

Impact of Bayesian Methods in Survey Sampling: An Appraisal

J. N. K. Rao

Carleton University, Ottawa, Canada

Invited Talk: Workshop on Bayesian Methods that Frequentists Should Know, April 30- May 1, 2008

Design-based Approach:

- Probability-sampling design and randomization inferences
- Early landmark contributions (1920-50):

Neyman (1934):

1. Stratified random sampling preferable to balanced samples.
2. Ideas of efficiency and optimal sample allocation.
3. Normal theory confidence intervals in large samples.
 - Frequency of errors in the confidence statements based on all possible stratified random samples that could be drawn does not exceed the limit prescribed in advance “**whatever the unknown properties of the population**”.
 - Any method of sampling that satisfies the above frequentist statement is called “**representative**”.

Mahalanobis(1944)

- **Mahalanobis:** As early as 1937, Mahalanobis in India used multi-stage sampling designs: villages, grids/villages, plots/grids for crop surveys.
- **Mahalanobis (1944):** Cost and variance functions for design of surveys, deep theoretical results on efficient design. He established the National Sample Survey of India.

US Census Bureau (Golden Era 1940-60):

- **Leaders:** Hansen, Hurwitz, Madow, Waksberg, Bershad, Tepping and others. Amstatnews, March 2008 says “Morris Hansen is the most influential statistician in the evolution of survey methodology in the 20th century.”
- **Hansen et al. (1943):** Stratified multi-stage designs with one PSU per stratum by PPS sampling, sub-sampling to ensure self-weighting with equal workloads, reduce variance without actually stratifying by PSU size.
- **Current Population Survey (CPS):** rotation sampling to reduce respondent burden, composite estimation.

Golden Era Contd..

- **Woodruff (1952)** intervals for quantiles using only estimated distribution function and associated SE. Valid for complex designs and excellent performance.
- Two volume **Wiley (1953) book** by Hansen, Hurwitz and Madow.
- **Other Leaders:** Cochran (Wiley 1953 book), Deming, Hartley, Jessen, Kish, Keyfitz, Lahiri, Dalenius and Sukhatme.

Features of design-based approach:

- Focus on sampling errors, large samples (large domains), efficiency and cost considerations
- Strategies (design and estimation) that appeared reasonable were entertained (accounting for costs) and relative properties carefully studied by analytical and/or empirical methods, mainly through comparison of MSE and anticipated MSE under plausible population models.
- Design unbiased estimators not insisted upon because it “often results in much larger MSE than necessary”. Instead **design consistency** is deemed necessary for large samples.

Measurement or response errors

- **Mahalanobis (1946)**: Interpenetrating subsamples assigned to different interviewers: Tests for interviewer differences and estimator of total variance.
- **Hansen et al. (1951)**: Basic theory for totals under additive measurement error models. Decomposition of total variance into sampling variance, simple response variance (**SRV**) and correlated response variance (**CRV**).

Response errors contd..

- CRV is of order $1/(\# \text{ interviewers})$. 1950 US Census Interviewer Variance Study showed large CRV for small areas. As a result, self-enumeration by mail adopted in the 1960 Census to greatly reduce CRV: **Success of theory influencing the practice.**

Remarks

- Much of the basic theory was developed by official statisticians or those closely associated with official statistics. **Theory was driven by the need to solve real problems.**
- Unfortunately academic statisticians paid little attention to survey sampling those days with some exceptions: Iowa State University under leadership of Cochran, Jessen, and Hartley. Horvitz and Thompson were Ph. D. students at ISU resulting in the well-known HT estimator (1952).

Remarks

- Smith (1994) named the design-based approach as “procedural inference” and defended its use in the public domain because procedures are laid down in advance, using sampler’s skills in choosing them, and inferences are free of political interference which in turn promotes trust in the statistics generated by the agency.

Model-dependent (or prediction) approach

- Brewer (1963) and Royall (1970): BLUP estimators
- Inferences conditional on the sample assuming a stochastic model for the item given predictor variables . Model assumed to hold for the sample. Focus on estimator, variance estimator and normal theory confidence intervals. Hence, only item means and covariance structure need to be specified.

Prediction Approach Contd..

- Royall and Pfeffermann (1982) studied **Bayesian inference** on the population mean assuming **normality** of item values and **flat (or diffuse)** prior on the model parameters: focus on posterior mean and posterior variance. Results are similar to Royall (1970) without the prior.
- Scott and Smith (1969): two-stage sampling.
- Malec and Sedransk (1985): three-stage sampling.

Hansen et al. 1983

- Model-dependent strategies can perform poorly in large samples under model misspecifications not easily detectable.
- True model: $E_m(y_i) = 0.4 + 0.25x_i$
- Model misspecification: $E_m(y_i) = \beta x_i$
- Stratified SRS with disproportionate allocation (business population), $n = 200$, nominal coverage of CI is 95%. **Model assisted** (strata weighted est.): 94.4% while **model dependent** (un-weighted ratio est.): 70%.
- Design inconsistency of the BLUP (un-weighted ratio estimator) is the root cause for the poor performance. Model-assisted estimator also performed well conditionally (Rao, 1999).
- Little (1983) proposed choosing models for which BLUP is design-consistent.

Model-assisted approach

- Särndal et al. (1992)
- Working linear regression model, generalized linear regression (GREG) estimator, design consistency. Results in a single weight for each unit provided the same model is used for all the items.
- Not much attention is given to choice of working model leading to poor performance of CI coverage for highly skewed x even for moderately large samples (Dorfman, 1994). Remedy is to pay some attention to working model (Rao et al., 2003): true model quadratic, working model linear.

Calibration estimators

- Satisfy user-specified population constraints (known totals of auxiliary variables).
- Resulting estimator may not be model-assisted but calibration leads to single weight for each unit. Hence calibration is very popular in federal agencies.
- Chi-squared distance measure for calibration leads to GREG on the user-specified covariates.

Calibration contd...

- “Optimal” linear regression calibration estimators with good conditional repeated sampling properties (Rao 1982, 84) and Casady and Valliant (1993).
- Calibration weight w_i equals design weight d_i multiplied by adjustment factor g_i derived from calibration constraints:

$$w_i = d_i g_i, d_i = 1 / \pi_i, \pi_i = \Pr(i \in s)$$

Särndal (2007):

- Calibration has established itself as an important methodological instrument in large-scale production of statistics. Several national statistical agencies have developed software designed to compute weights (GES, CALMAR).

Analysis of Survey Data

- **Design features:** clustering, unequal probabilities and stratification (stratified multi-stage designs used in large-scale socio-economic and health surveys). Ignoring design features can lead to erroneous inferences on model parameters.
- Why not include among the covariates all the design variables that define the selection process at the various levels and ignore design?
Answer: (1) Not all design variables may be known or accessible to the analyst. (2) Too many design variables. (3) Resulting model may no longer be of scientific interest to the analyst.
- Do not change the analyst's model. Instead use weights $\{w_i, i \in s\}$ to account for design features.

Population model leads to Population estimating equation (EE):

- $U(\theta) = \sum u(y_i, \theta) = 0 \Rightarrow \theta_c$ (“census” parameter)

- Example: Linear regression with mean specification

$$E_m(y_i) = z_i' \theta$$

$$u(y_i, \theta) = z_i'(y_i - z_i \theta), E_m\{u(y_i, \theta)\} = 0$$

- **Sample weighted EE:** $\hat{U}(\theta) = \sum_{i \in s} w_i u(y_i, \theta) = 0 \Rightarrow \hat{\theta}$

Bootstrap standard errors

- Use Rao-Wu (1988) bootstrap weights
 $\{w_i^{(b)}, i \in s\}, b = 1, \dots, B \Rightarrow \hat{U}^{(b)} = \sum_{i \in s} w_i^{(b)} u(y_i, \theta) = 0 \Rightarrow \hat{\theta}^{(b)}$
- Software packages that account for weights can be used to get $\hat{\theta}, \hat{\theta}^{(b)}, b = 1, \dots, B$
- Bootstrap variance estimator:

$$v_{BOOT}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^B [\hat{\theta}^{(b)} - \hat{\theta}][\hat{\theta}^{(b)} - \hat{\theta}]^t$$

- Data File: $\{y_i, z_i, w_i, w_i^{(b)}, b = 1, \dots, B, i \in s\}$
- **Note:** Analysis is done using only the data file and standard software allowing weights.

Bayesian Approach

- Inferences are conditional on the sample data
- Main hurdles: specification of likelihood and prior

Nonparametric likelihood:

- Parameter vector $\tilde{y} = (\tilde{y}_1, \dots, \tilde{y}_N)'$
- Sample data: $\{(i, y_i), i \in s\}$ minimal sufficient
- Likelihood function $L(\tilde{y})$ is flat: All possible unobserved \tilde{y}_i have the same $L(\tilde{y}) = p(s)$ if sample data consistent with \tilde{y} , = 0 otherwise.

Resolution

- **Likelihood route** (Hartley and Rao, 1968; Royall, 1968): Use reduced data set. For example, for SRS, suppress labels and use $(y_i, i \in s)$.
Resulting likelihood is informative and inferences depend on sampling design.
- **C. R. Rao (1970)**: “In situations where the full likelihood does not satisfy our purpose, we may have to depend on a statistic which for every observed value supplies information (however poor it may be) on parameters of interest. Unfortunately, no unique choice of may be possible.”

Scale-load approach (Hartley and Rao, 1968): SRS

- y_i belongs to a finite set of possible values $\{h_1, \dots, h_T\}$
- $N_t =$ scale load of $h_t \Rightarrow \bar{Y} = \sum p_t h_t$ where $p_t = N_t / N$
- $n_t =$ sample scale load (≥ 0)
- **Likelihood function:** $L(N)$ multivariate hypergeometric with support on scale loads $n_t > 0$. If sampling fraction is negligible, then $L(p)$ is approximated by a multinomial likelihood, now called **empirical likelihood** (Owen, 1988):

$$\log L(p) = \sum n_t \log p_t$$

Bayesian inference

- Combine $L(N)$ with a Dirichlet-multinomial prior (Hoadley, 1969) to get a translated Dirichlet-multinomial posterior distribution (Hartley and Rao, 1968).
- In the case of a diffuse prior we get posterior mean of \bar{Y} = sample mean \bar{y} and Posterior variance of \bar{Y} is

$$(1-f) \frac{s_y^2}{n} \frac{n-1}{n+1}$$

- Hence, normal approximation (NA) posterior credible intervals are almost identical to NA confidence intervals on \bar{Y}
- If the sampling fraction is small, we use a conjugate Dirichlet prior on p leading to Dirichlet posterior on the components p_t with $n_t > 0$. Easy to simulate from this posterior and hence posterior summaries for any parameter (function of p) can be computed.

Extensions (Rao and Ghangurde 1972):

- Stratified SRS, two-phase sampling, Hansen-Hurvitz non-response method: optimal sample allocation minimizing expected posterior variance based on informative priors or pilot samples and diffuse priors or multiple informative priors.
- Unfortunately, my work on Bayesian methods has been largely ignored by the Bayesians. However, see Aitkin (2008).

Informative priors

- **Ericson (1969)**: highly cited paper
- Exchangeable prior on \tilde{y} and flat likelihood leads to informative posterior that does not depend on the sampling design. Posterior mean and posterior variance are **identical** to those of Hartley and Rao (1968) given above. Binder (1982) used Dirichlet Process (DP) prior.

Meden & Vardeman (1991)

- For SRS, Polya posterior (PP) over the unobserved (scale-load set up) assuming “unseen are like the seen”. Posterior does not arise from a single prior: pseudo-posterior. PP is a flexible tool with reasonable frequentist properties. It is similar to **Bayesian bootstrap** (Lo, 1988). Samples are generated from PP to compute posterior summaries. For the mean, results are identical to those of Hartley and Rao (1968). For the median, PP performed well in terms of coverage probability (CP) but Woodruff method also does well.

Polya Posterior

- PP method can handle complex parameters, such as ratio of medians of a bivariate population, through simulation of the posterior.
- **Meden (1995)**: Auxiliary scalar information (x_1, \dots, x_N) utilized by assuming exchangeability of the ratios $r_i = y_i / x_i$. Under this exchangeability assumption PP leads to reasonable frequentist properties.
- **Meden (1999)**: Two-stage sampling (balanced case). Results for the mean are very close to standard results (Cochran, 1977).

Calibration

- It is difficult to generate samples from PP that can satisfy calibration constraints. A possible solution is to use the **likelihood route** of Hartley and Rao (1968) and get profile likelihood (**PL**) by maximizing the empirical likelihood (**EL**) subject to calibration constraints and estimating equation for the parameter of interest. Combine this likelihood with a diffuse prior to get pseudo-posterior summaries. Robust Bayesian inferences can also be handled using this approach (Greco et al., 2008).

Small Area Estimation

- Traditional domain-specific methods are not adequate for small domains (areas) because of small sample size or even zero sample size.
- Necessary to **borrow strength** from related areas through **linking models**. Hansen et al. (1983): “ If the assumed model accurately represents the state of nature, useful inferences can be based on quite small samples at least for certain models”.
- Explicit linking models through auxiliary information (census and administrative data) and linear, **generalized linear mixed models** with random effects leading to integration with mainstream statistical theory: Empirical Bayes (**EB**) and Hierarchical Bayes (**HB**) methods. See my Wiley book (Rao, 2003).

SAE contd...

- Major application of model-based methods:
SAIPE
- **HB** approach can handle complex models and can lead to **exact** inferences. Availability of powerful **MCMC** methods and **WinBUGS** software using default priors.
- Extensive model diagnostic tools available but not necessarily good for detecting model deviations:
Posterior P-value

Choice of Prior

- Basic area level (Fay-Herriot) model:

Sampling model: $\hat{Y}_i = Y_i + e_i, e_i \sim_{ind} N(0, D_i)$, known D_i

Linking model: $Y_i = z_i' \beta + v_i, v_i \sim_{iid} N(0, A), i = 1, \dots, m$

- Prior: $f(\beta, A) \propto f(A)$
- **Posterior summaries:** Posterior mean and posterior variance of the total Y_i , credible intervals on Y_i , using direct estimates \hat{Y}_i and auxiliary data z_i

Choice of Prior

- Flat prior $f(A) \propto 1$ is commonly used. But it lacks frequentist validity unless all the sampling variances are equal: $D_i = D$.
- Datta et al. (2005) derived **matching prior**

$$f(A) \propto (A + D_i)^2 \sum_{l=1}^m (A + D_l)^{-2}$$

- It tracks the **MSPE** in the sense that the posterior variance is nearly unbiased for **MSPE**. Also coverage probability (CP) of the normal theory interval based on posterior mean and posterior variance tracks the corresponding CP based on EB and **estimated MSPE**.

Choice of Prior

- Ganesh and Lahiri (2007) obtained a **single prior** such that the weighted posterior variance over areas tracks the corresponding weighted MSPE for given weights.
- In the SAIPE application, REML estimate of A sometimes turned out to be zero in the State model for poverty which means EB estimate gives zero weight to direct estimate even for a big state like California. Bell (1999) used the flat prior on A which gives positive posterior mean of A and hence positive weight to the direct estimate. However, in this application $\max D_i / \min D_i$ is as large as 20.

Small Area-contd...

- Morris (2006) proposed to multiply the residual likelihood of A by the factor A and maximize this adjusted likelihood. Resulting estimate of A is always positive and gets around the difficulty with REML.
- My 2003 book has the longest chapter on HB methodology for small area estimation. Surely, I cannot be an **anti-Bayesian!**

Concluding Remarks

- For domains with large samples, traditional design-based approach will remain as the preferred method in official statistics. Impact of Bayesian methods here is likely to be small.
- For small area estimation, **HB** offers a lot of promise because of its ability to handle complex models and provide exact inferences. However, the priors should be chosen to provide frequentist validity and this is not likely to be easy in practice.