Posterior distribution

Andrew Gelman

Volume 3, pp 1627–1628

in

Encyclopedia of Environmetrics
(ISBN 0471 899976)

Edited by

Abdel H. El-Shaarawi and Walter W. Piegorsch

Posterior distribution

The *posterior distribution* summarizes the current state of knowledge about all the uncertain quantities (including unobservable parameters and also missing, latent, and unobserved potential data) in a Bayesian analysis (*see* **Bayesian methods and modeling**). Analytically, the posterior density is the product of the prior density (*see* **Prior distribution**) and the likelihood. In a complicated analysis, the *joint posterior distribution* can be summarized by a set of L simulation draws of the vector of uncertain quantities $w_1, \ldots, w_J$, as illustrated in the following matrix:

$$\begin{array}{c|cccc} l & w_1 & w_2 & \cdots & w_J \\ \hline 1 & \cdot & \cdot & \cdots & \cdot \\ 2 & \cdot & \cdot & \cdots & \cdot \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ L & \cdot & \cdot & \cdots & \cdot \end{array}$$

The *marginal posterior distribution* for any unknown quantity w_l can be summarized by its column of L simulation draws. In many examples it is not necessary to construct the entire table ahead of time; rather, one creates the L vectors of posterior simulations for the parameters of the model and then uses these to construct posterior simulations for other unknown quantities of interest (e.g. predictions), as necessary.

Example

Suppose you are interested in R_i, the annual average living area concentration of radon in your home. **Radon** is a naturally occurring carcinogenic gas, and the US Government has recommended that you remediate your house if its radon level exceeds $4\,\mathrm{pCi\,l^{-1}}$. Any decision you make about remediating your home should use your posterior distribution for R_i.

A model has been fit to radon measurements in a national survey of houses, leading to an estimate of the distribution of radon levels in any set of houses given their location and other house characteristics. For example, suppose you live in Chester County, PA, and your house has a basement which is sometimes used as a living area. Then the model says that the radon levels of houses like yours have an approximate **lognormal distribution** with **geometric** mean M_i and geometric standard deviation S_i {that is, $\log R_i | M_i, S_i \sim \mathrm{N}[\log M_i, (\log S_i)^2]$}, with $S_i \approx 1.21$ and the posterior distribution for M_i itself being approximately lognormal with geometric mean 2.74 and geometric standard deviation 2.10. Averaging over the uncertainty in M_i yields a lognormal posterior distribution for R_i with geometric mean 2.74 and geometric standard deviation $\exp\{[(\log 1.21)^2 + (\log 2.10)^2]^{1/2}\} = 2.15$.

This is a *posterior* distribution with respect to the data from the national survey, and it can be summarized in various ways: for example, the posterior median of R_i is 2.74, the posterior expectation of R_i is $\exp[\log 2.74 + (\log 2.15)^2/2] = 3.67$, and the posterior probability that R_i exceeds $4\,\mathrm{pCi\,l^{-1}}$ is $\Phi[(\log 2.74 - \log 4)/\log 2.15] = 0.31$. Of these summaries, the posterior expectation is probably the most important since it estimates total radon exposure.

As is standard in Bayesian inference, the posterior distribution acts as a *prior* distribution for any analysis of further data. For example, suppose you measure the radon level in your home as $6.6\,\mathrm{pCi\,l^{-1}}$, and this sort of measurement is known to be lognormally distributed (actually, not an unreasonable assumption here at all) with multiplicative bias of 1.6 and geometric standard deviation of 1.8. Then the new posterior distribution for R_i is lognormal with geometric mean $\exp\{[\log 2.74/(\log 2.15)^2 + \log(6.6/1.6)/(\log 1.8)^2] / [1/(\log 2.15)^2 + 1/(\log 1.8)^2]\} = 3.54$ and geometric standard deviation $\exp\{(1/[1/(\log 2.15)^2 + 1/(\log 1.8)^2])^{1/2}\} = 1.59$. Thus the updated posterior expectation of R_i is $\exp[\log 3.54 + (\log 1.59)^2/2] = 3.94$, and so forth.

Literature

Recent theoretical and applied overviews of Bayesian statistics, including many examples and uses of posterior distributions, appear in [1]–[3]. The use of posterior distributions for decision-making about home radon exposure is discussed in [4].

Acknowledgments

Work supported in part by the US National Science Foundation Grant SBR-9708424 and Young Investigator Award DMS-9796129.

References

[1] Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory*, Wiley, New York.

[2] Carlin, B.P. & Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall, London.

[3] Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (1995). *Bayesian Data Analysis*, Chapman & Hall, London.

[4] Lin, C.Y., Gelman, A., Price, P.N. & Krantz, D.H. (1999). Analysis of local decisions using hierarchical modeling, applied to home radon measurement and remediation (with discussion), *Statistical Science* **14**, 305–337.

(*See also* **Bayesian computation**; **Hierarchical model**; **Markov chain Monte Carlo (MCMC)**)

ANDREW GELMAN