Journal of Official Statistics, Vol. 24, No. 4, 2008, pp. 495–506

The Morris Hansen Lecture 2007

Assessing the Value of Bayesian Methods for Inference about Finite Population Quantities

Joseph Sedransk[1]

Bayesian methodology is well developed and there are successful applications in many areas of substantive research. However, the use of such methodology in making inferences about finite population quantities is limited. I will describe several types of application where greater use of Bayesian methods is likely to be profitable and some where they are not. In addition, I will describe research whose successful completion should lead to improved analysis. The illustrations will come, primarily, from establishment surveys and a related area, providing public "report cards" for providers of medical care.

1. Introduction

Bayesian methodology is well developed and there are successful applications in many areas of substantive research. In many cases the advantage of using a Bayesian approach is that it avoids the necessity of using large-sample approximations while in others it permits a more appropriate but complicated probabilistic specification (e.g., in order restricted inference where one can include in the specification uncertainty about the existence and nature of the order restrictions; see Nandram, Sedransk, and Smith 1997). The formal structure of a Bayesian analysis makes it easier to incorporate available prior data, common in many repeated surveys. For some, the foundational arguments are compelling; see, e.g., DeGroot (1970) and Pratt, Raiffa, and Schlaifer (1964).

However, the use of such methods in survey sampling has been limited. The reasons for this include the presence of a well-developed methodology, i.e., design-based inference, the complexity of many survey designs, and the "observational" nature of the data. There are, though, many applications where a Bayesian approach should provide improved inferences. I shall describe a few of them.

At the U.S. Federal Reserve Board in 1972 there was a proposed redesign of a survey to provide early (weekly) estimates of several components of the money supply using data from a sample of member banks. An important feature was the vast amount of prior, *population*, data available (e.g., daily data for each bank in the population for the past eight years). These data could be used to improve estimation and to test alternative designs and estimators, and suggested that a Bayesian approach could be beneficial.

In Section 2 I discuss several types of problem where a Bayesian approach may be helpful: (a) analysis of data from establishment surveys, (b) small area inference, (c)

[1] Case Western Reserve University, Department of Statistics, 10900 Euclid Avenue, Cleveland, OH 44106-7054, U.S.A. Email: jxs123@cwru.edu

pooling data from disparate sources, and (d) general survey design. Section 3 has a discussion of situations where Bayesian methods are not likely to be useful, including a survey to estimate the prevalence of West Nile virus in Cuyahoga County, Ohio. Currently, providing "public" report cards for providers of medical care is a controversial topic. In Section 4 I use this as an example where Bayesian methods are useful, but the key issue is selection of the model. There is a brief summary in Section 5.

2. Applications

2.1. Inference for Establishment Surveys

At the U.S. Federal Reserve Board (FRB) in the 1970s it was desired to have timely and accurate estimates of various monetary aggregates for the population of about 5700 member banks of the Federal Reserve System using a sampling fraction of 10–15% with each bank reporting weekly. Since *all* member banks were required to report their data within three weeks, extensive prior data were available to improve estimation and to test samples and estimators using actual population data. Moreover, there is a plausible linear model relating each bank's current monetary aggregate to its value in a base period (where the value of the monetary aggregate is known for all units in the population). This situation is typical of many establishment surveys where there is a close relationship between the variable of interest, Y, and a covariate, X, and the values of X are known for all units in the population.

In the FRB example it is natural to use a stratified random sample with a large "certainty" stratum because of the extremely skewed distribution of Y for all monetary aggregates. Letting Y_{hi} denote the value of the monetary aggregate for bank i in stratum h for week t and X_{hi} the same aggregate for (i,h) for the base week, a reasonable population model is

$$Y_{hi} = \beta_h X_{hi} + \varepsilon_{hi}; \quad h = 1, \ldots, L, \quad i = 1, \ldots, N_h \tag{1}$$

where the ε_{hi} are independent with

$$\varepsilon_{hi} \sim N\left(0, \sigma_h^2 X_{hi}\right)$$

The finite population total, Y_T, can be written as

$$Y_T = \sum_{h=1}^{L} \sum_{i=1}^{N_h} Y_{hi} = \sum_{h=1}^{L} [n_h \bar{y}_h + (N_h - n_h)(\beta_h \tilde{x}_h + \tilde{\varepsilon}_h)] \tag{2}$$

where $\tilde{x}_h$ and $\tilde{\varepsilon}_h$ are the means for the $N_h - n_h$ nonsampled banks in stratum h and $\bar{y}_h$ is the sample mean in stratum h. Thus, the posterior expected value of Y_T is

$$E(Y_T|y_s) = \sum_{h=1}^{L} [n_h \bar{y}_h + (N_h - n_h)\bar{x}_h E(\beta_h|y_s)] \tag{3}$$

where y_s denotes the sample data.

For point estimation only $E(\beta_h|y_s)$ is needed; see (3). For inference we assumed that the σ_h^2 were known (i.e., to be replaced by point estimates – recall that this was in the early

1970s) and considered several prior distributions for the β_h. First, for a given week,

$$\begin{aligned} &\beta_1, \ldots, \beta_L | \nu \underset{\sim}{iid} N\left(\nu, \sigma_\beta^2\right) \\ &\pi(\nu) = \text{constant} \end{aligned} \tag{4}$$

where σ_β^2 is specified. Using (1) and (3) a straightforward extension of results in Scott and Smith (1969) leads to expressions for the posterior expected value and variance of Y_T. See Sedransk (1977) for details.

Second, for a given week,

$$\begin{aligned} &\beta_1, \ldots, \beta_L | \nu \underset{\sim}{iid} N\left(\nu, \sigma_\beta^2\right) \\ &\nu \sim N\left(\beta_0, \sigma_\nu^2\right) \end{aligned} \tag{5}$$

where $\left(\sigma_\beta^2, \beta_0, \sigma_\nu^2\right)$ are specified. The prior in (5) can be interpreted as follows (Sedransk 1977): "For any given current week t define J_t as the Julian date corresponding to the end of the week of interest. Then there is a conceptual time series $\{(J_t,\beta_{0t})\}$ where $1 \leq J_t \leq 366$. For given (J_t,β_{0t}), ν_t is an observation from $N\left(\beta_{0t}, \sigma_\nu^2\right)$ where σ_ν^2 is assumed known. Thus, $\{(J_t,\beta_{0t})\}$ is a further time series. Finally, given $\nu_t, (\beta_{1t}, \ldots, \beta_{Lt})$ is a random sample from $N\left(\nu_t, \sigma_\beta^2\right)$ with σ_β^2 specified."

We used the extensive historical data available (e.g., values of Y_T for each week for the previous ten years and microdata for the previous five years) to estimate $\left(\sigma_\beta^2, \beta_0, \sigma_\nu^2\right)$ in (5). Defining $\tilde{\beta}_h = \sum_{i=1}^{N_h} Y_{hi} / \sum_{i=1}^{N_h} X_{hi}$, the estimated regression coefficent for stratum h using data from *all* banks in the *population*, we used a cross-sectional random effects model for the $\tilde{\beta}_h$ to estimate β_0 and σ_β^2 for week t. Letting $\tilde{\beta}_{0t}$ denote the estimate of β_0 for week t we then used an ARIMA model on $\{\tilde{\beta}_{0t} : t = 1, \ldots, T\}$ to provide an estimate for σ_ν^2 and final estimates for $\{\beta_{0t} : t = 1, \ldots, T\}$.

We tested the alternative estimates by using the population data (Sedransk 1977, Sections 3.3 and 4). This investigation showed that the Bayesian methods were preferable to standard ones (e.g., the separate ratio estimator), but did not exhibit the full gains that could be realized. First, for (5), we estimated σ_ν^2 much too conservatively, so the estimates using (4) and (5) were similar. Second, we expect that the exchangeability assumption in (4) would continue to hold over time, but to be more beneficial as the structure of the finite population changed while the sample remained essentially the same (the latter due to reluctance to incur the expense of recruiting and training new sample banks).

A second example is estimation of total monthly sales of electricity at the Energy Information Administration (EIA) using a cutoff sample of companies. The current methodology is based on (1) where, for a *specified month*, Y_{hi} is the sales of electricity for company i in stratum h and X_{hi} is the *annual* sales of electricity for this company from a recent census. Inference at EIA uses model-based predictive inference (see Valliant et al. 2000), so the estimator, $\tilde{Y}_T$, of the finite population total, Y_T, in (2) is given by (3) with $E(\beta_h|y_s)$ replaced by $\tilde{\beta}_h$, the weighted least squares estimate. At EIA the strata are defined using regions based on climate, and there is ambiguity about how these should be chosen. An alternative is to use (4) or its frequentist counterpart. Recent

research (Sagatelov 2007) suggests that this is an appropriate model to use for estimation. However, inference can be improved by modifying (4) to accommodate the possibility that only subsets of the L regression coefficients are exchangeable. Three possibilities are using a Dirichlet process prior (e.g., Kleinman and Ibrahim 1998), "uncertain pooling" (Malec and Sedransk 1992; Evans and Sedransk 2001), or heavy tailed prior distributions (O'Hagan 1988). Research to compare these approaches and to extend the existing theory would be desirable. I know of no comparable extensions of (4) available in the frequentist mode of inference.

2.2. *Inference for Small Subpopulations*

There is, by now, a large literature on the theory and methodology for "small area" inference, and general agreement that model-based methods are needed for most applications. Three useful references are Rao (2003) and Jiang and Lahiri (2006), predominantly frequentist in approach, and Ghosh and Meeden (1997), a Bayesian treatment. As noted in Section 1, some may prefer the Bayesian approach because of foundational arguments. A more compelling reason for many is that use of large sample approximations, necessary in a frequentist analysis, can be avoided by using a Bayesian approach. Moreover, the theory necessary to obtain these large sample approximations may be formidable for many practical situations because of the complex models required (for example, see Jiang and Lahiri 2006, Section 4: Generalized linear mixed models.) To illustrated consider inference using data from the National Health Interview Survey (NHIS), for the proportion of individuals who made at least one visit to a doctor during the past year. Estimates were required for 72 age/race/sex categories for each county, i.e., about 3,000 x 72 subpopulations. To make these estimates Malec et al. (1997) used the following model.

Assume that each individual in the population is assigned to one of K mutually exclusive and exhaustive classes based on the individual's socioeconomic/ demographic status. Let Y_{ikj} denote a binary random variable for individual j in class k, cluster i where $i = 1, \ldots, L$, $k = 1, \ldots, B$, and $j = 1, \ldots, N_{ik}$. Here, $Y_{ikj} = 1$ if, and only if, individual (ikj) has made at least one doctor visit during the past year and the clusters are U.S. counties. Within cluster i and class k, and conditional on p_{ik}, the Y_{ikj} are assumed to be independent Bernoulli random variables with $\Pr(Y_{ikj} = 1 | p_{ik}) = p_{ik}$. A row vector of M covariates, $X_k = (X_{k1}, \ldots, X_{kM})$, is assumed to be the same for each individual j in class k, cluster i. Given X_k and a column vector of regression coefficients, $\beta_i = (\beta_{i1}, \ldots, \beta_{iM})'$, it is assumed that

$$logit(p_{ik}) = X_k \beta_i \tag{6}$$

Moreover, conditional on η and Γ, the β_i are independently distributed with

$$\beta_i \sim N(G_i \eta, \Gamma) \tag{7}$$

where each row of G_i is a subset of the cluster level covariates $(Z_{i1}, \ldots, Z_{ic})$, not necessarily related to X_k, η is a vector of regression coefficients, and Γ is an MxM positive

definite matrix. Reference prior distributions are assigned to η and Γ, i.e.,

$$p(\eta, \Gamma) \propto c \tag{8}$$

In this application (6) is a piecewise linear model, linear in age, with different forms for white males, white females, nonwhite males and nonwhite females. (An important benefit of this modelling exercise is that it demonstrated interesting relationships of the variable "probability of a doctor visit" to age, sex and race.) The elements of G_i are county level covariates such as county per capita income or education level. The assumptions in (6)–(8) provide a relatively simple specification that is concordant with the data (see Malec et al. 1997). The regression in (7) is important because it permits covariation between individuals in a cluster and provides the opportunity for increased precision.

The objective is to make inference about the finite population total, Y_T, the number of individuals in a specified small area and subpopulation who made at least one visit to a doctor during the past year; i.e.,

$$Y_T = \sum_{i \in I} \sum_{k \in K} \sum_{j=1}^{N_{ik}} Y_{ikj} \tag{9}$$

where I is the collection of clusters that define the small area, K is the collection of classes that define the subpopulation, and N_{ik} is the total number of individuals in cluster i, class k.

Now, let s_{ik} denote the set of sampled individuals in class k, cluster i that has size n_{ik}, and y_s the vector of sample observations. Then the posterior expected value of Y_T is

$$E(Y_T|y_s) = \sum_{i \in I} \sum_{k \in K} \sum_{j \in s_{ik}} y_{ikj} + \sum_{i \in I} \sum_{k \in K} \sum_{j \notin s_{ik}} E(p_{ik}|y_s) \tag{10}$$

and the second term in (10) can be written as

$$\sum_{i \in I} \sum_{k \in K} (N_{ik} - n_{ik}) E(p_{ik}|y_s)$$

where

$$p_{ik} = \frac{\exp[X_k \beta_i]}{1 + \exp[X_k \beta_i]}$$

Malec et al. (1997) show how to make these inferences using an exact method, and compare the results with approximations to the likelihood and posterior distribution and also to empirical Bayes and synthetic estimation. There is a large cross validation exercise to assess the quality of the fitted model.

One problem, perhaps not commonly recognized, is that the clusters used in the sample design may not be the most appropriate ones for the analysis. This should be a concern for a frequentist design-based analysis as well as model-based analyses – think of "post cluster analysis," analogous to post stratification. In the analysis described in Malec et al. (1997) the clusters in the model were chosen to be U.S. counties, which are more homogeneous than the primary sampling units. Other units (e.g., census tracts) were considered for analysis but rejected because the sample sizes were much too small.

2.3. Survey Design

One early line of development of Bayesian methodology was decision analysis. The objective is to make optimal decisions when data are observed and to choose the optimal experiment or sample survey. One approach requires specification of a loss function, $L(d, \theta)$, for decision d and parameter value θ, and may include the costs of various types of sampling. A good reference is Raiffa and Schlaifer (1961); Lindley (1971, Section 4) is a good summary.

With the current interest in reducing survey costs, using this framework may be a useful way to think about obtaining an improved survey design. This framework has been used to determine optimal strata sample sizes (Ericson 1965), optimal sample sizes from a two-phase stratified sample design (Smith and Sedransk 1982; Jinn, Sedransk, and Smith 1987), and the optimal choice of the number of nonrespondents to sample (Ericson 1967; Singh and Sedransk 1984).

While this list is incomplete, the papers cited provide a good introduction to the methodology. An alternative Bayesian approach to determination of optimal sample sizes is reviewed in Wang and Gelfand (2002). While the examples are about assessing performance under a given model and separating models, the methodology for use in sample surveys and related fields is apparent.

When designing a sample survey it is important to assess the precision that will be attained. At the point where the survey design has been selected but the actual sample has not been chosen even a stalwart Bayesian practitioner should use design-based principles. In the FRB example (Section 2.1) I selected a set of stratified random samples from a finite population constructed from a prior year's census data. Estimates from this set of samples provided the basis for the evaluation of the expected precision of estimation.

2.4. Pooling Data from Several Sources

With a greater demand for estimates for smaller subpopulations and reduced budgets, there is increased interest in using the data from several related sources. Raghunathan et al. (2007) consider the use of data from both the National Health Interview Survey (NHIS) and the Behavioral Risk Factor Surveillance System (BRFSS) to estimate, at the county level, four cancer risk factors and the use of two types of cancer screening. The situation is a typical one: The BRFSS is a much larger survey and almost all counties are included, but the nonresponse rates are much larger than in the NHIS and the BRFSS does not include subjects in households with no telephones. By contrast, the NHIS is much smaller but is a personal interview survey with lower nonresponse rates.

Estimates are required for each of four years for each U.S. county. This is a very large number of subpopulations, although not as large as the set in the NHIS example in Section 2.2. Raghunathan et al. (2007) start with (transformed) *direct*, *design-based* point estimates and estimates of variance, the latter adjusted by design effects to make the two surveys comparable. There are three estimates: (a) BRFSS, (b) NHIS for households with telephones, and (c) NHIS for households without telephones. The sampling model has means $(1 + \delta_{jt})\theta_{jt}$ for (a), θ_{jt} for (b) and ϕ_{jt} for (c) where j indexes counties and t years.

Letting $w_{jt} = (\theta_{jt}, \phi_{jt}, \delta_{jt})$

$$w_{jt} \sim N(U_{jt}\beta, \Sigma)$$

where U_{jt} is a vector of covariates. Finally, diffuse, proper priors are assumed for β and Σ.

The complexity of microdata models needed to provide state-of-the-art small area estimates, as exemplified by the problem described in Section 2.2, indicates the need for approximate methods. The approach in Raghunathan et al. (2007) is one possibility. Noting that in the version of NHIS used by Raghunathan et al. (2007) direct estimates are available only for about 25% of the U.S. counties, and that county sample sizes will be very small in many cases (especially for the nontelephone households), the quality of the "small area" estimates is uncertain and that of the estimates of variance is even more questionable. Here, a comparison between the use of direct estimates as the starting point, and an analysis based on modelling the microdata, as in Section 2.2, would be helpful in acquiring knowledge about the types of approximation that are acceptable.

A further question is whether $(\theta_{jt}, \phi_{jt}, (1 + \delta_{jt})\theta_{jt})$, adjusted for the county and year level covariates, might be regarded as exchangeable. If not, one of the approaches noted in Section 2.1 (i.e., uncertain pooling, Dirichlet process prior or heavy tailed prior distribution) may be an appropriate choice.

In a general setting with data from several, say k, sources one might consider the approaches noted above where parameters corresponding to each source are treated as "partially exchangeable," i.e., the observed data identify the grouping of the parameters corresponding to the k data sources.

3. Limitations

A sample survey of persons residing in Cuyahoga County, Ohio was conducted in December 2002 to estimate the prevalence of West Nile virus (Kippes and Sedransk 2004; Mandalakis et al. 2005). The survey also had two other objectives. The first was to obtain a large number of cases so that there could be a meaningful analysis of the relationship between presence/absence of West Nile virus and underlying risk factors. The second objective was to survey areas of the county where there was *no* suspected activity. Since previous surveys of West Nile virus (e.g., Mostashari et al. 2001) were limited to the epicenters of the disease, the Cuyahoga County investigators wanted to study the prevalence of the disease beyond the epicenter.

The survey required personal visits because each participant gave a blood sample. Given the conflicting objectives described above and the need to reduce travel time, a stratified, several-stage cluster design was used. The sampling fractions varied substantially, the stratification having been based on known information about the West Nile virus epidemic (i.e., numbers and locations of reported human cases and prevalence of infected mosquitos in traps placed across the county). To try to obtain a larger number of cases we oversampled in areas with large expected incidence of West Nile virus. With this complexity and limited time for providing an estimator (and estimator of variance), a Bayesian approach was infeasible, although the design-based estimator of variance also included several approximations.

The principal limitation with regard to the use of Bayesian methodology, shared with the frequentist model-based approach, is the cost of developing the models to be used for inference. An example is a specialized survey such as the West Nile virus survey just described. It is also costly to develop models for many population-based, complex, multivariate surveys, especially those requiring personal interviews. Use of design-based methods in the latter case often offers a reasonable alternative if inference is required only for moderate to large subpopulations. But this may not be true if inference is required for small subpopulations, if large sample approximations used in a design-based analysis are inappropriate or assumptions do not hold. Regarding the last two points a concern is about stratified designs with one or two primary sampling units selected per stratum where the between primary sampling unit variation within strata is a large component of the total variation and differs across strata. In cases such as this, large sample approximations commonly used may be inappropriate: One cannot provide adequate measures of variability and assumptions about the sampling distribution of the pivotal quantity (used to provide a confidence, or other, interval) may be questionable.

While there is no magic solution to the problem of developing appropriate models, there is, I think, a reasonable way to proceed. By fitting many models such as those described in Section 2.2 for the NHIS, experience will be gained and simpler models that are acceptable approximations will emerge. At the same time analytical approximations for the posterior distributions will also be identified, saving computer effort and time.

It is important to note that modelling for many establishment surveys is relatively easy. Furthermore, there may be large gains in precision by using hierarchical models and the more advanced methods that permit data-based assignment of the units to groups that are then assumed to be (internally) exchangeable (see e.g., Kleinman and Ibrahim 1998; Malec and Sedransk 1992; Evans and Sedransk 2001; O'Hagan 1988).

4. Provider Profiling

Provider profiling is the evaluation of the performance of hospitals, doctors, and other medical practitioners to enhance the quality of care. Jurisdictions such as Scotland, Ontario, California, Pennsylvania, Massachusetts, and New York have released "public report cards" comparing hospital or physician-specific outcomes. The objective of provider profiling may be to seek corrective measures when a provider's performance is thought to be unsatisfactory or to increase public awareness so individuals or institutions can make an informed decision about medical care. Given the practical importance of provider profiling it is essential to study the methodology used and to suggest alternatives.

Since 1989 the New York State Department of Health has evaluated the performance of the hospitals in New York which are licensed to perform coronary artery bypass graft surgery (CABG). The methodology they use is the one most commonly used for provider profiling. An elementary assessment of the performance of hospital i is to compare its Risk Adjusted Mortality Rate with the overall Statewide Mortality Rate. Then the New York State Department of Health (NYS DOH) obtains a $100(1 - \alpha)\%$ confidence interval for the expected value of the Risk Adjusted Mortality Rate for hospital i, and declares this hospital to be an outlier if the Statewide Mortality Rate falls outside this interval.

The basis for this analysis is the following model for patient j in hospital i: For the binary variable Y, $Y_{ij} = 1$ if, and only if, patient (ij) died in the hospital. It is assumed that

$$p_{ij} = Pr(Y_{ij} = 1); \quad i = 1, \ldots, m, \quad j = 1, \ldots, n_i \tag{11}$$

and

$$logit(p_{ij}) = X'_{ij}\beta \tag{12}$$

where, given the p_{ij}, the Y_{ij} are mutually independent, β is a vector of regression coefficients, and X_{ij} is a vector of preoperative risk factors for patient (ij). The Risk Adjusted Mortality Rate (RAMR) for hospital i is defined by

$$RAMR_i = (SMR_i)(SR)$$

where the Statewide Mortality Rate, SR, is $\sum_{i=1}^{m}\sum_{j=1}^{n_i} y_{ij} / \sum_{i=1}^{m} n_i$.

Also, $SMR_i = \sum_{j=1}^{n_i} y_{ij} / \tilde{E}_{s(i)}$ with $\tilde{E}_{s(i)} = \sum_{j=1}^{n_i} \tilde{p}_{ij}$, $\tilde{p}_{ij} = [1 + \exp(-X'_{ij}\hat{\beta})]^{-1}$ and $\hat{\beta}$ is the maximum likelihood estimator of β.

The NYS DOH procedure for identifying an outlier is awkward, requiring a confidence interval for the expected value of the RAMR. Moreover, large sample approximations are required (see Racz and Sedransk 2007). A more natural approach is to obtain the predictive distribution of the total mortality in hospital i, $\sum_{j=1}^{n_i} Y_{ij}$, assuming the statewide model in (11) and (12). One can then obtain a $100(1 - \alpha)\%$ predictive interval for $\sum_{j=1}^{n_i} Y_{ij}$ and regard hospital i as an outlier if, and only if, the observed mortality in hospital i lies outside this interval.

Racz and Sedransk (2007) use the model in (11) and (12), together with a uniform prior distribution for β, to obtain the posterior predictive interval for $\sum_{j=1}^{n_i} Y_{ij}$.

Normand et al. (1997) and Shahian et al. (2005) have advocated considering the hospital effects as random. Using (11), they assume

$$logit(p_{ij}) = X'_{ij}\beta + \delta_i \tag{13}$$

with

$$\delta_1, \ldots, \delta_m \underset{\sim}{iid} N(0, \gamma^2) \tag{14}$$

The Bayesian approach is to use (11), (13) and (14) together with prior distributions on β and γ^2 (e.g., Racz and Sedransk 2007 use uniform priors). For assessments one may use the $(\alpha/2)100th$ and $(1 - (\alpha/2))100th$ percentile values of the posterior distribution of the δ_i. Using (11), (13) and (14), an analogous frequentist method of assessment can easily be constructed (see Racz and Sedransk 2007 for details).

Racz and Sedransk used New York State's Cardiac Surgery Reporting System data for eight years to compare four methods: (a) New York State's Indirect Standardization, (b) Bayesian predictive inference patterned after (a), (c) Bayesian random effects, and (d) frequentist random effects. Note that the CSRS contains demographic variables, patients' clinical risk factors and complications, dates of admission, surgery, and discharge, and discharge status. There is also a separate analysis using constructed data, based on the CSRS data, where a specific hospital is targeted as an outlier.

We found that (a) and (b) selected exactly the same outliers (both "high" and "low") over the eight years, about 250 cases in total (approximately 32 hospitals per year). This is very surprising because the intervals are quite different (since they are intervals for different quantities). However, both compare the observed mortality in hospital i with mortality expected under the statewide model. Methods (c) and (d) identified similar outliers. The differences seem to be due to the use of a standard, symmetric confidence interval for (d) while the predictive interval for (c), based on the posterior predictive distribution, does not have to be symmetric – an advantage for the latter in this situation.

Comparing (a) with (d) (or (b) with (c)), many fewer outliers are detected using the random effects model, as one would expect. This is especially true for the low volume hospitals. This a major concern because Racz and Sedransk show, via published accounts, that several hospitals identified as outliers using indirect standardization (i.e., (a) or (b)) but *not* by (c) or (d) did, in fact, have serious deficiencies in their treatment of patients.

In Provider Profiling, as described here, there are advantages to a Bayesian approach. First, the Bayesian indirect standardization (Method (b)) avoids the large sample approximations inherent in the analogous frequentist (i.e., NYS DOH) method. The predictive approach is a more natural approach than using a confidence interval for an expected rate. Second, if one was to use a *standard* random effect approach it is unlikely that, the resulting symmetric confidence interval would be as satisfactory as a nonsymmetric one. However, the main issue here is the extent to which a random effect (or hierarchical) model should be used in the Provider Profiling setting and caution is advised.

5. Summary

For many establishment surveys models can be developed easily, and concordance of the model and data can be evaluated using census data. Using Bayesian methods in such cases can lead to improved inferences in several ways. First, intervals of plausible values of the finite population quantities do not have to rely on large sample approximations or on assumptions of symmetry such as $\hat{\theta} \pm z[\hat{SE}(\hat{\theta})]$. Second, there may be gains in precision, possibly substantial, from using hierarchical models. Extensions of such models to ones that do not require full exchangeability should make inferences more consonant with the data. The provider profiling problem provides a cautionary note that model validation is important, especially when there are serious consequences if one makes a poor decision.

The situation for population based surveys is more problematical, especially when inferences are needed for small to moderate-sized subpopulations. The limiting factor is usually the time needed to develop appropriate models, usually more complex than those for establishment surveys. I suggest that doing substantially more modelling will, inevitably, lead to acquiring greater skills in doing so. Ultimately, shortcuts will be discovered, e.g., methods for linking models for a set of different questions in a single sample survey.

6. References

DeGroot, M. (1970). Optimal Statistical Decisions. New York: McGraw-Hill.

Ericson, W. (1965). Optimum Stratified Sampling Using Prior Information. Journal of the American Statistical Association, 60, 750–771.

Ericson, W. (1967). Optimal Sample Design with Nonresponse. Journal of the American Statistical Association, 62, 63–78.

Evans, R. and Sedransk, J. (2001). Combining Data from Experiments that May Be Similar. Biometrika, 88, 643–656.

Ghosh, M. and Meeden, G. (1997). Bayesian Methods for Finite Population Sampling. London: Chapman and Hall.

Jiang, J. and Lahiri, P. (2006). Mixed Model Prediction and Small Area Estimation. Test, 15, 1–96.

Jinn, J., Sedransk, J., and Smith, P. (1987). Optimal Two-phase Stratified Sampling for Estimation of the Age Composition of a Fish Population. Biometrics, 43, 343–353.

Kippes, C. and Sedransk, J. (2004). Design, Implementation, and Analytical Methods for a Countywide West Nile Virus Seroprevalence Survey. Technical Report.

Kleinman, K. and Ibrahim, J. (1998). A Semiparametric Approach to the Random Effects Model. Biometrics, 54, 921–938.

Lindley, D. (1971). Bayesian Statistics, A Review. Philadelphia: SIAM.

Malec, D. and Sedransk, J. (1992). Bayesian Methodology for Combining the Results from Different Experiments when the Specifications for Pooling Are Uncertain. Biometrika, 79, 593–601.

Malec, D., Sedransk, J., Moriarity, C., and LeClere, F. (1997). Small Area Inference for Binary Variables in the National Health Interview Survey. Journal of the American Statistical Association, 92, 815–826.

Mandalakis, A., Kippes, C., Sedransk, J., Marfin, A. et al. (2005). West Nile Virus Epidemic, Northeast Ohio, 2002. Emerging Infectious Diseases, 11, 1774–1777.

Mostashari, F., Bunning, M., Kitsutani, P. et al. (2001). Epidemic West Nile Encephalitis, New York, 1999; Results of a Household-based Seroepidemiological Survey. Lancet, 358, 261–264.

Nandram, B., Sedransk, J., and Smith, S. (1997). Order-restricted Bayesian Estimation of the Age Composition of a Population of Atlantic Cod. Journal of the American Statistical Association, 92, 33–40.

Normand, S., Glickman, M., and Gatsonis, C. (1997). Statistical Methods for Profiling Providers of Medical Care: Issues and Applications. Journal of the American Statistical Association, 92, 803–814.

O'Hagan, A. (1988). Modeling with Heavy Tails. Bayesian Statistics 3, J. Bernardo, M. DeGroot, D. Lindley, and A.F.M. Smith (eds). Oxford: Clarendon Press, 345–355.

Pratt, J., Raiffa, H., and Schlaifer, R. (1964). The Foundations of Decision under Uncertainty: An Elementary Exposition. Journal of the American Statistical Association, 59, 353–375.

Racz, M. and Sedransk, J. (2007). Bayesian and Frequentist Methods for Provider Profiling Using Risk-adjusted Assessments of Medical Outcomes. Technical Report.

Raghunathan, T., Xie, D., Schenker, N., Parsons, V., Davis, W., Dodd, K., and Feuer, E. (2007). Combining Information from Two Surveys to Estimate County-level Prevalence Rates of Cancer Risk Factors and Screening. Journal of the American Statistical Association, 102, 474–486.

Raiffa, H. and Schlaifer, R. (1961). Applied Statistical Decision Theory. Boston: Division of Research, Graduate School of Business Administration, Harvard University.

Rao, J.N.K. (2003). Small Area Estimation. New York: Wiley.
Sagatelov, R. (2007). Survey Design and Analysis for Energy Statistics. Master's Paper. Cleveland: Case Western Reserve University.
Scott, A. and Smith, T.M.F. (1969). Estimation in Multistage Surveys. Journal of the American Statistical Association, 64, 830–840.
Sedransk, J. (1977). Sampling Problems in the Estimation of the Money Supply. Journal of the American Statistical Association, 72, 516–522.
Shahian, D., Torchiana, D., Shemin, R., Rawn, J., and Normand, S. (2005). Massachusetts Cardiac Surgery Report Card: Implications of Statistical Methodology. Annals of Thoracic Surgery, 80, 2106–2113.
Singh, B. and Sedransk, J. (1984). Bayesian Inference and Sample Design for Regression Analysis when There Is Nonresponse. Biometrika, 71, 161–170.
Smith, P. and Sedransk, J. (1982). Bayesian Optimization of the Estimation of the Age Composition of a Fish Population. Journal of the American Statistical Association, 77, 707–713.
Valliant, R., Dorfman, A., and Royall, R. (2000). Finite Population Sampling and Inference. A Prediction Approach. New York: Wiley.
Wang, F. and Gelfand, A. (2002). A Simulation-based Approach to Bayesian Sample Size Determination for Performance under a Given Model and for Separating Models. Statistical Science, 17, 193–208.

Received April 2008