This is obvious because probabilities sum to 1. But, this flies in the face of our experience as scientists, modelers and statisticians. To quote George Box (1976) "All models are wrong." For us, if all models are wrong, what sense does it make to believe in any of them? If you don't believe in models, what sense does it make to depend on a statistical system predicated on belief in models? However, doubt about belief is not share uniformly by scientists as evidenced by this quote from an unpublished manuscript by an ecologist.

> "Frequentists never explicitly state how their metrics such as P-values and confidence intervals should be translated into belief about the strength of evidence, although such translation is clearly being done (otherwise data analysis is pointless if it is not informing belief). This is why I view the frequentist approach as subjective; there is no theory for how frequentist metrics should be translated into belief, so clearly the interpretation of frequentist metrics in terms of strength of evidence and belief must be subjective."

This ecologist believes in belief so strongly as to essentially accuse frequentists of lying when they say they don't.

Interestingly, some Bayesian statisticians concur with us. Gelman and Shalizi (2013) state: "It is hard to claim that the prior distributions used in applied work represent statisticians' states of knowledge and belief before examining their data, if only because most statisticians do not believe their models are true, so their prior degree of belief in all of Θ is not 1 but 0." Clearly, for these statisticians Bayesian statistics simply represents a very convenient calculation engine. G.A. Barnard (1949) made a more psychological point when he said:

> "To speak of the probability of a hypothesis implies the possibility of an exhaustive enumeration of all possible hypotheses, which implies a degree of rigidity foreign to the true scientific spirit. We should always admit the possibility that our experimental results may be best accounted for by a hypothesis which never entered our own heads."
>
> G.A. Barnard (1949)

What does it do to us as scientists to continually condition ourselves to believe that our little systems comprehend reality?

Bayesian aspects of Akaike weights

Akaike weights are very important in so called frequentist model averaging (Burnham and Anderson 2002). They are the weights used in averaging models. However, as pointed out by Burnham and Anderson (2004) Akaike weights are posterior probabilities based on subjective priors of the form

$$q_i = C \cdot \exp\left(\frac{1}{2} K_i \log(n) - K_i \right),$$

where q_i is the prior for model i, C is a normalization constant, K_i is the number of parameters in the model, and n is the number of observations. This prior is a b-

probability, and as consequence so are Akaike weights. Thus, Burnham and Anderson's model averaging depends on a subjectively chosen prior, and as such inherits all of the justified criticism of such priors.

Burnham and Anderson like this prior a great deal. They call it a *savvy prior* (their emphasis). The prior they favor captures the Burnham and Anderson world-view very well. Plotting this prior as a function of the number of parameters in model *i*, it is easy to see that if you have more than 8 observations this prior is in fact an "anti-parsimony" prior, where models of more parameters are being favored *a priori* over models with fewer.

Priors as practical regularization devices

A class of intractable estimation problems using likelihood inference can be rendered tractable using subjective Bayesian statistics. Suppose we were wishing to estimate both, the probability of success p in a binomial trial whose total number of trials is unknown. In such cases, and depending on the values of p, the profile likelihood for the total number of trials N may not be well behaved and result in confidence limits with an infinite upper bound (Montoya 2008). In that case, as in similar species richness estimation problems (Christen and Nakamura 2000), subjective prior elicitation results in reliable inferences that have found applications in planning of biodiversity studies (Christen and Nakamura 2000).

This is not to say the only way to control a badly behaving likelihood is through a prior. Moreno and Lele (2010) were able to greatly improve the performance of site

occupancy estimation using penalized likelihood. Some statisticians claim that penalized likelihood is equivalent to some prior (Wang and Lindsay, 2005). In Moreno and Lele's case, they penalized to an alternative estimator based on the same data so no belief or prior information was involved.

Using the paradigms

Statistics as a means to clarify arguments

There is a strong impulse among ecologists to seek a statistical paradigm that is true and exact and will make all their analyses beautiful. No such paradigm exists. No paradigm is bullet proof, and no paradigm applies to all situations. Science works by making demonstrations through evidence based arguments (Gelman and Hennig 2015). Statistics functions in science to quantify and clarify those arguments. Different statistical paradigms can be applied to different scientific arguments.

Scientists are not used to thinking about the merits of statistical paradigms usefully. Scientist judge scientific theories by how well they match an external reality. But, all statistical methods exist in the mind only, there is no external reality against which to judge them. Statistical methodologies are to be judged as tools. Are they useful in the construction of sound scientific arguments or are they not?

The Central Task of Science

We hold the view that models carry the meaning in science (Frigg, 2006; Giere, 2004; 2008). Less radical views, of models such as that they represent reality Giere, 1988; 1999; 2004; Hughes, 1997; Morgan, 1999; Suppe, 1989; van Fraassen, 1980; 2002) or

1209 serve as tools for learning about reality (Giere, 1999; Morgan, 1999) all still give a very
1210 central place to models in science.
1211 Consequently, the job of scientists is to replace old (possibly good models) with
1212 new better models. When we have taught courses in both ecological modeling and
1213 statistical modeling our primary instruction is always: "Never fall in love with your
1214 model – it should not be a long relationship." Even if a scientist's interest is primarily in
1215 parameter values, Model identification is paramount. Without a good model, parameter
1216 estimation will be faulty.
1217 Evidential statistics gives the scientist tools to choose among the models he has
1218 and motivation to formulate new ones. Evidential statistics is a complete framework. It
1219 encompasses: The design of experiments and the control of error, post data assessment of
1220 the strength of inference, model identification, the comparison of models, assessment of
1221 model uncertainty, parameter estimation, and assessment of estimate uncertainty.
1222
1223 *Communicating about models: Public versus Personal Epistemology*
1224 Science knows much more than any individual scientist. Science has learned much more
1225 than any individual scientist has ever learned. This knowledge has accumulated over
1226 thousands of years through a complex web of transmission, colleague to colleague and
1227 teacher to student. Science is a public epistemology.
1228 Belief is personal and difficult to transfer. Belief also depends strongly on such
1229 individual things as cultural background and present mood. Evidence, on the other hand,
1230 is independent of the individual, transferable, and can accumulate. As such it is much
1231 better suited to form the basis of a public epistemology than is belief. Personal belief,

1232 although critically important for conducting first-person epistemology, needs to be
1233 strengthened with incorporation of data and information gathered from objectively
1234 grounded research to meet the demand of ever-growing science. Scientific epistemology,
1235 on the other hand is public, and is based on the transferrable and accumulation of
1236 information from many people and over great periods of time (See Strevens, 2010).
1237 However, the growth of scientific knowledge is not divorced from personal beliefs.
1238 Scientists are people, and create their research programs informed by their personal
1239 beliefs.
1240
1241 *The Character and Contributions of Statistical Paradigms.*
1242 Each of the statistical paradigms discussed has its own character and can make
1243 contributions to science. Error statistics, for instance, has been the backbone of science
1244 for a hundred years. Undoubtedly, it will continue to make major contributions in the
1245 21st century. There are inherent conservative biases in error statistics generated by the
1246 focus on the null hypotheses and the pre-specification of error rates. This conservative
1247 bias makes error statistics well suited for application in regulatory situations, medical
1248 science, and legal testimony, all fields that ethically mandate a similar bias.
1249 Evidential statistics, while still retaining control of error, places all models on
1250 equal footing. These properties and its focus on models make us feel that the evidential
1251 paradigm is best suited for the prosecution of general science. Nevertheless, when we are
1252 consulting for people answering to regulatory agencies, all of our evidential statistics get
1253 packed away, and out comes an error statistical tool kit.

Although we personally find the belief based philosophical foundations of Bayesian statistics unsound to support science as a public epistemology (this includes both subjective and objective Bayesian approaches), a lot of good work has been done with Bayesian statistics. A Bayesian analysis unchecked by frequentist methods runs the risk of undetected catastrophic failure, but in practice, much of the time it will be fine. Even if one seeks to avoid the use of a belief-based probability definition, an understanding of Bayesian methods in the analysis of hierarchical models is absolutely necessary. Most of the alternative methods for solving complex problems in science, empirical Bayes, data cloning, and empirical Jeffreys' priors all require a solid grounding in Bayesian methods.

It is our opinion that the epistemological high ground is now held by evidential statistics. We look forward to developments that will further evidential statistics, and someday lead to something that supplants it. Currently, most of the purported advantages of both error statistics and Bayesian statistics are now held by evidential statistics. This is by design; the framers of evidential statistics have ruthlessly borrowed what was good and rejected what was faulty. Many of the key ideas in evidential statistics were pioneered by its predecessors.

The central theme of this essay is that there is no magic wand for scientists in statistics. If one wants to use statistics effectively in science, then one needs to learn how to clarify scientific arguments with statistical arguments. To do that one needs to understand how the statistical arguments work. In many ways, this is a much harder task than mastering statistical methods. There are a number of excellent sources to help with

this task. As a beginning, we suggest: Royall, 1997; Barnett, 1999; Sprott, 2000; Taper and Lele, 2004; Thompson, 2007, and Bandyopadhyay et al. 2015.

Acknowledgements: We thank Dr. Yukihiko Toquenaga for inviting MLT to present this article in a plenary symposium of the 30th Annual Meeting of the Society of Population Ecology in Tsukuba, Japan. We also grateful to the Society and to the University of Tsukuba for providing funding for MLT travel expenses and our publication costs. MLT was partially supported by US National Science Foundation grant # DUE-1432577. JMP was partially supported by US National Institute of Health grant # R01 GM103604. We thank Ian Ausprey, Juan Pablo Gomez, Brian Dennis, and Robert Holt for insightful comments and useful suggestion helping to improve earlier drafts of this manuscript. We also would like to thank Jack Sullivan for his questions about information criteria, and Tessa Barton for her questions about the subjectivity of model choice. JMP would like to thank MLT for inviting him to co-author this paper. MLT would like to thank Prasanta Bandyopadhyay and Gordon Brittan for many discussions on the philosophy of statistics during the production of Bandyopadhyay et al. This paper and that work were produced simultaneously and ideas have filtered between the two. Any precedence will be an accident of publication. The authors wish to thank also the constructive critiques of Michael J. Lew and another anonymous reviewer.

References:

Aho, K., D. Derryberry, and T. Peterson. 2014. Model selection for ecologists: the worldviews of AIC and BIC. Ecology **95:**631-636.

Akaike, H. 1974. A new look at statistical-model identification. IEEE Transactions on Automatic Control AC19:716-723.

Bandyopadhyay, P. S., M. L. Taper, and G. J. Brittan. 2015 (anticipated). Belief, Evidence, and Uncertainty: Problems of Epistemic Inference. Springer.

Barnard, G. A. 1949. Statistical Inference. Journal of the Royal Statistical Society Series B-Statistical Methodology **11:**115-149.

Barnard, G.A. 1967. The use of the likelihood function in statistical practice. Proceedings of the 5th Berkeley symposium on Mathematical Statistics and Probability, Vol. I. Eds. L. Le Cam and J. Neyman.

Barnett, V. 1999. Comparative Statistical Inference. Third edition. John Wiley & Sons, LTD, Chinchester.

Basu, A., H. Shioya, and C. Park. 2011. Statistical Inference: The Minimum Distance Approach. Chapman & Hall (CRC Press).

Beaumont, Mark A., and Bruce Rannala. 2004. The Bayesian revolution in genetics. Nature Reviews Genetics **5(4):** 251-261.

Blume, J., and J. F. Peipert. 2003. What your statistician never told you about P-values. Journal of the American Association of Gynecologic Laparoscopists **10:**439-444.

Box, G. E. P. 1976. Science and Statistics. Journal of the American Statistical Association **71:**791-799.

1320 Bozdogan, H. 1987. Model Selection and Akaike Information Criterion (AIC) - the
1321 General-Theory and Its Analytical Extensions. Psychometrika **52:**345-370.
1322 Burnham, K. P., and D. R. Anderson. 2002. Model Selection and Multi-model Inference:
1323 A Practical Information-Theoretic Approach. 2nd edition. Springer-Verlag, New
1324 York.
1325 Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference - understanding AIC
1326 and BIC in model selection. Sociological Methods & Research 33:261-304.
1327 Burnham, K. P., D. R. Anderson, and K. P. Huyvaert. 2011. AIC model selection and
1328 multimodel inference in behavioral ecology: some background, observations, and
1329 comparisons. Behavioral Ecology and Sociobiology 65:23-35.
1330 Chatfield, C. 1995. Model Uncertainty, Data Mining and statistical Inference. Journal of
1331 the Royal Statistical Society A **158:**419-466.
1332 Clark, J. S. 2005. Why environmental scientists are becoming Bayesians. Ecology
1333 Letters **8:**2-14.
1334 Cohen, J. E. 2004. Mathematics is biology's next microscope, only better; Biology is
1335 mathematics' next physics, only better. Plos Biology **2:**2017-2023.
1336 Christen, J. A., & Nakamura, M. 2000. On the analysis of accumulation curves.
1337 Biometrics ***56*(3):** 748-754.
1338 Dennis, B. 2004. Statistics and the scientific method in ecology (with commentary). The
1339 nature of scientific evidence: statistical, philosophical, and empirical
1340 considerations. University of Chicago Press, Chicago, Illinois, *USA*, 327-378.
1341 Dennis, B., and J. M. Ponciano. 2014. Density-dependent state-space model for
1342 population-abundance data with unequal time intervals. Ecology **95:**2069-2076.

1343 Dennis, B., J. M. Ponciano, S. R. Lele, M. L. Taper, and D. F. Staples. 2006. Estimating
1344 density dependence, process noise, and observation error. Ecological Monographs
1345 **76:**323-341.
1346 Dorazio, R. M. 2015. TITLE AND CITATION TO BE SUPPLIED BY PE EDITORS
1347 Edwards, A. W. F. 1992. Likelihood. Expanded Ed. Johns Hopkins University Press,
1348 Baltimore.
1349 Efron, B. 2010. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing,
1350 and Prediction Institute of Mathematical Statistics Monographs, Cambridge Univ.
1351 Press, Cambridge, UK.
1352 Efron, B. 2013. Bayes' Theorem in the 21st Century. Science 340:1177-1178.
1353 Ellison, A. M. 2004. Bayesian inference in ecology. Ecology Letters **7**:509-520.
1354 Fisher, R. A. 1922. On the mathematical foundations of theoretical statistics.
1355 Philosophical Transactions of the Royal Society, A 222:309-368.
1356 Fisher, R. A. 1956. Statistical Methods and Scientific Inference. Oliver and Boyd,
1357 London.
1358 Fisher, R. A. 1971. The Design of Experiments. Eighth edition. Hafner Publishing
1359 Company, New York.
1360 Frigg, R. 2006. Scientific Representation and the Semantic View of Theories. Theoria
1361 **55:**49-65.
1362 Gause, G. F. 1934. The struggle for existence. Williams and Wilkins, Baltimore,
1363 Maryland, USA.
1364 Gelman, A., and C. R. Shalizi. 2013. Philosophy and the practice of Bayesian statistics.
1365 British Journal of Mathematical & Statistical Psychology **66**:8-38.

Gelman, A., and C. Hennig. 2015. Beyond subjective and objective in statistics. Columbia University Department of Statistics technical report.

Giere, R. 1988. Explaining Science. University of Chicago Press, Chicago.

Giere, R. N. 1999. Science without laws (Science and Its Conceptual Foundations). University of Chicago Press, Chicago.

Giere, R. N. 2004. How models are used to represent reality. Philosophy of Science **71:**742-752.

Giere, R. N. 2008. Models, Metaphysics, and Methodology.in S. Hartmann, L. Bovens, and C. Hoefer, editors. Nancy Cartwright's Philosophy of Science. Routledge.

Gimenez, O., S. T. Buckland, B. J. T. Morgan, N. Bez, S. Bertrand, R. Choquet, S. Dray, M.-P. Etienne, R. Fewster, F. Gosselin, B. Merigot, P. Monestiez, J. M. Morales, F. Mortier, F. Munoz, O. Ovaskainen, S. Pavoine, R. Pradel, F. M. Schurr, L. Thomas, W. Thuiller, V. Trenkel, P. de Valpine, and E. Rexstad. 2014. Statistical ecology comes of age. Biology letters **10**:20140698.

Guttorp, P. 1995. Stochastic Modeling of Scientific Data. Chapman & Hall, London.

Hacking, I. 1965. Logic of statistical inference. Cambridge University Press., Cambridge.

Hájek, A., "Interpretations of Probability", The Stanford Encyclopedia of Philosophy (Winter 2012 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/win2012/entries/probability-interpret/

Hannan, E. J., and B. G. Quinn. 1979. Determination of the order of an autoregression. Journal of the Royal Statistical Society Series B-Methodological 41:190-195.

Hughes, R. I. G. 1997. Models and Representation. Philosophy of Science (Proceedings) 64: 325-336.

1389 Hurvich, C. M., and C. L. Tsai. 1989. Regression and Time-Series Model Selection in
1390 Small Samples. Biometrika 76:297-307.
1391 Kass, R. E. and A. E. Raftery. 1995. Bayes factors. Journal of the American Statistical
1392 Association 90: 773-795.
1393 Kalbfleisch, J.G. 1985. Proability and Statistical Inference. Volume II: Statistical
1394 Inference. 2nd Edition. Springer-Verlag.
1395 Kolmogorov, A. N., 1933, *Grundbegriffe der Wahrscheinlichkeitrechnung, Ergebnisse*
1396 *Der Mathematik*; translated as *Foundations of Probability*, New York: Chelsea
1397 Publishing Company, 1950.
1398 Kuhnert, P. M., T. G. Martin, and S. P. Griffiths. 2010. A guide to eliciting and using
1399 expert knowledge in Bayesian ecological models. Ecology Letters **13**:900-914.
1400 Lele, S. R. 2004a. Evidence Functions and the Optimality of the Law of Likelihood.*in* M.
1401 L. Taper and S. R. Lele, editors. The Nature of Scientific Evidence: Statistical,
1402 Philosophical and Empirical Considerations. The University of Chicago Press,
1403 Chicago.
1404 Lele, S. R. 2004b. Elicit Data, Not Prior: On Using Expert Opinion in Ecological
1405 Studies.*in* M. L. Taper and S. R. Lele, editors. The Nature of Scientific Evidence:
1406 Statistical, Philosophical and Empirical Considerations. The University of
1407 Chicago Press, Chicago.
1408 Lele, S. R. 2015 (submitted). Is non-informative Bayesian analysis appropriate for
1409 wildlife management: survival of San Joaquin Kit Fox and declines in amphibian
1410 populations. Methods in Ecology and Evolution.

1411 Lele, S. R., and K. L. Allen. 2006. On using expert opinion in ecological analyses: a
1412 frequentist approach. Environmetrics **17**:683-704.
1413 Lele, S. R., B. Dennis, and F. Lutscher. 2007. Data cloning: easy maximum likelihood
1414 estimation for complex ecological models using Bayesian Markov chain Monte
1415 Carlo methods. Ecology Letters **10:**551–563.
1416 Lele, S. R., K. Nadeem, and B. Schmuland. 2010. Estimability and Likelihood Inference
1417 for Generalized Linear Mixed Models Using Data Cloning. Journal of the
1418 American Statistical Association **105**:1617-1625.
1419 Lele, S. R. and M. L. Taper. Information Criteria in Ecology. In Sourcebook in
1420 Theoretical Ecology. 2012. A. Hastings and L. Gross editors University of
1421 California Press.
1422 Lindblom, C. E. 1959. The Science of Muddling Through. Public Administration Review
1423 **19**:79-88.
1424 Lindley, D. V. 1957. A Statistical Paradox. Biometrika **44**:187-192.
1425 Lindley, D. V. 2000. The philosophy of statistics. Journal of the Royal Statistical Society
1426 Series D-the Statistician **49:**293-319.
1427 Lindsay, B. G. 2004. Statistical distances as loss functions in assessing model adequacy.
1428 Pages 439-488 in M. L. Taper and S. R. Lele, editors. The nature of scientific
1429 evidence: Statistical, philosophical and empirical considerations. The University
1430 of Chicago Press, Chicago.
1431 Mayo, D. G. 1996. Error and the Growth of Experimental Knowledge. University of
1432 Chicago Press, Chicago.

1433 Mayo, D. G., and D. R. Cox. 2006. Frequentist statistics as a theory of inductive
1434 inference. Pages 77-97 in Optimality: The 2nd Lehmann Symposium. Institute of
1435 Mathematical Statistics Rice University.
1436 Mayo, D. G., and A. Spanos. 2006. Severe testing as a basic concept in a Neyman-
1437 Pearson philosophy of induction. British Journal for the Philosophy of Science
1438 **57:**323-357.
1439 Montoya, J.A. 2008. La verosimilitud perfil en la inferencia estadística. Doctoral
1440 Dissertation, Center for Research in Mathematics, Guanajuato, México.
1441 Moreno, M., and S. R. Lele. 2010. Improved estimation of site occupancy using
1442 penalized likelihood. Ecology **91:**341-346.
1443 Montoya, J. A., Díaz-Francés, E. and D. A. Sprott. 2009. On a criticism of the profile
1444 likelihood function.Statistical Papers 50: 195-202.
1445 Morgan, M. 1999. Learning from Models. Pages 347-388 in M. Morrison and M.
1446 Morgan, editors. Models as Mediators: Perspectives on Natural and Social
1447 Science. Cambridge University Press, Cambridge.
1448 Newman, K. B., Buckland, S. T., Morgan, B. J., King, R., Borchers, D. L., Cole, D. J., ...
1449 & Thomas, L. (2014). Modelling Population Dynamics. Springer.
1450 Neyman, J., and E. S. Pearson. 1933. On the problem of the most efficient tests of
1451 statistical hypostheses. Philosophical Transactions of the Royal Society of
1452 London Series A 231:289-337.
1453 Pawitan, Y. 2001. In All Likelihood: Statistical Modeling and Inference Using
1454 Likelihood. Oxford University Press, Oxford.

1455 Peirce, C. S. 1878. Illustrations of the Logic of Science III —The doctrine of chances. .
1456 Popular Science Monthly 12:604-615.
1457 Ponciano, J. M., G. Burleigh, E. L. Braun, and M. L. Taper. 2012. Assessing parameter
1458 identifiability in phylogenetic models using Data Cloning. Systematic Biology
1459 **61**:955-972.
1460 Ponciano, J. M., M. L. Taper, B. Dennis, and S. R. Lele. 2009. Hierarchical models in
1461 ecology: confidence intervals, hypothesis testing, and model selection using data
1462 cloning. Ecology **90:**356-362.
1463 Popper, K. R. 1959. The propensity interpretation of probability. British Journal for the
1464 Philosophy of Science 10:25-42.
1465 Rannala, Bruce. 2002. Identifiability of parameters in MCMC Bayesian inference of
1466 phylogeny. Systematic Biology 51: 754-760.
1467 Raftery, A. E. 1995. Bayesian model selection in social research. Sociological
1468 methodology 25: 111-164.
1469 Rice, J.A. 1995. Mathematical Statistics and Data Analysis. 2nd edition. Duxbury Press,
1470 Belmont, California.
1471 Royall, R. M. 1986. The effect of sample-size on the meaning of significance tests.
1472 American Statistician 40:313-315.
1473 Royall, R. 1997. Statistical Evidence: A likelihood paradigm. Chapman & Hall, London.
1474 Royall, R. 2000. On the Probability of Observing Misleading Statistical Evidence.
1475 Journal of the American Statistical Association **95:**760-780.
1476 Royall, R. M. 2004. The Likelihood Paradigm for Statistical Evidence. Pages 119-152 *in*
1477 M. L. Taper and S. R. Lele, editors. The Nature of Scientific Evidence:

1478 Statistical, Philosophical and Empirical Considerations. The University of
1479 Chicago Press, Chicago.
1480 Royle, J. A., and R. M. Dorazio. 2008. Hierarchical Modeling and Inference in Ecology:
1481 The Analysis of Data from Populations, Metapopulations and Communities.
1482 Academic Press, San Deigo.
1483 Solymos, P. 2010. dclone: Data Cloning in R. The R Journal 2:29-37.
1484 Schwarz, G. 1978. Estimating the dimension of a model. Annals of Statistics 6:461-464.
1485 Sprott, D. A. 2000. Statistical Inference in Science. Springer-Verlag, New York.
1486 Sprott, D. A. 2004. What is optimality in scientific inference? Pages 133-152 in J. Rojo
1487 and V. PerezAbreu, editors. First Erich L. Lehmann Symposium - Optimality.
1488 Strevens, M. (2010): "Reconsidering authority: Scientific expertise, bounded rationality,
1489 and epistemic backtracking." In T. Szabo Gendler and J. Hawthorne, eds., Oxford
1490 Studies in Epistemology Vol. 3, Chap. 13, New York: Oxford University Press.
1491 Suppe, F. 1989. The Semantic Conception of Theories and Scientific Realism. University
1492 of Chicago Press. , Chicago.
1493 Taper, M. L. 2004. Model identification from many candidates. Pages 448-524 in M. L.
1494 Taper and S. R. Lele, editors. The Nature of Scientific Evidence: Statistical,
1495 Philosophical and Empirical Considerations. The University of Chicago Press,
1496 Chicago.
1497 Taper, M. L., and S. R. Lele. 2004. The nature of scientific evidence: A forward-looking
1498 synthesis. Pages 527-551 in M. L. Taper and S. R. Lele, editors. The Nature of
1499 Scientific Evidence: Statistical, Philosophical and Empirical Considerations. The
1500 University of Chicago Press, Chicago.

1501 Taper, M. L., and S. R. Lele. 2011. Evidence, Evidence Functions, and Error
1502 Probabilities. Pages 513-532 in P. S. Bandyopadhyay and M. R. Forster, editors.
1503 Philosophy of Statistics.
1504 Taper, M. L., D. F. Staples, and B. B. Shepard. 2008. Model structure adequacy analysis:
1505 selecting models on the basis of their ability to answer scientific questions.
1506 Synthese 163:357-370.
1507 Thompson, B. 2007. The Nature of Statistical Evidence. Springer, New York.
1508 Underwood, A. J. 1997. Experiments in Ecology: Their Logical Design and Interpretation
1509 Using Analysis of Variance. Cambridge University Press, Cambridge.
1510 van der Tweel, I. 2005. Repeated looks at accumulating data: To correct or not to correct?
1511 European Journal of Epidemiology **20**:205-211.
1512 van Fraassen, B. 1980. . The Scientific Image. Oxford University Press, Oxford.
1513 Van Fraassen, B. 2002. The Empirical Stance. Yale University Press, New Haven and
1514 London.
1515 Venn, J., 1876, *The Logic of Chance*, 2nd edition, London: Macmillan; reprinted, New
1516 York: Chelsea Publishing Co., 1962.
1517 von Mises, R. 1928 (German 1st edition), 1957 (English Translation of 3rd edition).
1518 Probability, Statistics, and Truth. George Allen & Unwin LTD. London.
1519 Walker, A. M. 1969. On asymptotic behaviour of posterior distributions. Journal of the
1520 Royal Statistical Society Series B-Statistical Methodology 31:80-&.
1521 Wang, J. P. Z., and B. G. Lindsay. 2005. A penalized nonparametric maximum likelihood
1522 approach to species richness estimation. Journal of the American Statistical
1523 Association **100:** 942-959.

Wilks, S.S. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. The Annals of Mathematical Statistics 9:60-62.

Yamamura, K. 2015 TITLE AND CITATION TO BE SUPPLIED BY PE EDITORS.