A Two-Hundred-and-Fifty-Year Argument

Bradley Efron*
Stanford University

Abstract

This is a brief commentary on the theme of the 30th Leeds Annual Statistical Research Workshop, "Next Generation Statistics in Biosciences." It raises the question of whether objective Bayes results enjoy the favorable properties of Bayesian analysis based on genuine prior distributions. This is pertinent to the Workshop theme since current activity in biomedical statistics has a strong objective Bayesian strain. Two methodologies, empirical Bayes and the bootstrap, are mentioned as connecting links between the frequentist and Bayesian worlds.

I am grateful to Professor Mardia for the chance to add a few brief comments on "Next Generation Statistics," the pregnant theme of the 30th Anniversary LASR Workshop. Let me begin facing in the wrong direction, back toward the beginning of statistical inference.

The year 2013 will mark the 250th anniversary of Bayes rule. The rule has been influential over the entire period, and controversial over most of it. Its popularity describes a rollercoaster graph: lofty in the time of Laplace, dipping low after Venn's 1866 *Logic of Chance*, back up by 1900, and then down again under Fisher and Neyman's massive influence. Most of the Twentieth Century was down time, especially in applications, but we now see a healthy revival of interest that is certain to influence the Next Generation.

Almost nobody questions Bayes rule in the presence of a genuine prior distribution, or for personal decision-making in the Savage–DeFinetti subjective prior framework. Controversy begins when the rule is applied to scientific problems in which prior information, often absent or untrustworthy, is replaced by "uninformative" priors of the type advocated by Jeffreys (1961). The healthy revival just mentioned is moving very much along Jeffreys' line. Markov chain Monte Carlo, MCMC, the key technology for computing posterior distributions in complicated circumstances, feasts on conventional uninformative priors.

Here is a simple (but true) story illustrating my point of concern. A physicist friend of mine found out, via sonogram, that she was going to have twin boys, and asked me what was the probability that they would be *identical* rather than *fraternal*. Her doctor had told her that only one-third of twin pairs are identical. Starting from this prior distribution, and the fact that a same-sex sonogram is twice as likely for identicals (since they are always same-sex whereas it's 50-50 for fraternals), Bayes rule gives the answer

$$\Pr\{\text{Identical}|\text{Same Sex}\} = 1/2. \tag{1}$$

A key assumption here, and the one that makes calculation (1) convincing, is that the doctor's one-third/two-thirds prior distribution was based on some huge registry of previous

*Max H. Stein Professor of Statistics and Biostatistics (Health Research and Policy)

twin births. Suppose, though, he had been speaking from limited personal experience, say just three previous twin deliveries, one identical and two fraternal. A conventional Jeffreys analysis might assign p, the population proportion of identical twins, a beta(1,2) distribution, with density

$$g(p) = 2 \cdot (1 - p) \qquad \text{for } 0 \le p \le 1, \tag{2}$$

obtained by beginning with an improper beta(0,0) "Haldane" hyperprior, updated to (2) on the basis of the doctor's experiences. (Jeffreys actually preferred starting from a beta($\frac{1}{2}, \frac{1}{2}$) hyperprior.) An entertaining application of Bayes rule starting from (2), which I only got wrong twice, yields Pr{Identical|Same Sex} = 1/2, the same as before.[1]

My point of concern is whether this "1/2" has the same logical force as that in (1). The beta(1,2) prior density (2) seems reasonable enough, but in fact there is nothing medical or twin-related about form (2). It yields other interesting conclusions — for instance, that the posterior probability of p exceeding 0.5 is 0.313 — which to a non-Jeffreysonian might seem suspiciously precise.

There is more at stake here than this little example suggests. Convenience priors of all types, "invariant," "uninformative," "reference," "conjugate," "objective," are featured in our journals, most often in complicated data analyses where it is difficult to trace their effect on conclusions. *Next Generation Statistics*, the subject here, is certain to include a substantial Bayesian component of the objective sort. In order to avoid another dip of the Bayesian rollercoaster we need a deeper understanding of the operational consequences of objective Bayesianism.

Besides its convenience and satisfying output (as seen in the twins example), Bayesian methodology offers an array of tempting theoretical advantages. A two-sample study, for instance, might include an over-ample supply of possible covariates, which have to be winnowed down before making the final comparison. Frequentist methods must take account of the complicated winnowing-down process, while the Bayesian's prior distribution is presumed to incorporate all such decisions. Similarly, interim looks at the data in a clinical trial require frequentist adjustment of the significance level, while the Bayesian need only consider the final comparison. "Selection bias," the data-based choice of interesting-looking cases, is no problem at all in the Bayesian world.

All of these properties follow logically and mathematically from Bayes theorem. The concern here is whether they are scientifically valid when a prior distribution, like (2), is not fully anchored in past experience. It seems important to me that objective Bayesian methods, which are certain to play an important Next Generation role, enjoy at least some frequentist support. (Bayarri and Berger (2004) pursue this point much more thoroughly.)

What I am hoping for from the Next Generation are ideas and methods that connect the two worlds, Bayesian and frequentist. *Empirical Bayes* is an attractive connecting technology, as discussed at length in Efron (2010). Hierarchical Bayes implementations of EB usually begin with uninformative priors at the top level, again raising the question of how far we can trust their formal Bayesian properties. (In this vein, Efron (2009) investigates freedom from selection bias in empirical Bayes estimation models.)

[1] The physicist's twins turned out to be fraternal, not identical. If the doctor were updating his prior from (2), it would now become beta(1,3), leading to a lower estimate than 1/2 if he were asked the same question about his next *twins* delivery.

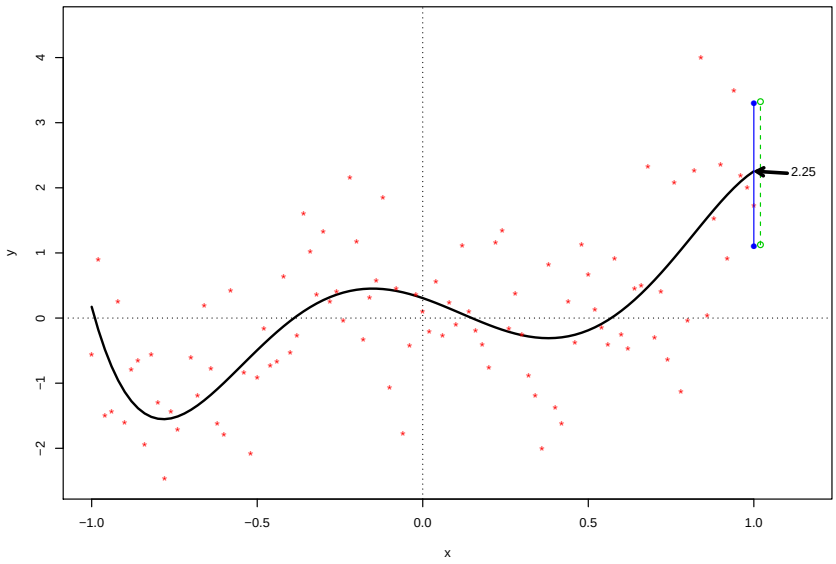

Figure 1: Points are 101 independent observations from regression model (3); solid line is fifth-degree polynomial chosen according to C_p criterion (6); $\hat{\theta} = 2.25$ estimates $f(1)$, (4). Dashed vertical line shows 90% central percentile interval for θ; reweighting the bootstrap replicates yields a central 90% Bayes credible interval, solid vertical line.

The bootstrap, usually presented in stark frequentist terms, also has its Bayesian connections. Figure 1 concerns an artificial but not completely unrealistic example. Independent normal responses from an unknown regression model $f(x)$ have been observed,

$$y_i \overset{\text{ind}}{\sim} \mathcal{N}\left(f(x_i), 1\right) \qquad \text{for } i = 1, 2, \ldots, 101 \tag{3}$$

with the goal of inferring

$$\theta = f(1), \tag{4}$$

the regression value at the far right end of the x scale.

A sequence of polynomial models

$$f^{(J)}(x) = \sum_{j=0}^{J} \beta_j x_i^j \qquad J = 0, 1, 2, \ldots, 8 \tag{5}$$

were fit by ordinary least squares, and compared in terms of the C_p (or AIC) criterion

$$C_p^{(J)} = \sum_{i=1}^{101} \left(y - \hat{f}^{(J)}(x_i)\right)^2 + 2 \cdot (J+1). \tag{6}$$

$J = 5$ gave a clear minimum, yielding the 5th-degree polynomial $\hat{f}(x)$ shown as the solid curve in Figure 1, and the estimate

$$\hat{\theta} = \hat{f}(1) = 2.25. \tag{7}$$

The assessment of confidence intervals after model selection is an exemplary Next Generation problem. One approach is via a parametric bootstrap. $B = 10,000$ (many more than necessary) bootstrap replications $\hat{\theta}^*$ were generated by each time taking

$$y_i^* \overset{\text{ind}}{\sim} \mathcal{N}\left(\hat{f}(x_i), 1\right) \qquad \text{for } i = 1, 2, \ldots, 101, \tag{8}$$

finding the minimum C_p polynomial $\hat{f}^*(x)$ based on the y^*'s, and setting $\hat{\theta}^* = \hat{f}^*(1)$. The vertical dashed line at the right shows the 90% central percentile interval $\theta \in [1.13, 3.32]$, with endpoints at the fifth and ninety-fifth percentiles of the 10,000 $\hat{\theta}^*$'s.

Newton and Raftery, in an under-appreciated 1994 paper, show how bootstrap replications, nonparametric in their case, can be reweighted to give Bayes posterior distributions (using a form of importance sampling). The solid vertical line is the central Bayes 90% credible interval $\theta \in [1.10, 3.30]$, based on a prior distribution following Smith and Spiegelhalter's 1980 prescription.

The reweighting values, which are mild and easy to compute in this case, give one a nice feeling for the relation between frequentist and Bayesian inferences. I expect, or at least hope, for an upsurge of progress along the Bayes/frequentist boundary in the near future.

References

Bayarri, M. J. and Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statist. Sci.* 19: 58–80.

Efron, B. (2009). Empirical bayes estimates for large-scale prediction problems. *J. Amer. Statist. Assoc.* 104: 1015–1028.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Institute of Mathematical Statistics Monographs *1*. Cambridge: Cambridge University Press.

Jeffreys, H. (1961). *Theory of probability*. Third edition. Clarendon Press, Oxford.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *J. Roy. Statist. Soc. Ser. B* 56: 3–48, with discussion and a reply by the authors.

Smith, A. F. M. and Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *J. Roy. Statist. Soc. Ser. B* 42: 213–220.