Bayes: Radical, Liberal, or Conservative?[1]

As a lifetime member of the International Chinese Statistical Association, I am pleased to introduce a volume of Bayesian articles. I remember that in graduate school, Xiao-Li Meng, now co-editor of this journal, told me they didn't teach Bayesian statistics in China because the idea of a prior distribution was contrary to Mao's quotation, "truth comes out of empirical/practical evidence." I have no idea how Thomas Bayes would feel about this, but Pierre-Simon Laplace, who is often regarded as the first applied Bayesian, was active in politics during and after the French Revolution.

In the twentieth-century Anglo-American statistical tradition, Bayesianism has certainly been seen as radical. As statisticians, we are generally trained to respect conservatism, which can sometimes be defined mathematically (for example, nominal 95% intervals that contain the true value more than 95% of the time) and sometimes with reference to tradition (for example, deferring to least-squares or maximum-likelihood estimates). Statisticians are typically worried about messing with data, which perhaps is one reason that the Current Index to Statistics lists 131 articles with "conservative" in the title or keywords and only 46 with the words "liberal" or "radical."

Like many political terms, the meaning of conservatism depends on its comparison point. Does the Democratic Party in the U.S. represent liberal promotion

[1] Based on joint research of Andrew Gelman and Aleks Jakulin, Department of Statistics and Department of Political Science, Columbia University, New York. We thank the National Science Foundation and National Institutes of Health for financial support.

of free expression or a conservative perpetuation of government bureaucracy? Do the Republicans promote a conservative defense of liberty and property or a radical revision of constitutional balance? And where do we place seemingly unclassifiable parties such as the Institutional Revolutionary Party in Mexico or the pro-Putin party in Russia?

Such questions are beyond the scope of this essay, but similar issues arise in statistics. Consider the choice of estimators or prior distributions for logistic regression. Table 1 gives an example of the results of giving specified doses of a toxin to 20 animals. Racine et al. (1986) fit a logistic regression to these data assuming independent binomial data with the logit probability of death being a linear function of dose. The maximum likelihood estimate for the slope is 7.8 with standard error of 4.9, and the corresponding Bayesian inference with flat prior distribution is similar (but with a slightly skewed posterior distribution; see Gelman et al. 2003, Section 3.7).

This noninformative analysis would usually be considered conservative—perhaps there would be some qualms about the uniform prior distribution (why defined on this particular scale), but with the maximum likelihood estimate standing as a convenient reference point and fallback. But now consider another option.

Instead of a uniform prior distribution on the logistic regression coefficients, let us try a Cauchy distribution centered at 0 with a scale of 2.5, assigned to the coefficient of the standardized predictor. This is a generic prior distribution that encodes the information that it is rare to see changes of more than 5 points on the logit scale (which is what it would take to shift a probability from 0.01 to 0.5, or from 0.5 to 0.99). Similar models have been found useful in the information retrieval literature (Genkin, Lewis, and Madigan, 2006). Combining the data in Table 1 with this prior distribution yields an estimated slope of 4.4 with standard error 1.9. This is much different from the classical estimate; the prior distribution has made a big difference.

Table 1. Bioassay data from Racine et al. (1986), used as an example for fitting logistic regression.

Dose (log g/ml)	Number of animals	Number of deaths
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

Is this new prior distribution conservative? When coming up with it (and using it as the default in our bayesglm function in R), we thought so. The argument was that true logistic regression coefficients are almost always quite a bit less than 5 (if predictors have been standardized), and so this Cauchy distribution actually contains less prior information than we really have. From this perspective, the uniform prior distribution is the most conservative, but sometimes too much so (in particular, for datasets that feature separation, coefficients have maximum likelihood estimates of infinity), and this new prior distribution is still somewhat conservative, thus defensible to statisticians.

But from another perspective—that of prediction—our prior distribution is not particularly conservative, and the flat prior is even less so! Let us explain. We took the software of Genkin, Lewis, and Madigan (2005), which fits logistic regressions with a variety of prior distributions and found that a Gaussian prior distribution with center 0 and scale 2.5 performed quite well as measured using predictive error from five-fold cross validation, generally beating the corresponding Cauchy model (as well as the maximum likelihood estimate) in predictive error, when evaluated on a large corpus of datasets. The conclusion may be that the Gaussian distribution is better than the Cauchy at modeling the truth, or at least that this particular Gaussian prior distribution is closer in spirit to what cross-validation is doing: hiding 20% of the data and trying to make predictions using the model built on the other 80%.

This result is consistent with the hypothesis that our Cauchy prior distribution has more dispersion than the actual population of coefficients that might be encountered. But is it conservative? From the computer scientist's standpoint of prediction, it is the Gaussian prior distribution that is conservative, in yielding the lowest expected predictive error for a new dataset (to the best of our knowledge).

Thinking about binary data more generally, the most conservative prediction of all is 0.5 (that is, guessing that both outcomes are equally likely). From this perspective, one starts with the prior distribution and then uses data to gain efficiency, which is the opposite of the statistician's approach of modeling the data first. Which of these approaches makes more sense depends on the structure of the data, and more generally, one can use hierarchical approaches that fit prior distributions from data. Our point here is that, when thinking predictively, weak prior distributions are not necessarily conservative at all, and as statisticians we should think carefully about the motivations underlying our principles.

Statistical arguments, like political arguments, sometimes rely on catchy slogans. When I was first learning statistics, it seemed to me that proponents of different statistical methods were talking past each other, with Bayesians promoting "efficiency" and "coherence" and non-Bayesians bringing up principles such as "exact inference" and "unbiasedness." We cannot, unfortunately, be both efficient and unbiased at the same time (unless we perform unbiased *prediction* instead of *estimation*, in which case we are abandoning the classical definition of unbiasedness that conditions on the parameter value).

Statistics, unlike (say) physics, is a new field, and its depths are close to the surface. Hard work on just about any problem in applied statistics takes us to foundational challenges, and this is particularly so of Bayesian statistics. Bayesians have sometimes been mocked for their fondness of philosophy, but as Bayes (or was it Laplace?) once said, "with great power comes great responsibility," and, indeed, the power of Bayesian inference—probabilistic predictions about everything—gives us a special duty to check the fit of our model to data and to our substantive knowledge.

In the great tradition of textbook writers everywhere, I know nothing at all about the example of Racine et al. (1986) given in Table 1, yet I feel reasonably confident that the doses in the experiment do not take the true probability of death from 0.003 to 0.999 (as would result from the odds ratio implied by the maximum likelihood estimate of 7.8). It seems much more conservative to me to suppose this extreme estimate to have come from sampling variation, as is in fact consistent with the model and data. Ultimately, it would be even better to have more realistic models that appropriately combine information from multiple experiments—a goal that is facilitated by technical advances such as those presented in the papers in this volume.

References

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). *Bayesian Data Analysis*, 2nd edition. CRC Press, London.

Genkin, A., Lewis, D. D. and Madigan, D. (2005). BBR: Bayesian logistic regression software. Center for Discrete Mathematics and Theoretical Computer Science, Rutgers University. http://www.stat.rutgers.edu/~madigan/bbr/

Genkin, A., Lewis, D. D. and Madigan, D. (2006). Large-scale Bayesian logistic regression for text categorization. *Technometrics* (To appear).

Racine, A., Grieve, A. P., Fluhler, H. and Smith, A. F. M. (1986). Bayesian methods in practice: experiences in the pharmaceutical industry (with discussion). *Appl. Statist.* **35**, 93-150.

— Andrew Gelman