Ecological Applications, 6(4), 1996, pp. 1090–1094

COMMENT: THE FIRST DATA ANALYSIS SHOULD BE JOURNALISTIC[1,2]

DON EDWARDS
Department of Statistics, University of South Carolina, Columbia, South Carolina 29208 USA

Abstract. Bayesian statistical methods can be considered an attempt at mathematical formalization of the natural scientific process of interpretation of data in light of preexisting information. As such, their use, and the degree to which they are used, is largely a question of efficiency. In some instances it may be appropriate to incorporate prior information into an analysis to the extent that this information is deemed reliable by all concerned; it will not often be the case in an ecological study, however, that information satisfying these constraints is substantial. In the most important distributional setting a frequentist confidence interval is identical to a noninformative Bayesian credible interval, and it is asserted that in most other cases where noninformative priors are used, these two will be very similar; the primary data analysis in an ecological study should probably be of one of these two forms. It is conjectured that, in order to be mathematically tractable, decision theoretic methods (Bayesian or not) will often deal with a dangerously short action–space time frame. Finally, Empirical Bayesian methods and hierarchical models in general are powerful new methods that should be used, with caution, to the extent that their superstructural assumptions are reliable.

Key words: Bayesian inference; confidence intervals; decision theory; Empirical Bayes methods; frequentist statistics; hierarchical statistical models; hypothesis tests; statistical ecology.

A QUESTION OF EFFICIENCY

To quote I. J. Good (1983), University Distinguished Professor Emeritus of Statistics and Philosophy at Virginia Tech, "People who don't know they are Bayesians are called *non-Bayesians.*" At first, this may seem like a "slam" on non-Bayesians, but there is another interpretation: perhaps Bayesians and non-Bayesians are not as different as they think they are. How has science always been done? A study is begun with some existing beliefs on the state of nature, with differing amounts of confidence in the various existing theories (if any); after observing data, the scientist revises these beliefs and his/her confidence in the candidate theories. A talented scientist instinctively makes good decisions along the way, in much the same way a talented card player intuits probabilities accurately.

Bayesian statistical methods can be regarded as an attempt to formalize this age-old process by quantifying preexisting beliefs with a (somewhat) handy mathematical device called a prior probability distribution. Bayes' theorem provides the formula by which existing beliefs are updated as a reaction to data, resulting in revised beliefs quantified by the posterior probability distribution. If this formal mathematical process can be done, and done well, it results in a more *efficient* scientific process: human errors in judgment, especially in assessing the importance of the data relative to existing information, are reduced. However, if the Bayesian formalization of the scientific process is not done well, it can most definitely make matters worse. Moreover, in the science of ecology, with the current state of data-analytic technology, it often cannot be done at all by the scientists who have access to the best prior information. It is naive and a little bit arrogant to think that knowledge that required an ecologist years or decades of study to acquire can be passed intact to a statistician as one would pass the salt at the dinner table.

CAN'T WE ALL JUST GET ALONG?

Ludwig (1996) states that "the fundamental difference between . . . frequentist . . . statistical theory and . . . Bayesian statistics is in the interpretation of the term 'probability'." Though the relative-frequency interpretation of probability seems incompatible with a Bayesian's subjective, degree-of-belief interpretation, there are statistical methods that can be interpreted under both philosophies: the common ground is the device frequentists call a confidence interval. Consider the most common setting, the "normal prior + normal data" case, as in the hypothetical example used by Ellison (1996). His quantity of interest is the unknown fraction β of foliar area of red spruce affected by a predefined concentration of acid deposition. An individual's prior information on β is quantified using a normal distribution with some mean and standard deviation. The 10 observations (given β) are assumed to be taken from a normal distribution with mean β and some standard deviation σ. The posterior distribution for β in this setting is also normal, with expressions for pos-

[1] Manuscript received 29 December 1995; accepted 27 March 1996.

[2] For reprints of this group of papers on Bayesian inference, see footnote 1 on p. 1034.

TABLE 1. Ellison's red spruce acid deposition example, continued (Ellison 1996).

Ecologist	Prior mean	Prior SD	Posterior mean	Posterior SD	95% credible interval
A	0.4	0.05	0.318	0.021	(0.276, 0.360)
B	0.2	0.10	0.295	0.023	(0.249, 0.340)
C: non-informative	any value	infinite	0.300	0.024	(0.253, 0.347)
D	0.05	0.02	0.154	0.015	(0.124, 0.184)
E	0.8	0.02	0.592	0.015	(0.562, 0.622)

Note: Sample mean = 0.3, $s \approx \sigma = 0.075$, $n = 10$.

terior mean and variance given by Ellison's equations 8 and 9. Ellison then assumes $\sigma = s = 0.075$, though with only 9 degrees of freedom the sample standard deviation is fairly unreliable as a point estimate of σ; for the sake of discussion, we play along in this assumption.

Usually a normal posterior distribution would be reported as a credible interval (using 95% probability here):

posterior mean $\pm$ 1.96(posterior standard deviation).

Table 1 demonstrates the effects of several different choices of prior mean and standard deviation on the ultimate results for this example. Ellison's Bayesian Ecologists A and B have only slightly different prior beliefs, and are not overly confident in these beliefs, as reflected by relatively large prior standard deviations. Their posterior means are not very different from each other, or from the sample mean. The third entry of Table 1 shows a credible interval identical to a frequentist "σ known" 95% confidence interval for β. This interval corresponds to Ecologist C, who admits to the crime of having a completely open mind a priori (infinite prior standard deviation). The reported intervals for Ecologists A, B, and C are clearly not very different. One has to wonder whether the slight decrease in interval length for Ecologists A and B is worth the controversy that will be generated by the incorporation of their opinions into the data analysis.

In this setting and many others, a frequentist's confidence interval could be regarded by an open-minded, pragmatic Bayesian (assuming the intersection of these three groups of individuals is not empty) as a "journalistic" credible interval. It reports the posterior information flowing solely from the data, uncolored by the data analyst's own attitudes (except for distributional assumptions, which must be made under either philosophy). To spend a lot of time and energy arguing about whether we should say "confidence" or "probability" is to earn a reputation for being a useless academic: it's the same interval. Report the interval, and let each reader interpret it as he/she prefers, in the name of religious freedom. Care should be taken that the reported information is complete enough to allow any reader to formally incorporate prior information if he/she chooses to do so.

Of course, a Bayesian credible interval will not always be equivalent to a frequentist confidence interval. In the "journalistic" case of large prior variance, they are usually similar enough. However, a very small prior variance results in an editorial: the results are mostly opinion. This is the case for both Ecologists D and E in Table 1. These two individuals had strong and opposing views prior to collection of the data. By choosing very small prior variance, they have allowed themselves to ignore, for the most part, the sample information. Notice that for these dogmatic ecologists, even though the sample information strongly contradicts their prior beliefs, they have nonetheless mixed the sample and prior information and seem to have more confidence than ever in their posterior beliefs (their posterior standard deviations are smaller than their prior standard deviations).

HYPOTHESIS TESTS VS. INTERVAL ESTIMATION

Much of the discussion in Ellison (1996) is in fact arguing against hypothesis testing methodology, not frequentist methods in general. Prediction and estimation methods are available under both frequentist and Bayesian philosophies. Hypothesis tests are overused because they are so simple; the fact that an insignificant P value does not necessarily imply the hypothesis to be "true," and that a very small P value does not necessarily imply it to be meaningfully false, is well known to well-educated scientists. There is no question that interval estimates of ecologically meaningful quantities are more informative and honest than hypothesis tests, but one need not be a Bayesian to calculate them. In addition, Ellison's assertion that the coverage probability of a Student's t confidence interval is "substantially less than the expected coverage . . . for most parameterized, non-normal distributions" is grossly inaccurate. Procedures based on Student's t distribution have often been studied and are always found to be remarkably robust to nonnormality (Pearson [1931]; Box [1953]; Scheffé [1959:337]; Sokal and Rohlf [1981:414]). Ellison's lone supporting reference for this statement, Robinson (1975), considers only 50% confidence intervals, under three closely related, highly unrealistic (a.k.a. bizarre) distributional settings, and Robinson's intervals are not Student's t intervals.

POINT ESTIMATION VS. INTERVAL ESTIMATION

Ludwig (1996) calculates point estimates of near-extinction probabilities for three species under a simple

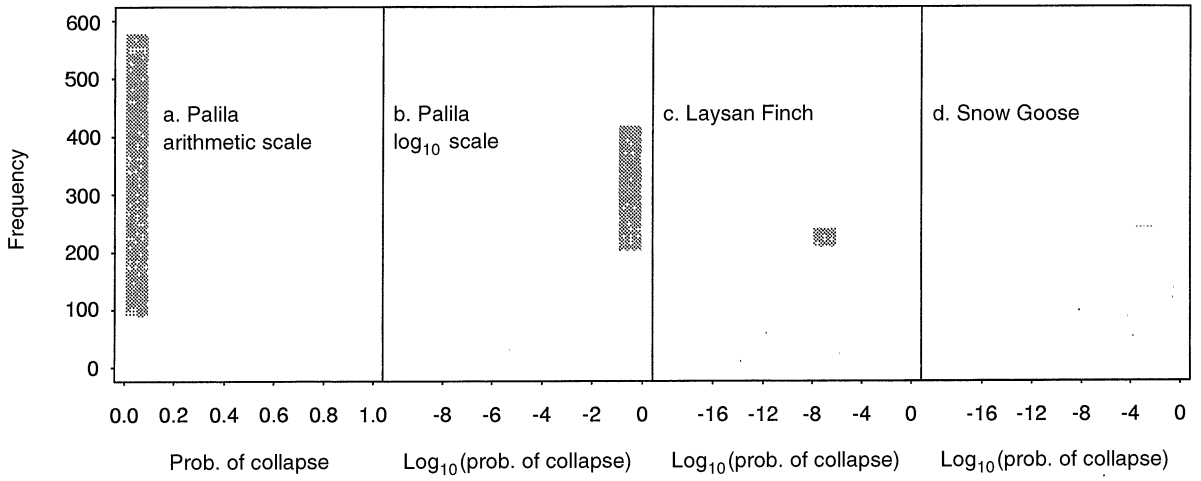

FIG. 1. Frequency distributions for 1000 realizations of the early-collapse probability under each of the posterior distributions derived by Ludwig (1996).

linear regression model using both the frequentist (maximum likelihood) method and a Bayesian noninformative prior approach. He shows that the estimated probabilities as computed by these two methods are quite different, in order to "point out the large differences between . . . [methods] . . . that ignore our uncertainty and those that take uncertainty into account," implying that he prefers the Bayesian estimates. It is ironic that, in the first draft of his paper, Ludwig reported point estimates with no accompanying measure of *their* uncertainty; the posterior distributions of his Fig. 1 provide this, but these were added in the second draft. The moral of his story *should* be that a point estimate (Bayesian or frequentist) with no accompanying measure of its accuracy is worse than no estimate at all. Instead he tries to argue that the Bayesian point estimates are more reliable than the frequentist ones. It is obvious from his posterior distributions and the supporting analysis provided below that neither point estimate is adequate as a full report of the information in the data.

Since Ludwig did not originally provide any measure of the accuracy of his estimates, in writing this comment I endeavored to do so by simulating his early-collapse probability $u(x, x_0, x_1, \beta_1, \beta_2, \sigma)$, as suggested by Ver Hoef (1996). Under the Bayesian approach, for any fixed choice of the x's, this quantity is a random variable whose probability distribution is determined by the joint posterior distribution of β_1, β_2, and σ. Ludwig rigorously and lucidly derives this distribution and displays it in terms of transformed quantities ρ, θ, and w. These quantities are a posteriori independent, with distributions up to multiplicative constants given by the $F(2, n-2)$, Uniform$(0,2\pi)$, and $\chi^2(n)$ distributions, respectively. A thousand realizations of the quantities ρ, θ, and w were generated using Splus (Becker et al. 1988), and the value of u computed for each using the numerical details provided by Ludwig.

The results of the simulation are displayed as histograms in my Fig. 1. These approximate the posterior density functions of the early-collapse probability, i.e., the derivatives of the functions shown in Ludwig's Fig. 1. The story told in each figure is the same, though many readers will see it more clearly using the histograms. Clearly there is a great deal of variability remaining in each posterior distribution, so it should come as no surprise that point estimates by different methods seem very different. As pointed out by Ludwig in his second draft, the probability of collapse for the Palila and Snow Goose species is indeterminate, possibly small but possibly quite large as well; the probability of collapse for Laysan Finch is clearly small, on the order of 0.001 or less. Both the Bayesian point estimates and the maximum likelihood estimates are well within the range of plausible values described by the posterior distributions.

It would be very interesting to compute and compare frequentist confidence intervals for the early collapse probabilities, for comparison with the Bayesian credible intervals. This is a much more difficult problem, however, and time constraints did not allow its solution for this comment.

DECISION THEORY

The elements of Bayesian decision theory are well described by Wolfson et al. (1996). It is a stimulating and logically beautiful theory, but many stimulating and logically beautiful theories do not work well in the real world. For example, the instructions state that "all possible actions must be enumerated" in the action space. Over what time frame? The difficulty with specifying an action space is that the time frame considered must be limited, since considering the great number of possible actions over a long time frame results in a mathematically intractable problem. This will have the effect of blocking out useful strategies and blinding the

analyst to long-term effects of decisions; we become chess players who cannot think beyond their next move. For example, in the first case study of Wolfson et al., one obvious strategy was omitted: taking the sample sequentially, a few observations at a time, was not considered. Confronted with this particular situation, a sequential sampling scheme seems a very obvious and viable compromise.

The specification of a loss function is also problematic, involving the assignment of "utility," a jargonese euphemism for dollar value, to outcomes that are qualitatively very different, and whose relative costs are very difficult to weigh. Here is a tragic true story of a former Mayor of Columbus, Ohio: when asked what his administration had done about the serious problem of rape around the Ohio State campus, he blurted out a frustrated reply to the effect that there were "many serious problems on campus, for example the littering problem." He lost the election. To all decision theorists: how many Baby Ruth wrappers equals one rape? Whatever the Mayor's optimal strategy, it would most likely *not* include the publicizing of his own loss function. That's a good way to alienate registered voters. This would often be the case for a politician or political entity, such as the EPA. The optimal decision-theoretic strategies for a political entity will place excessive weight on strategies improving the chances of its self-preservation, again ignoring long-term consequences.

As another specific example in this feature, to aid in specification of loss functions, Wolfson et al. specify some typical elements of a loss function related to health and environmental issues: monetary costs, loss or gain of goodwill, increase in life expectancy, medical problems, loss of quality of life, potential threat of litigation. Though they point out that this is not an exhaustive list, it is clearly focused on short-term human losses. It must be especially disappointing to ecologists that the very serious long-term loss due to low-level, steady degradation of the Earth's ecosystem did not make the list! This is blatantly obvious in the second example, where a local cemetery was not considered for remediation simply because no human beings lived there at present. In saying this, it is assumed that "remediation" means remedying (cleaning up) the contamination, as opposed to paying off any current residents who might sue the company. This latter activity would more accurately be termed "remuneration."

Hierarchical Models

Empirical Bayes methods, as discussed by Ver Hoef (1996), are considered by many to be the best idea to come along in statistics in the past 40 yr. They are examples of a class of exciting new statistical methods known as hierarchical models; other examples include meta-analysis (Gurevitch and Hedges 1993) and certain mixed-effects ANOVA models. As Ver Hoef points out, to call them "Empirical Bayes" models is somewhat misleading, since one need not adopt the subjective interpretation of probability to use them; they are so named because they use the Bayesian conjugate-prior mathematical framework as the top level of the hierarchical model structure.

Ver Hoef's first example is an excellent backdrop for discussion. Yearly estimates of harbor seals are not particularly accurate, individually. It is reasonable to assume some similarity in the true (unobservable) population size over consecutive years. In this case, Ver Hoef assumes a linear trend. Thus, the analytical model has a hierarchy of assumptions, one set for the unobservable true population sizes over years, another set for the sampled data within years. Unknowns at all levels of the hierarchy are estimated with the data. If the assumptions at all levels are fairly accurate, individual estimates (now called "predictions") of yearly population sizes are improved. Though the improvement in this example is modest, in some cases it can be quite dramatic.

There is much to be gained in the use of these sophisticated models. Is anything lost? Most of the following criticisms apply to Bayesian methods as well as frequentist hierarchical models: (1) the models are so complex as to be outside the reach of individuals who are not operating at the M.S. level or above in statistical knowledge; (2) they are "assumption heavy": the more structure adopted, the greater is the chance of serious misspecification; these dangers mount exponentially with added assumptions, not linearly; (3) assumptions are much more difficult to check in these models, especially assumptions made on unobservables. All the papers in this symposium, with the possible exception of Wolfson et al., are disappointing in their lack of reference to model diagnostic checks; (4) perhaps most seriously, as examples of "smoothing" methods, these models trade increased bias for reduced variance. If the top-level structure is too rigid, the data are oversmoothed, and as a result important discoveries may be missed. For example, the true population of harbor seals may be trending, but the trend is surely not rigidly linear. By assuming it so, bias is introduced into yearly population estimates. Referring to Ver Hoef's Fig. 1, the seal population in 1991 seems to be unusually high relative to the line. Had some unusual event occurred that allowed the seals to thrive that year? In his third example, the assumption of only two underlying mean densities, α and β, may be leading to a nondiscovery, in that the vegetation densities at low transects 1–35 seem consistently different as a group from those at high transects 70–100 (see Ver Hoef's Fig. 3).

These criticisms are stated in the spirit of devil's advocacy. In this author's opinion, the gains that are possible using carefully built hierarchical models far outweigh the concerns. To the extent that assumed structure is reliable, it should be used. Since the appropriate data analytic model will vary with the situ-

ation, the case-study approach taken by Ver Hoef (1996), Wolfson et al. (1996), Ludwig (1996), and Taylor et al. (1996) is necessary, and these authors are to be commended for doing the hard work required by this approach. These case studies are somewhat less than ideal, however, since in the real-data cases there is no way to know the truth, and in the single simulated-data case (the second example in Ver Hoef) the setting is somewhat artificial, and no investigation into effects of misspecified assumptions was made. The most convincing sort of evidence for any statistical methodology is the use of real data in a "verifiable" case study, one where the population truth is available. For example, we could subsample from a complete GIS layer, apply the various methods to the sample, and see how they perform in estimation of overall layer parameters, which are known. Several excellent examples of verifiable case studies are offered in Efron and Morris (1975), one of the pioneering Empirical Bayes works. In one example from that article, batting averages for the first 45 at-bats of the 1970 season were used to estimate final season batting averages for 18 major league baseball players. The Empirical Bayes approach assumed a normal distribution of (unobservable) innate abilities for the players, using the first 45 at-bats for all players to estimate the mean and variance of this ability distribution. The resulting empirical Bayes estimates of final batting averages were a dramatic improvement over the use of each player's 45-at-bats average, closer for 15 of the 18 players. The overall improvement in accuracy was equivalent to increasing the sample size by more than threefold.

Conclusion: Editorials Have Their Place, but . . .

It is not uncommon to find staunch, opposing attitudes to open research questions in ecology. Audiences are sometimes hostile; at the very least, they will aggressively challenge any arbitrary decisions made in the analysis and interpretation of data. Instances in which one could incorporate substantial amounts of prior information into an analysis, without rousing a shouting argument from an audience of ecologists, seem hard to imagine to this author. In the reporting of an ecological study, as in a newspaper, there may be a place for the author's opinion, but it is not on the front page. The first responsibility is to present the data without opinion (with the exception of clearly stated structural assumptions), in a form that will allow readers, formally or informally, to "update their own priors." To this end, a frequentist confidence interval analysis or a noninformative Bayesian credible interval analysis will be the most useful and appropriate primary analysis. These will provide similar interpretations, as was discovered in the section entitled: *Can't we all just get along,* and by Taylor et al. (1996).

Acknowledgments

Contribution Number 1080 from the Belle W. Baruch Institute for Marine Biology and Coastal Research, University of South Carolina, Columbia, South Carolina 29208.

Literature Cited

Becker, R. A., J. M. Chambers, and A. R. Wilks. 1988. The new S language. Wadsworth, Pacific Grove, California, USA.

Box, G. E. P. 1953. Non-normality and tests on variances. Biometrika **40**:318–335.

Efron, B., and C. Morris. 1975. Data analysis using Stein's estimator and its generalizations. Journal of the American Statistical Association **70**:311–319.

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

Good, I. J. 1983. The robustness of a hierarchical model for multinomials and contingency tables. Pages 191–211 *in* G. E. P. Box, T. Leonard, and C. Wu, editors. Scientific inference, data analysis, and robustness. Academic Press, New York, New York, USA.

Gurevitch, J., and L. V. Hedges. 1993. Meta-analysis: combining the results of independent experiments. Chapter 17 *in* S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments. Chapman and Hall, New York, New York, USA.

Ludwig, D. 1996. Uncertainty and the assessment of extinction probabilities. Ecological Applications **6**:1067–1076.

Pearson, E. S. 1931. The analysis of variance in cases of non-normal variation. Biometrika **23**:114–133.

Robinson, G. K. 1975. Some counterexamples to the theory of confidence intervals. Biometrika **62**:155–161.

Scheffé, H. 1959. The analysis of variance. John Wiley and Sons, New York, New York, USA.

Sokal, R. R., and F. J. Rohlf. 1981. Biometry. W. H. Freeman, San Francisco, California, USA.

Taylor, B. L., P. R. Wade, R. A. Stehn, and J. F. Cochrane. 1996. A Bayesian approach for classification criteria for Spectacled Eiders. Ecological Applications **6**:1077–1089.

Ver Hoef, J. M. 1996. Parametric empirical Bayes methods for ecological applications. Ecological Applications **6**: 1047–1055.

Wolfson, L. J., J. B. Kadane, and M. J. Small. 1996. Bayesian environmental policy decisions: two case studies. Ecological Applications **6**:1056–1066.