ISSN : 0008-0683

CALCUTTA STATISTICAL ASSOCIATION

BULLETIN

CONTENTS

Applications of Statistics

Vol. 60, September and December 2008, Nos. 239 - 240

CALCUTTA STATISTICAL ASSOCIATION BULLETIN

Founding Editor : *Hari Kinkar Nandi*

Editors : **K. Das, U. Bandyopadhyay**
Coordinating Editor : **A. K. Chattopadhyay**
Associate Editors : **T. Banerjee, A. Dewanji,**
Aditya Chatterjee, P. Das

Each volume of this quarterly publication will comprise four numbers. Normally one double number (March - June) is published in June and another (September - December) in December every year.

Subscription rates for *current* volume are as follows :

Inland	Rs.400.00	(inclusive of postage)
Foreign	U.S.$ 80.00	(inclusive of postage)

Back Volumes :

Inland	Rs.500.00	(inclusive of postage)
Foreign	U.S.$ 100.00	(inclusive of postage)

Journal would be always despatched by registered Post (Inland)/Registered Airmail (Foreign).

The BULLETIN is supplied free of charge to the members of the Association.

The Council of Calcutta Statistical Association is not responsible for the views expressed in the BULLETIN.

All communications (other than those relating to publication of articles) are to be addressed to :

The Secretary, Calcutta Statistical Association,
C/o Department of Statistics, Calcutta University
35, Ballygunge Circular Road, Kolkata - 700 019, India

Communications relating to publication of articles are to be addressed to :

The Editor, **Calcutta Statistical Association Bulletin**
C/o Department of Statistics, Calcutta University
35, Ballygunge Circular Road, Kolkata - 700 019, India

website : www.calcuttastatisticalassociation.org

Calcutta Statistical Association Bulletin
Vol. 60, September & December 2008, Nos. 239-240

WEIGHTING AND PREDICTION IN SAMPLE SURVEYS

RODERICK J. LITTLE
University of Michigan, USA.

ABSTRACT : A fundamental technique in survey sampling is to weight included units by the inverse of their probability of inclusion, which may be known (as in the case sampling weights) or estimated (as in the case of non-response weights or post-stratification). The technique is closely associated with the design-based approach to survey inference, with the idea that units in the sample are representing a certain number of units in the population. I discuss weighting from a modelling perspective. Some common misconceptions of weighting will be addressed, including the idea that modelers can ignore the sampling weights, or that weighting necessarily reduces bias at the expense of increased variance, or that units entering the calculation of nonresponse weights should be weighted by their sampling weights. A robust model-based perspective suggests that selection weights cannot be ignored, but there may be better ways of incorporating them in the inference than via the standard Horvitz-Thompson estimator and its variants.

Keywords and phrases : Bayesian methods, Design-based inference, Sampling weights, Regression, Robustness, Survey sampling.

1. Introduction

It is an honor to write an article in celebration of the diamond jubilee of the Calcutta Statistical Association Bulletin, a venerable statistical institution, and to acknowledge the profound contribution of Indian statisticians to progress in our field. Historically, this is clear when we consider the influence of major Indian statisticians like Basu, Gnanadesikan, Mahalanobis, and more recently C. R. Rao, not to mention the distinguished Rao's with other initials, and many others. Personally, my career has been enhanced by numerous friendships and encounters with Indian statisticians; my boss in my first real job at World Fertility Survey was the demographer VC Chidambaram (Chid to all who knew him) who was a sympathetic colleague and strong leader; another fine colleague at World Fertility Survey was Vijay Verma, an outstanding student of Leslie Kish who played a leading role in sampling activities in that large study. More recently, I have since collaborated extensively with my colleague Trivellore Raghunathan at Michigan, on topics of sampling inference and missing data. Indeed Biostatistics at Michigan has a strong Indian representation in terms of faculty and students.

I write about on the role of weights in the analysis of survey samples. Probability sampling is one of the key contributions of statistics, and this is an area where Indian statisticians have made seminal contributions (e.g. Mahalanobis 1943; Godambe 1955; Basu 1971; Rao 1997, 2003). Many of the key aspects of probability sampling, including stratification and multistage sampling, were first implemented on a large scale in India. It has interested me since my time working at the World Fertility Survey, where the virtues of probability sampling were widely touted by Sir Maurice Kendall and Leslie Kish, and the question of making analytic inferences that incorporated the survey design was of great interest. As a statistician drawn to the Bayesian paradigm for survey inference, sample surveys are a challenge since the prevailing paradigm of survey sample inference is design-based, and survey samplers have a widespread distrust of models.

2. Survey Weighting, Prediction, and Design vs. Model-Based Inference

The clash between two approaches to weighting survey data puzzled me as a student of statistics. Early on we learn about linear regression, fitted by ordinary least squares (OLS), which is optimal for a model that assumes that the residual variance is constant for all values of the covariates. If the variance of the residual for unit i is σ^2/u_i for some known constant u_i, then better inferences are obtained by weighted least squares, with unit i assigned a weight proportional to u_i. This form of weighting is model-based, since the linear regression model for the outcome (say Y) has been modified to incorporate a non-constant residual variance.

Later I took a course in survey sampling, and learnt about a different form of weighting, based on the selection probabilities. If unit i is sampled with selection probability π_i, then the survey sampler replaces OLS by weighted least squares, weighting the contribution of unit i to the least squares equations by $w_i \propto 1/\pi_i$, the inverse of the probability of selection. This form of weighting is design-based, with π_i relating to the selection of units : since unit i "represents" $1/\pi_i$ units of the population, it receives a weight proportional to $1/\pi_i$ in the regression.

Both forms of weighting seem plausible, but they are not necessarily the same. So which is correct? The answer is not obvious - the role of sampling weights in regression has been extensively debated in the literature - see for example Konijn (1962), Brewer and Mellor (1973), Dumouchel and Duncan (1983), Smith (1988), Little (1991), Pfeffermann (1993), Korn and Graubard (1999). In fact, it rests fundamentally on whether one adopts a design-based or model-based perspective on statistical inference.

The design-based approach to survey inference (e.g. Hansen, Hurwitz and Madow 1953, Kish 1965, Cochran 1977) has the following main features. For a population with N units, let $Y = (y_1, \dots, y_N)$ where y_i is the set of survey variables for unit i, and let $I = (I_1, \dots, I_N)$ denote the set of *inclusion indicator variables,* where $I_i = 1$ if unit i is included in the sample and $I_i = 0$ if it is not included. Design-based inference for a finite population quantity $Q = Q(Y)$ involves the choice of an estimator $\hat{q} = \hat{q}(Y_{\text{inc}}, I)$, a function of the observed part Y_{inc} of Y, that is unbiased or approximately unbiased for Q with respect to the distribution I; and the choice of a variance estimator $\hat{v} = \hat{v}(Y_{\text{inc}}, I)$ that is unbiased or approximately unbiased for the variance of $\hat{q}$ with respect to the distribution of I. Inferences are then generally based on normal large sample approximations. For example, a 95% confidence interval for Q is $\hat{q} \pm 1.96\sqrt{\hat{v}}$.

The model-based approach to inference bases inference on the distribution of Y, and usually does not overtly consider a distribution for I; while assumptions of randomization lurk in the background, they are not the basis for the inference. The model for the survey outcomes Y is used to predict the non-sampled values of the population, and hence finite population quantities Q. There are two major variants : superpopulation modelling and Bayesian modelling. In superpopulation modelling (e.g. Royall 1970; Thompson 1988; Valliant, Dorfman and Royall 2000), the population values of Y are assumed to be a random sample from a "superpopulation", and assigned a probability distribution $p(Y \mid \theta)$ indexed by fixed parameters θ. Bayesian survey inference (Ericson 1969, 1988; Basu 1971; Scott 1977; Binder 1982; Rubin 1983, 1987; Ghosh and Meeden 1997, Little 2004) requires the specification of a prior distribution $p(Y)$ for the population values. Inferences for finite population quantities $Q(Y)$ are then based on the posterior predictive distribution $p(Y_{\text{exc}} \mid Y_{\text{inc}})$ of the non-sampled values (say Y_{exc}) of Y, given the sampled values Y_{inc}. The specification of the prior distribution $p(Y)$ is often achieved via a parametric model $p(Y \mid \theta)$ indexed by parameters θ, combined with a prior distribution $p(\theta)$ for θ, that is :

$$p(Y) = \int p(Y \mid \theta)p(\theta)d\theta.$$

The posterior predictive distribution of Y_{exc} is then

$$p(Y_{\text{exc}} \mid Y_{\text{inc}}) \propto \int p(Y_{\text{exc}} \mid Y_{\text{inc}}, \theta)p(\theta \mid Y_{\text{inc}})d\theta$$

where $p(\theta \mid Y_{\text{inc}})$ is the posterior distribution of the parameters, computed via Bayes' Theorem :

$$p(\theta \mid Y_{\text{inc}}) = p(\theta)P(Y_{\text{inc}} \mid \theta)/p(Y_{\text{inc}}),$$

where $p(\theta)$ is the prior distribution, $p(Y_{\text{inc}} \mid \theta)$ is the likelihood function, viewed as a function of θ, and $p(Y_{\text{inc}})$ is a normalizing constant. This posterior distribution induces a posterior distribution $p(Q \mid Y_{\text{inc}})$ for finite population quantities $Q(Y)$.

The specification of $p(Y \mid \theta)$ in this Bayesian formulation is the same as in parametric superpopulation modelling, and in large samples the likelihood based on this distribution dominates the contribution from the prior for θ. As a result, large-sample inferences from the superpopulation modelling and Bayesian approaches are often similar.

Example 2.1 Estimating a Mean from a Stratified Sample

Consider the simple case of estimation of a finite population mean $\overline{Y}$ from a stratified random sample. Suppose the population is divided into J strata, and let N_j be the known population count in stratum j and $\overline{Y}_j$ the unknown population mean in stratum j. The quantity of interest is $Q = \overline{Y} = \sum_{j=1}^{J} P_j \overline{Y}_j$, where $P_j = N_j/N$ is the proportion of the population in stratum j. We assume that a random sample of size n_j of the N_j units are sampled in stratum j, and let $\{y_{ji}, i = 1, \ldots, n_j\}$ denote the set of sampled Y-values in stratum j. Then $Y_{\text{inc}} = \{y_{ji}, j = 1, \ldots, J; i = 1, \ldots, n_j\}$. Stratified random sampling is defined by :

$$Pr(I_{ji} = 1) = \left[\binom{N_j}{n_j}\right]^{-1}, \text{ if } \sum_{i=1}^{N_j} I_{ji} = n_j, \text{ and 0 otherwise .}$$

The usual estimator of $\overline{Y}$ in this setting is the stratified mean

$$\hat{q} = \overline{y}_{\text{st}} \equiv \sum_{j=1}^{J} P_j \, \overline{y}_j = \left(\sum_{j=1}^{J} n_j \, \overline{y}_j / \pi_j\right) / \left(\sum_{j=1}^{J} n_j / \pi_j\right), \qquad (2.1)$$

where $\overline{y}_j$ is the sample mean in stratum j. The estimator (2.1) is the weighted mean of the sampled units, where units in stratum j are weighted by the inverse of their selection probability $\pi_j = n_j / N_j$.

Consider now a model-based approach. Suppose we assume the model

$$y_{ji} \sim_{\text{ind}} \text{ Nor } (\mu, \sigma^2 / u_j) \qquad (2.2)$$

where Nor (a, b) denotes the normal distribution with mean a, variance b, u_j is known, and the non-informative prior distribution

$$p(\mu, \log \sigma^2) = \text{ const.} \qquad (2.3)$$

The posterior mean of the population total is

$$\overline{y}_u = \left(\sum_{j=1}^{J} n_j \, u_j \, \overline{y}_j\right) / \left(\sum_{j=1}^{J} n_j u_j\right), \qquad (2.4)$$

which weights cases in stratum j by u_j, rather than $1/\pi$.

The application of design weights in this example is not controversial, and the stratified mean is difficult to beat as an estimator except in unusual situations. Indeed, the model-based estimator (2.4) is not recommended, since it is vulnerable to the assumption that the stratum means are equal. If the model (2.2) - (2.3) is changed to allow a separate mean in each stratum :

$$y_{ji} \sim_{\text{ind}} \text{Nor}\ (\mu_j, \sigma^2/u_j) \tag{2.5}$$

$$p(\mu_j, \log \sigma^2) = \text{const.}\ , \tag{2.6}$$

the posterior mean is then the stratified mean (2.1), so the design and model-based estimates correspond. Usually allowing a separate mean in each stratum is sensible, since strata are generally chosen to be related to survey outcomes; we do not determine strata by the toss of a coin.

In other settings, the design-weighted Horvitz-Thompson estimator (Horvitz and Thompson 1952) can lead to nonsensical estimates. Basu (1971) gave the following famous and amusing example :

Example 2.2 Basu's Elephants.

"The circus owner is planning to ship his 50 adult elephants and so he needs a rough estimate of the total weight of the elephants. As weighing an elephant is a cumbersome process, the owner wants to estimate the total weight by weighing just one elephant. Which elephant should he weigh? So the owner looks back on his records and discovers a list of the elephants' weights taken 3 years ago. He finds that 3 years ago Sambo the middle-sized elephant was the average (in weight) elephant in his herd. He checks with the elephant trainer who reassures him (the owner) that Sambo may still be considered to be the average elephant in the herd. Therefore, the owner plans to weigh Sambo and take $50y$ (where y is the present weight of Sambo) as an estimate of the total weight $Y = Y_1 + Y_2 + \ldots Y_{50}$ of the 50 elephants. But the circus statistician is horrified when he learns of the owner's purposive sampling plan. "How can you get an unbiased estimate of Y this way?" protests the statistician. So, together they work out a compromise sampling plan. With the help of a table of random numbers they devise a plan that allots a selection probability of 99/100 to Sambo and equal selection probabilities of 1/4900 to each of the other 49 elephants. Naturally, Sambo is selected and the owner is happy. "How are you going to estimate Y?", asks the statistician. "Why? The estimate ought to be $50y$ of course," says the owner. "Oh! No! That cannot possibly be right," says the statistician, "I recently read an article in the *Annals of Mathematical Statistics* where it is proved that the Horvitz-Thompson estimator is the unique hyperadmissible estimator in the class of all generalized polynomial unbiased estimators. "What is the Horvitz-Thompson estimate in this case?" asks the owner, duly impressed. "Since the selection probability for Sambo in our plan was 99/100," says the statistician, "the proper estimate of Y is $100y/99$ and not $50y$." "And, how would you have estimated Y," inquires the incredulous owner, "if our sampling plan made us select, say, the big elephant Jumbo?" "According to what I understand of the Horvitz-Thompson estimation method," says the unhappy statistician, "the proper estimate of Y would then have been $4900y$, where y is Jumbo's weight." That is how the statistician lost his circus job (and perhaps became a teacher of statistics!)"

Design-based statisticians groan when modelers bring up Basu's example, since they view it as a caricature : no sensible design-based statistician would use the HT estimator in this case. Basu was using the example to make a theoretical point; the HT estimator has the useful property of design-unbiasedness in large samples, but no single estimator is optimal in all situations, and weighted estimators can do very badly, particularly in small samples. As a more realistic example, design-based statisticians deviate from strict weighting when outlying observations receive large weights, and dominate the estimator.

Slavish adoption of the design-weighted estimator without attention to whether the underlying model is reasonable is not wise. How can we tell when the HT estimator is not going to work? One approach is to consider the model for the population implied model by weighting. Specifically, consider creating an estimate of the population by replicating sample observation i $1/\pi_i$ times. Is the resulting population sensible as an approximation for the problem at hand? Clearly the answer is "yes" in Example 2.1, and "no" in Example 2.2. When the answer is no, better estimates exist.

The population that replicates the sample is a kind of model, and design-based statisticians cannot avoid models. On the other hand, model-based statisticians cannot avoid weights, since a model that ignores the survey weights is likely to be poorly calibrated, given the realities of model misspecification as exemplified by the absence of stratum means in (2). For other examples, see Kish & Frankel (1974), Hansen, Madow & Tepping (1983), Holt, Smith, and Winter (1980), and Pfeffermann and Holmes (1985).

My own philosophy of survey sampling inference, as for statistics in general, is calibrated Bayes, where inferences are Bayesian and based on models for Y, but models need to be calibrated in the sense of having good design-based properties in repeated sampling from the distribution of I (Box 1980, Rubin 1984, Little 2006). The calibrated Bayes philosophy leads to prediction models with relatively noninformative prior distributions, which incorporate design features appropriately, seeking both efficiency and robustness to model misspecification. My work in this area has been guided by this underlying principle.

For calibrated Bayesians, both the distribution of Y and the distribution of I are important – indeed a useful and unifying conceptual device is to formulate the model in terms of the joint distribution of both Y and I. The early literature of surveys focused either on the distribution of Y or the distribution of I, rather than the joint distribution of Y and I. This tended to lead to compartmentalization into design-based and model-based advocates. To my knowledge, the first person to explicitly model I and Y seems to be Rubin (1978), in a paper that was more focused on estimating treatment effects but also modelled the selection mechanism.

The joint modelling of Y and I in the survey context is well described in the book by Gelman et al. (2003). The following description is from Little (2003a). The model can be formulated as :

$$p(y_U, i_U \mid z_U, \theta, \phi) = p(y_U \mid z_U, \theta) \times p(i_U \mid z_U, y_U, \phi),$$

where U denotes universe as opposed to sample, y_U denotes the survey data, i_U the sample inclusion indicators, z_U denotes design variables, such as strata indicators, and θ, ϕ are unknown parameters. The likelihood of θ, ϕ based on the observed data (z_U, y_{inc}, i_U) is then :

$$L(\theta, \phi \mid z_U, y_{inc}, i_U) \propto p(y_{inc}, i_U \mid z_U, \theta, \phi) = \int p(y_U, i_U \mid z_U, \theta, \phi) dy_{exc}.$$

The more usual likelihood does not include the inclusion indicators i_U as part of the model. Specifically, the likelihood *ignoring the selection process* is based on the model for y_U alone :

$$L(\theta \mid z_U, y_{inc}) \propto p(y_{inc} \mid z_U, \theta) = \int p(y_U \mid z_U, \theta) dy_{exc}.$$

Applying Rubin's (1976) theory, sufficient conditions for ignoring the selection mechanism are :

Selection at Random (SAR) : $p(i_U \mid z_U, y_U, \phi) = p(i_U \mid z_U, y_{inc}, \phi)$ for all y_{exc}.

Distinctness : θ, ϕ have distinct parameter spaces.

Probability sample designs are generally both ignorable and known, in the sense that :

$$p(i_U \mid z_U, y_U, \phi) = p(i_U \mid z_U, y_{inc}),$$

where z_U represents known sample design information, such as clustering or stratification information. Thus the sampling mechanism can be ignored, provided the sample design information in z_U is included in the model. In the case of weighting, this means conditioning on the design variables that lead to differential weights. This analysis also provides a justification for randomization in design, since other forms of sampling, like quota sampling or purposive selection, do not necessarily satisfy the SAR assumption. Extensions to handle survey nonresponse are given in Little (1982, 2003b).

The sampling weights in Examples 2.1 and 2.2 are determined solely by the probabilities of selection. More generally, survey weights also involve components for survey nonresponse and for post-stratification to match known population distributions. The standard approach creates a composite weight for unit i of the form

$$w_i \propto w_{is} \times w_{in}(w_{is}) \times w_{ip}(w_{is}, w_{in}) \qquad (2.7)$$

where w_{is} is the sampling weight, $w_{in}(w_{is})$ is a nonresponse weighting factor and $w_{ip}(w_{is}, w_{in})$ is a post-stratification adjustment. In the remainder of this article I'll give some additional illustrations of prediction models that features like selection probabilities and survey nonresponse.

3. Weights that Incorporate Population Information

In Example 2.1 we noted that the weighting and prediction approaches yield the stratified mean in the case of the stratified example. Post-stratification is a closely related example :

Example 3.1 Inference for the Mean with Categorical Post-Strata.

Another situation where the design and model-based approaches intersect is estimation of the population mean of a variable Y from a simple random sample, given a categorical post-stratum variable Z with known distribution in the population. Let y_{ji} denote the value of Y for sampled unit i in post-stratum $Z = j$. Assume the model of Equations (2.5) - (2.6). The posterior distribution of the population mean has mean

$$\overline{y}_{\text{mod}} = \overline{y}_{\text{wt}} = \sum_{j=1}^{J} P_j \, \overline{y}_j = \sum_{j=1}^{J} w_j \, n_j \, \overline{y}_j / \sum_{j=1}^{J} w_j \, n_j, \qquad (3.1)$$

where in post-stratum $Z = j, P_j$ is the population proportion, n_j is the sample size, $\overline{y}_j$ is the sample mean, and $w_j = n \, P_j / n_j$. The estimate (3.1) is the post-stratified mean, also obtained in the design-based approach by applying post-stratification weights w_j to the sampled units in post-stratum j.

Asymptotically (3.1) works fine, but in small samples it is unstable. The situation here differs from stratification on Z, where the stratum counts $\{n_j\}$ are under the control of the sampler. With post-stratification, the $\{n_j\}$ are determined by which units happen to fall into post-stratum j. The post-stratum counts n_j in one or more post-strata may become very small, yielding large weights w_j; indeed (3.1) is not defined if for any j, $n_j = 0$, and it does not have a well-defined sampling distribution in repeated samples unless $\{n_j\}$ are constrained to be positive; for discussion of this point see Holt and Smith (1979) and Little (1993). Design-based approaches modify the weights, for example by pooling small post-strata. However, from a prediction perspective, the problem lies not in the weights, but in the unstable predictions $\overline{y}_j$ of the means in post-strata with small counts. The associated proportions P_j are, after all, known!

From a Bayesian perspective, the posterior distribution of $\overline{Y}$ for the model (2.5) - (2.6) is a mixture of t distributions, and as such incorporates t corrections from estimating the variance that are not available under the design-based approach, which is basically asymptotic. Concerning the instability of (3.1), the Bayesian solution is to modify the prior distribution (2.6) to allow borrowing of strength across post-strata. One such modification is

$$\mu_j \sim_{\text{ind}} N\left(\mu, \tau^2\right), p\left(\mu, \log \sigma^2, \tau^2\right) = \text{const.},$$

which yields predictions that effectively shrink the weights w_j to a constant. This approach to weight shrinkage is discussed in Little (1993), and extensions in the presence of covariates are discussed in Lazzeroni and Little (1998) and Elliott and Little (2000).

Example 3.2 Categorical Strata and Post-Strata.

Suppose now that we have a stratified sample, with stratifier Z_1 with population distribution $\{P_{1j}, j = 1, \ldots, J\}$, and we also know the population distribution $\{P_{2k}, k = 1, \ldots, K\}$ of a post-stratification variable Z_2. The traditional weighting approach (2.7) is to post-stratify the stratification weights so that the weighted sample counts match the population distribution of Z_2. That is, the composite weight for units in stratum j, post-stratum k is

$$w_{jk} = w_{1j} \times w_{2k \cdot j},$$

where $w_{1j} = n P_{1j}/n_{1j}$ and $w_{k \cdot j} = n P_{2k} w_{1j} / \sum_l w_{1\,l}$. Interestingly, these weights lead to stratum counts that do not match the population distribution of Z_1. From a modelling perspective, the data about the joint distribution of Z_1 and Z_2 consists of the sample counts $\{n_{jk}\}$ and the known marginal distributions of Z_1 and Z_2. A saturated model for the joint distribution of Y, Z_1 and Z_2 takes the form :

$$\begin{aligned} &\{n_{jk}\} \sim \text{ MNOM } (n, P_{jk}); \\ &y_{jki} \sim \text{ Nor } (\mu_{jk}, \sigma^2_{jk}), p(\mu_{jk}, \log \sigma^2_{jk}) = \text{ const.} \end{aligned} \tag{3.2}$$

Maximum likelihood estimates $\{\hat{P}_{jk}\}$ of $\{P_{jk}\}$ are obtained by ranking the sample counts to match the Z_1 and Z_2 margins by iterative proportional fitting, yielding weights that match both of these margins. The maximum likelihood estimate of the population mean of Y is then

$$\overline{y}_{\text{mod}} = \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{P}_{jk} \overline{y}_{jk}. \tag{3.3}$$

Classification by both Z_1 and Z_2 increases the likelihood of small counts $\{n_{jk}\}$ in some cells, so modifications of (3.2) for predicting the cell means may be important. One possibility is to replace the saturated model by

$$y_{jki} \sim \text{ Nor } \left(\mu + \alpha_j + \beta_k + \gamma_{jk}, \sigma^2_{jk}\right),$$
$$\sum_{j=1}^{J} \alpha_j = \sum_{k=1}^{K} \beta_k = 0, \gamma_{jk} \sim \text{ Nor } \left(0, \tau^2\right) \tag{3.4}$$

which results in shrinkage of the sample mean $\overline{y}_{jk}$ towards the fitted mean for the additive model relating Y to Z_1 and Z_2. In summary, adopting a prediction perspective (a) corrects the usual estimator to match both stratum and post-stratum margins; (b) provides t corrections for estimating the variance, as in Example 3.1; and (c) allows modifications of the estimator (3.3) in small samples by modifying the prior distribution of the cell means.

Example 3.3 Probability Proportional to Size (PPS) Sampling.

The weights in Examples 3.1 and 3.2 incorporate information from categorical variables in the population. Sometimes sample designs involve stratifiers that are continuous variables. A common design with a continuous stratifier is PPS sampling, where units are selected with probability proportional to a size variable Z known for all units in the population. The standard design-based estimator in this setting is the HT estimator

$$\overline{y}_{\text{wt}} = \frac{1}{N}\left(\sum_{i=1}^{n} y_i/\pi_i\right) \tag{3.5}$$

where π_i is the probability of selection for unit i. From a modelling perspective, the objective is to base estimates on predictions from a regression model for the distribution of Y given Z. The estimator (3.1) is approximately the prediction estimator for the "HT model"

$$y_i \mid z_i \sim \text{ Nor } \left(\beta z_i, \sigma^2 z_i^2\right). \tag{3.6}$$

The estimator (3.1) tends to be efficient when the HT is satisfied, but does poorly when this model is seriously violated. Zheng and Little (2003, 2004, 2005) consider predicting the non-sampled values using the more flexible penalized spline model

$$y_i \sim \text{ Nor } (f(z_i, \beta), \sigma^2 z_i^k),$$

where f is a spline function :

$$f(z_i, \beta) = \beta_0 + \sum_{j=1}^{p} \beta_j \, z_i^j + \sum_{l=1}^{m} \beta_{l+p} \, (z_i - \kappa_l)_+^p, i = 1, \ldots, N.$$

Here $k \geq 0$ is a constant reflecting the knowledge of the error variance and the constants $\kappa_1 < \ldots < \kappa_m$ are selected fixed knots, and $(u)_+^p = u^p$ if $u > 0$, and 0, otherwise; and$(\beta_{p+1} \ldots, \beta_{p+m})^T$ are assumed $\text{Nor}(0, \tau^2 I_m)$. This model relaxes the assumption that the relationship between Y and Z is linear. Zheng and Little (2005) show by simulation that prediction inferences based on this model yield gains over the HT estimator in both efficiency and confidence coverage when the HT model (3.6) is violated, while sacrificing little in terms of efficiency when the HT model is satisfied. Chen, Elliott and Little (2009) develop Bayesian inference for a population proportion from unequal probability samples, where the probit of the probability that $y_i = 1$ is modelled as penalized spline of the size variable. They also show gains in terms of efficiency and confidence coverage compared with the HT estimator, and generalized regression extensions of the HT estimator.

4. Unit and Item Nonresponse

In the context of survey nonresponse, weighting adjustments are common in the case of unit nonresponse, as in the following example.

Example 4.1 Unit Nonresponse in Surveys

Suppose that respondents and nonrespondents are classified into C adjustment cells based on covariates X observed for both. The nonresponse weight in cell c is then the inverse of the estimated response rate in that cell. This is also the prediction estimator for a model that assumes a different mean for the outcome in each adjustment cell. Some comments on this approach follow :

1. Given extensive covariate information, adjustment cells should be chosen that are predictive of both the survey outcomes and

of nonresponse. Adjustment cell weighting, and extensions based on models for the propensity to respond, tend to focus on good predictors of response, but Little and Vartivarian (2005) argue that having a good predictor of the outcome is more important; these can actually improve efficiency of estimation, and good predictors of nonresponse that are not related to the outcome simply increase variance without reducing bias.

2. When the sampling weights are not constant within adjustment cells, it is common practice to compute the nonresponse weight as the inverse of the weighted response rate, where units are included in the rate weighted by their sampling weights. This "weight squared" approach does not correct for bias when the outcome is related both to the adjustment cell variable and the stratification variable, as is demonstrated by simulations in Little and Vartivarian (2003).

3. Since nonresponse is not under the control of the sampler, highly variable nonresponse weights are possible, as when the fraction of respondents in an adjustment cell is small. Thus shrinkage of the nonresponse weights may be attractive, and this is accomplished by putting a proper prior on the adjustment cell means, as was done in Example 3.1 in the case of post-stratification.

Example 4.2 Item Nonresponse in Surveys.

Item nonresponse occurs when particular items in the survey are missing, because they were missed by the interview, or the respondent declined to answer particular questions. For item nonresponse the pattern of missing values is general complex and multivariate, and substantial covariate information is available to predict the missing values in the form of observed items. These characteristics make weighting adjustments unattractive, since weighting methods are difficult to generalize to general patterns of missing data (Little 1988) and make limited use of information in the incomplete cases.

A common practical approach to item missing data is imputation, where missing values are filled in by estimates and the resulting data are analyzed by complete-data methods. In this approach incomplete cases are retained in the analysis. Imputation methods until the late 1970's lacked an underlying theoretical rationale. Pragmatic estimates of the missing values were substituted, such as unconditional or conditional means, and inferences based on the filled-in data. A serious defect with the method is that it "invents data". More specifically, a single imputed value cannot represent all of the uncertainty about which value to impute, so analysis that treat imputed values just like observed values generally underestimate uncertainty, even if nonresponse is modelled correctly. Rubin's (1987) theory of multiple imputation (MI) put imputation on a firm theoretical footing, and also provided simple ways of incorporating imputation uncertainty into the inference. Instead of imputing a single set of draws for the missing values, a set of Q (say $Q = 10$) datasets are created, each containing different sets of draws of the missing values from their predictive distribution given the observed data. The analysis of interest is then applied to each of the Q datasets and results are combined using simple multiple imputation combining rules (Rubin 1987; Little and Rubin, 2002). An alternative to multiple imputation is to use sample re-use methods that reimpute the data on each replicate sample (Rao 1996).

5. Conclusion

The above examples suggest that weighting provides a useful all-purpose approach to large sample estimation in surveys, but Bayesian predictive models yield useful extensions and refinements, provided careful attention is paid to incorporating the survey design. Some advantages of the Bayesian approach are :

(1) It provides a unified approach to survey inference, aligned with mainline statistics approaches in other application areas such as econometrics.

(2) In large samples and with uninformative prior distributions, results can parallel those from design-based inference, as we have seen in the case of stratified and post-stratified sampling in Examples 1.1 and 2.1.

(3) The Bayesian approach is well equipped to handle complex design features such as clustering through random cluster models (Scott and Smith 1969), stratification through covariates that distinguish strata, nonresponse (Little 1982; Rubin 1987; Little and Rubin 2002) and response errors.

(4) The Bayesian approach may yield better inferences for small sample problems where exact frequentist solutions are not available, by propagating error in estimating parameters. For example, the posterior distribution of the mean for inference from normal stratified samples in Example 2.1 is a mixture of t distributions that propagates uncertainty in estimating the stratum variances. On the other hand, the standard design-based inference based on the normal distribution assumes that the stratum variances are estimated without error from the sample.

(5) The Bayesian approach allows prior information to be incorporated, when appropriate; and

(6) Likelihood-based approaches like Bayes or maximum likelihood have the property of large-sample efficiency, and hence match or outperform design-based inferences if the model is correctly specified.

An alternative to a direct Bayesian modelling approach for incorporating auxiliary information is model-assisted estimation, where a model is applied to predict the non-sampled values, and then the predictions are calibrated by applying the HT estimator to the residuals from that model (Särndal, Swensson and Wretman 1992). Specifically, the generalized regression estimator of T takes the form :

$$\hat{T}_{\text{gr}} = \sum_{i=1}^{N} \hat{y}_i + \sum_{i \text{ sampled}} (y_i - \hat{y}_i)/\pi_i, \tag{5.1}$$

where $\hat{y}_i$ is the prediction from a linear regression model relating Y to the covariates. While this approach is popular and yields design-consistent (Isaki and Fuller 1982) estimates, my personal preference is to choose robust models that yield design-consistent estimates, that is, to correct the model rather than to correct the estimator. It is relatively easy to find models that yield design consistent estimates (e.g. Firth and Bennett (1998), and there is little evidence that calibration yield better inferences than direct model estimates when the latter are design consistent.

A criticism of the model-based approach is that it is impractical for large-scale survey organizations : the work in developing strong models, and the computational complexity of fitting them, is not suited to the demands of "production-oriented" survey analysis. However, attention to models is needed in model-assisted approaches, even when the basis for inference is the sample design. Also, computational power has expanded dramatically since the days of early model versus randomization debates, and much can be accomplished using software for mixed models in the major statistical packages (SAS 1992; Pinheiro and Bates 2000) or Bayesian software based on MCMC methods such as BUGS. (Spiegelhalter, Thomas, and Best 1999). Bayesian software targeted at complex survey problems would increase the utility of this approach for practitioners. Also, guidance on "off-the-shelf" models for routine application to standard sample designs would be useful, although no statistical procedure, design or model-based, should be applied blindly without any attention to diagnostics of fit to the data.

REFERENCES

Box, G. E. P. (1980). Sampling and Bayes inference in scientific modelling and robustness (with discussion), *J. Royal Statist. Soc.* A **143** 383 - 430,

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part 1, in *Found. Statist. Inf.* Toronto : Holt, Rinehart and Winston, pp. 203 - 242.

Binder, D. A. (1982). Non-parametric Bayesian Models for Samples from Finite Populations, *J. Royal Statist. Soc.* B **44** 3, 388 - 393.

Breidt, F. J. and Opsomer, J. D. (2000). Local Polynomial Regression Estimators in Survey Sampling, *Annals. Statist.* **28** 1026 - 1053.

Brewer, K. R. W. and Mellor, R. W. (1973). The Effect of Sample Structure on Analytical Surveys, *Australian J. Statist.* **15** 145 - 152.

Chen, Q., Elliott, M. R. and Little, R. J. (2009). Bayesian Penalized Spline Model-Based Estimation of the Finite Population Proportion for Probability-Proportional-to-Size Samples, Submitted to *Survey Methodology*

Cochran, W. G. (1977). *Sampling Techniques* 3rd Edition, New York : Wiley.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics,* London : Chapman Hall.

Dumouchel, W. H. and Duncan, G. J. (1983). Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples, *J. American Statist. Assoc.* **78** 535 - 543.

Elliott, M. R. and Little, R. J. A. (2000). Model-Based Alternatives to Trimming Survey Weights, *J. Official Statist.* **16** No. 3, 191 - 209.

Ericson, W. A. (1969). Subjective Bayesian Models in Sampling Finite Populations, *J. Royal Statist. Soc.* B **31** 195 - 234.

Ericson, W. A. (1988). Bayesian Inference in Finite Populations, *in Handbook of Statistics* **6** Amsterdam : North-Holland, pp. 213 - 246.

Firth, D. and Bennett, K. E. (1998). Robust Models in Probability Sampling, *J. Royal Statist. Soc.* B **60** 3 - 21.

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

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

Godambe, V. P. (1955). A Unified Theory of Sampling from Finite Populations, *J. Royal Statist. Soc.* B **17** 269 - 278.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953). *Sampling Survey Methods and Theore* Vols. I and II, New York : Wiley.

Hansen, M. H., Madow, W. G. and Tepping, B. J. (1983). An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys, *J. American Statist. Assoc.* **78** 776 - 793 (with discussion).

Holt, D. and Smith, T. M. F. (1979). Poststratification, *J. Royal Statist. Soc.* A **142** 33 - 46.

Holt, D., Smith, T. M. F. and Winter, P. D. (1980). Regression Analysis of Data from Complex Surveys, *J. Royal statist. Soc.* A **143** 474 - 487.

Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of Sampling without Replacement from a Finite Universe, *J. American Statist. Assoc.* **47** 663 - 685.

Isaki, C. T. and Fuller, W. A. (1982). Survey Design Under the Regression Superpopulation Model, *J. American Statist. Assoc.* **77** 89 - 96.

Kish, L. (1965). *Survey Sampling,* New York : Wiley.

Kish, L. and Frankel, M. R. (1974). Inferences from Complex Samples (with discussion), *J. Royal Statist. Soc.* B **36** 1 - 37.

Konijn, H. S. (1962). Regression Analysis in Sample Surveys, *J. American Statist. Assoc.* **57** 590 - 606.

Korn, E. L. and Graubard, B. I. (1999). *Analysis of Health Surveys,* New York : Wiley.

Lazzeroni, L. C. and Little, R. J. A. (1998). Random-Effects Models for Smoothing Post-Stratification Weights, *J. Official Statist.* **14** 61 - 78.

Little, R. J. A. (1982). Models for Nonresponse in Sample Surveys, *J. American Statist. Assoc.* **77** 237 - 250.

Little, R. J. A. (1988). Missing Data in Large Surveys, *J. Bus. and Econ. Statist.* **6** 287 - 301. (with discussion).

Little, R. J. A. (1991). Inference with Survey Weights, *J. Official Statist.* **7** 405 - 424.

Little, R. J. A. (1993). Post-Stratification : a Modeler's Perspective, *J. American Statist. Assoc.* **88** 1001 - 1012.

Little, R. J. A. (2003a). The Bayesian Approach to Sample Survey Inference, *Anal. Survey Data* R. L. Chambers & C. J. Skinner, eds., pp. 49 - 57. Wiley : New York.

Little, R. J. A. (2003b). Bayesian Methods for Unit and Item Nonresponse. In *Analysis of Survey Data* R. L. Chambers & C. J. Skinner, eds., pp 289 - 306. Wiley : New York.

Little, R. J. A. (2004). To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling, *J. American Statist. Assoc.* **99** 546 - 556.

Little, R. J. A. (2006). Calibrated Bayes : A Bayes/Frequentist Roadmap. *The American Statist.* **60** 3, 213 - 223.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data,* 2nd edition, New York : Wiley.

Little, R. J. A. and Vartivarian, S. (2003). On weighting the Rates in Nonresponse Weights, *Statist. Medicine* **22** 1589 - 1599.

Little, R. J. A. and Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology* **31** 161 - 168.

Mahalanobis, P. C. (1943). Recent Experiments in Statistical Sampling in the Indian Statistical Institute. *J. Royal Statist. Soc.* **109** 325 - 378.

Pfeffermann, D. (1993). The Role of Sampling Weights when Modelling Survey Data, *International Statist. Rev.* **61** 317 - 337.

Pfeffermann, D. and Holmes, D. J. (1985). Robustness Considerations in the Choice of Method of Inference for Regression Analysis of Survey Data, *J. Royal Statist. Soc.* A **148** 268 - 278.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus,* New York : Springer.

Rao, J. N. K. (1996). On Variance Estimation with Imputed Survey Data, *J. American Statist. Assoc.* **91** 499 - 506.

Rao, J. N. K. (1997). Developments in Sample Survey Theory : An Appraisal, *Canadian J. Statist.* **25** 1 - 21.

Rao, J. N. K. (2003). *Small Area Estimation* New York : Wiley.

Royall, R. M. (1970). On Finite Population Sampling Under Certain Linear Regression Models, *Biometrika* **57** 377 - 387.

Rubin, D. B. (1976). Inference and Missing Data, *Biometrika* **53** 581 -592.

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician, *Annals Statist.* **12** 1151 - 1172.

Rubin, D. B. (1983). Comment on "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys", by M. H. Hansen, W. G. Madow and B. J. Tepping, *J. American Statist. Assoc.* **78** 803 - 805.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys,* Wiley : New York.

Särndal, C. E., Swensson, B. and Wretman, J. H. (1992). *Model Assisted Survey Sampling,* Springer Verlag : New York.

SAS (1992). The Mixed Procedure, in *SAS/STAT Software : Changes and Enhancements, Release 6.07,* Technical Report P - 229, SAS Institute, Inc., Cary, NC.

Scott, A. J. (1977). Large-Sample Posterior Distributions for Finite Populations, *Annals. Math. Statist.* **42** 1113 - 1117.

Scott, A. J. and Smith, T. M. F. (1969). Estimation in Multistage Samples, *J. American Statist. Assoc.* **64** 830 - 840.

Smith T. M. F. (1988). To Weight or not to Weight, that is the Question, in *Bayesian Statistics* 3, J. M. Bernado, M. H. DeGroot and D. V. Lindley, eds., Oxford, UK : Oxford University Press, pp. 437 - 451.

Spiegelhalter, D. J., Thomas, A. and Best, N. J. (1999). *WinBUGS Version 1.2 User Manual,* MRC Biostatistics Unit, Cambridge, UK.

Thompson, M. E. (1988). Superpopulation Models, *Encyclopedia Statist. Sci.* (Vol. 1) **9** 93 - 99.

Valliant, R., Dorfman, A. H. and Royall, R. M. (2000). *Finite Population Sampling and Inference : A Prediction Approach,* New York : Wiley.

Zheng, H. and Little, R. J. A. (2003). Penalized Spline Model-Based Estimation of the Finite Population Total from Probability-Proportional-To- Size Samples, *J. Official Statist.* **19** 2 99 - 117.

Zheng, H. and Little. R. J. A. (2004). Penalized Spline Nonparametric Mixed Models for Inference about a Finite Population Mean from Two-Stage Samples, *Survey Methodology* **30** 2 209 - 218.

Zheng, H. and Little, R. J. A. (2005). Inference for the Population Total from Probability-Proportional-to-Size Samples Based on Predictions from a Penalized Spline Nonparametric Model, *J. Official Statist.* **21** 1 - 20.

DISCUSSION

ANDREW GELMAN
Department of Statistics
and Department of Political Science,
Columbia University, USA

Survey weights, like sausage and legislation, are best appreciated by those who are placed a respectable distance from their manufacture. For those of us working inside the factory, vigorous discussion of methods is welcome. I enjoyed Rod Little's review of the connections between modeling and survey weighting and have just a few comments.

I like Little's discussion of model-based shrinkage of post-stratum averages, which, as he notes, can be seen to correspond to shrinkage of weights. I would only add one thing to his formula at the end of his Example 3, which is that his regression model can include poststratum-level predictors; for example, if poststrata are indexed by sex, age, ethnicity, and education, the model could include indicators for each of these factors, and even two-way effects as necessary. This seems to be where he is leading in his Example 4.

I also found Little's discussion of probability proportional to size (pps) sampling helpful; this is a problem that I have found difficult to attack using model-based methods. The spline model for the response given stratum size seems like a good way to go. My only comment here is that I have always associated pps sampling with two-stage cluster sampling, in which clusters are sampled pps and then a fixed-size sample is drawn from each cluster. In this case, the classical pps unit weights are all equal, and it is hard for me to believe that a model-based approach can improve much upon this, at least in settings in which the measures of size used in the sampling are not far from the actual sizes of the clusters.

As Little emphasizes, weights and other survey adjustment procedures are intended to correct for known differences between sample and population. I would rephrase his claim that "model-based statisticians cannot avoid weights", and instead say that statisticians cannot avoid adjustment, but this adjustment could take other forms, such as my personal favorite of model-based poststratification (Gelman and T. C. Little, 1997, Gelman, 2007).

Don Rubin once told me he would prefer to do all survey adjustment using multiple imputation; for example, in a survey of 1000 American adults, he would impute the missing responses for the other 250 million. I asked him if that was impractical, and he replied that the imputation could only realistically be performed conditional on information available on all 250 million; i.e. Census demographics, and thus the imputation would in fact be equivalent to fitting a regression model of the response conditional on key demographic variables recorded in the survey and then summing over Census numbers to get national estimates. Depending on the method used to estimate the regression, it might be possible to approximate such an estimate as a weighted average over the sample (Little, 1993, Gelman, 2006) but it would be stretching it to call this a use of weights. In addition, under this approach, the approximate weights depend on the fitted model and thus on the outcome being modeled. Having a different weight for each question on the survey would seem to go beyond the usual conception of survey weighting.

Even in the design-based world, survey weights are not always based on selection probabilities. Consider the following poststratification example : A national survey of American adults is conducted and yields 600 female respondents and 400 males. The standard poststratified estimate is to take 0.52 times the average response for the women plus 0.48 times the average for the men, which corresponds to unit weights of 0.52/0.60 for each woman and 0.48/0.40 for each man. These are not inverse selection probabilities but rather are based on the known proportions of men and women in the sample and population. The weights are not even estimated inverse selection probabilities, a fact which we can see by noting that, even the actual selection probabilities were given to us, we would not use them: the poststratification weights are better. Which is perfectly consistent with the points Little makes in his article.

ADDITIONAL REFERENCES

Gelman, A. (2007). Struggles with survey weighting and regression modelling (with discussion). *Statistical Science.*

Gelman, A., and Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology* **23** 127 - 135.

DISCUSSION

DANNY PFEFFERMAN
Jerusalem and Southampton
Statistical Science Research Institute,
Hebrew University, UK

Reading the work of Rod Little is always very challenging and instructive, and this article is no exception. On first reading I thought that I actually agree with everything said, but after a second and more thorough reading I found some points that are probably worthy further discussion, which I do below.

Little starts his discussion by raising the question of whether when fitting regression models to survey data, one should weight by the inverse of the variances under the model, or by the inverse of the sample selection probabilities. The first form of weighting is 'model-based'; the second form is 'design-based'. Little points out that both forms of weights are "plausible", but to me they actually represent different kind of conditioning. Suppose that we consider the population values $Y_U = (Y_1 \ldots, Y_N)$ as being generated from a superpopulation model ξ (the regression model in Little's example). Denote the sampling design by $p = p(I)$, where $I = (I_1 \ldots, I_N)$ and I_i defines the sample inclusion indicator. We can now view the sample data as being the outcome of the two random processes ξ, and then p. Going back to the regression example, estimating the regression coefficients under the combined $p\ \xi$ distribution suggests weighting by both the inverse of the model variances and by the inverse of the selection probabilities. Assuming non-informative sampling and conditioning on I suggests weighting by only the inverse of the variances as the optimal weighting, whereas conditioning on Y_U suggests weighting by only the inverse of the selection probabilities. When the sample selection is informative, a third conditioning becomes plausible. Following a result by Pfeffermann and Sverchkov (1999), weighting by both the inverse of the variances and the inverse of the sample selection probabilities is the optimal weighting (in a least squares sense) under the conditional distribution of the sampled y-values, given the selected sample (see also below).

Later on, Little advocates the joint modelling of Y and I as part of the "calibrated Bayesian approach", and hence the use of the *full likelihood*, $L(\theta, \phi \mid z_U, y_{inc}, i_U) \prec \int p(i_U \mid z_U, y_U; \phi) p(y_U \mid z_U; \theta) dy_{exc}$ for inference, where (z_U, y_{inc}, i_U) defines the observed data and y_{exc}

the unobserved y-values. I fully agree that this should be the preferred likelihood under either the frequentist approach or as part of a Bayesian model because it encompasses all the design information, when available, and generally guarantees that the sampling mechanism can be ignored in the inference process. Unfortunately, the full likelihood is not always operational in a secondary analysis, because some or all of the values of the design variables z may not be known, notably for the nonsampled units. This is usually the case when, for example, the design variables include a size variable used for PPS sampling (Example 5 in the paper). So, even if $p(i_U \mid z_U, y_U; \phi) = p(i_U \mid z_U, y_{inc}; \phi)$, guaranteeing sampling ignorability, the use of the full likelihood requires knowledge of z_U or an adequate summary of it. When, in addition, $p(y_{inc} \mid z_U; \theta) = p(y_{inc} \mid z_{inc}; \theta)$, one can estimate θ, but knowledge of the design variables values for the nonsampled units is still required for predicting the unobserved (excluded) y-values via the predictive model $p(y_{exc} \mid z_U, y_{inc})$. In theory, one could integrate the likelihood or the predictive model over the joint distribution of the missing design variables, but when there are many of them, this might not be practical.

In the present article Little restricts to the estimation (prediction) of finite population totals. Often, however, survey data are used for the fitting of structural models per se, such as the regression model mentioned before, and not for estimating population totals. Denoting the independent variables in the model by x, the focus of inference is in this case the model $f_\xi(y \mid x)$, and not the model $f_\xi(y \mid x, z)$, which may not even be interpretable. Here again, the sampling mechanism can be ignored if z_U is included among the model covariates, but fitting the model $f_\xi(y \mid x)$ requires then integrating the extended model $f_\xi(y \mid x, z_U)$ with respect to the distribution of $z_U \mid x$, which could be formidable.

How can we deal with these problems? One possibility, not the only one or necessarily the best one, is to consider instead of the full likelihood the conditional sample likelihood, or more generally, to base the inference on the conditional sample distribution (hereafter, the sample model). Following Pfeffermann et al. (1998) and Pfeffermann and Sverchkov (1999), the sample model is defined as,

$$f_s(y_i \mid x_i) \overset{def}{=} f(y_i \mid x_i, i \in s) = \frac{Pr(i \in s \mid y_i, x_i) f_\xi(y_i \mid x_i)}{Pr(i \in s \mid x_i)}$$

$$= \frac{E_s(w_i \mid x_i) f_\xi(y_i \mid x_i)}{E_s(w_i \mid y_i, x_i)}, \qquad (1)$$

where $f_\xi(y_i \mid x_i)$ defines, as before, the 'superpopulation' model, $w_i = 1/Pr(i \in s)$ is the (base) sampling weight and $E_s(\cdot)$ is the expectation under the sample model (the model holding for the sample data). Notice that when $Pr(i \in s \mid y_i, x_i) = Pr(i \in s \mid x_i)$ for all $y_i, f_s(y_i \mid x_i) = f_\xi(y_i \mid x_i)$. On the other hand, $Pr(i \in s \mid y_i, x_i)$ is generally not the same as $\pi_i = Pr(i \in s)$, which depends on the design values z_U. However, the use of the sample model only requires modelling $E_s(w_i \mid y_i, x_i)$, (which is not always trivial in practice), thus avoiding the need to know all the values of the design variables and incorporate them in the model. Pfeffermann et al. (1998) establish mild conditions under which if the outcomes are independent under the population model, they are also 'asymptotically independent' under the sample model when increasing the population size but holding the sample size fixed. Pfeffermann and Sverchkov (2003) discuss alternative likelihood-based approaches of estimating the parameters underlying the model (1).

Returning to the issue of different forms of weighting when fitting regression models, the use of the sample model suggests a third set of weights. Let the regression model be $y_i = x_i\beta + \varepsilon_i; E_\xi(\varepsilon_i) = 0; Var_\xi(\varepsilon_i) = \sigma_i^2, E_\xi(\varepsilon_i\varepsilon_j) = 0$ for $i \neq j$. Under this model, $\beta = \arg \min_{\tilde{\beta}} E_\xi[(\frac{y_i - x_i'\tilde{\beta}}{\sigma_i})|x_i]^2 = \arg \min_{\tilde{\beta}} E_s[w_i \frac{(y_i - x_i'\tilde{\beta})^2}{E_s(w_i|x_i)\sigma_i^2}|x_i]$. Estimating the external expectation in the right hand side by the corresponding sample mean yields the estimator $\hat{\beta}_q = [\sum_{i\in s} q_i x_i x_i'/\sigma_i^2]^{-1} \sum_{i\in s} q_i x_i y_i/\sigma_i^2$, where $q_i = w_i/E_s(w_i \mid x_i)$. The weights $\{q_i\}$ account for the net sampling effects on the conditional target model $f_\xi(y_i \mid x_i)$, and the estimator $\hat{\beta}_q$ is therefore less variable than the probability weighted estimator $\hat{\beta}_w = [\sum_{i\in s} w_i x_i x_i'/\sigma_i^2]^{-1} \sum_{i\in s} w_i x_i y_i/\sigma_i^2$, which uses the sampling weights w_i. As noted before, the estimator $\hat{\beta}_w$ is obtained as the optimal estimator in the least square sense when dropping the conditioning on x_i in the definition of β. As with $\hat{\beta}_w$, under mild conditions $\hat{\beta}_q$ is design consistent for the census vector $B = [\sum_{k=1}^N x_k x_k'/\sigma_k^2]^{-1} \sum_{k=1}^N (x_k y_k/\sigma_k^2)$.

How can the sample model be used for predicting finite population means? For this, one needs to use the sample-complement model defined as,

$$f_c(y_i \mid x_i) \overset{def}{=} f(y_i \mid x_i, i \notin s) = \frac{Pr(i \notin s \mid y_i, x_i) f_\xi(y_i \mid x_i)}{Pr(i \notin s \mid x_i)}$$
$$= \frac{E_s[(w_i - 1) \mid y_i, x_i] f_s(y_i \mid x_i)}{E_s[(w_i - 1) \mid x_i]}, \quad (2)$$

with the second equality shown in Sverchkov and Pfeffermann (2004). Note that the sample-complement model is again a function of the sample model and the expectation $E_s(w_i \mid y_i, x_i)$, and thus can be estimated from the sample data. The optimal predictor of the population total under a quadratic loss function is now,

$$\hat{T} = \sum_{i \in s} y_i + \sum_{j \notin s} E(y_j \mid j \notin s) = \sum_{i \in s} y_i + \sum_{j \notin s} E_c(y_j \mid x_j)$$
$$= \sum_{i \in s} y_i + \sum_{j \notin s} \frac{E_s[(w_j - 1) y_j \mid x_j]}{E_s[(w_j - 1) \mid x_j]}. \quad (3)$$

The last equality follows from (2), with the sample expectations in the numerator and the denominator either being modelled based on the sample data or simply estimated by the corresponding sample means. As shown in Sverchkov and Pfeffermann (2004), familiar estimators of finite population means are obtained as special cases of this theory. For example, consider the case of no covariates ($x_i = 1$). Then, by (3), $\hat{T} = \sum_{i \in s} y_i + (N - n) \hat{E}_s[\frac{w_j - 1}{\hat{E}_s(w_j) - 1} y_j]$. Estimating the two sample expectations by the respective sample means yields the estimator, $\hat{T}_m = \sum_{i \in s} y_i + \frac{(N-n)}{\sum_{i \in s}(w_i - 1)} \sum_{i \in s}(w_i - 1) y_i$. Interesting enough, using this estimator in Basu's elephants example (Example 2 in the present paper) yields the estimator $50 \times y$ where y is the weight of the selected elephant, and in particular, the estimator $50 \times y_{jambo}$ when jambo is selected!!

I conclude my discussion by commenting on Little's strong opinion that "one should choose robust models that yield design-consistent estimators, that is, to correct the model rather than to correct the estimator." This proposition is not new but I think that one needs to be cautious in its application. I am familiar with the famous saying that "no model is correct but some are useful", and I obviously agree that one should try using robust models, but the question is what is meant by a 'corrected' model, keeping in mind that the randomization distribution under which the predictor is expected to be consistent does not constitute an alternative plausible model. If the idea is to use a less stringent model, for example, allowing for different expectations in different strata instead of assuming a common expectation, or assume a polynomial expectation with an intercept instead of a simple regression through the origin, such that the extended model is basically still correct, then I can see the merit of this approach. But if correcting the model implies, for example, changing the distribution of the error terms, then I start worrying because other than predicting the population quantity of interest, one has to produce also an estimator for the variance, and possibly also set up a confidence interval (credibility interval under the Bayesian approach). I presume that these are supposed to be computed under the corrected model as well. Are we guaranteed that they are sufficiently accurate under this model? Do we need to robustify them separately? I hope that Little can shed some more light on this issue in his rejoinder, if there is one.

Let me finish with what I started. I truly enjoyed reading this article and I hope that it will generate further discussion and possibly new research on this very important topic.

ADDITIONAL REFERENCES

Pfeffermann, D., Krieger, A. M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica* 8 1087 - 1114.

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya* **61** 166 - 186.

Pfeffermann, D. and Sverchkov, M. (2003). Fitting generalized linear models under informative probability sampling. In, Analysis of survey Data, Eds. C. Skinner and R. Chambers, New York : Wiley, 175 - 195.

Sverchkov, M. and Pfeffermann, D. (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology* **30** 79 - 92.

DISCUSSION

J. N. K. RAO
Department of Mathematics and Statistics
Carleton University, Ottawa, Canada

It is a great pleasure to discuss this thought-provoking paper by R. J. Little (henceforth RJL) on inferential issues in sample surveys. RJL has made seminal contributions to missing data problems and his Wiley book with Don Rubin on this topic is widely used. More recently, RJL has been working on the role of survey design weights from a model-based Bayesian perspective and coming up with impressive solutions to practical problems. This paper illustrates his approach through a series of practically relevant examples. I will attempt to provide some comments based on my own views on inference from survey data.

It is indeed very nice of RJL to recognize the seminal contributions of some Indian samplers. I would like to add to this partial list the following names: P. V. Sukhatme, G. R. Seth and R. D. Narain from the Indian Agricultural Research Statistics Institute (IARSI) and D. B. Lahiri, M. N. Murthy and Des Raj from the Indian Statistical Institute (ISI). Narain (1951) developed designs with inclusion probabilities π_i proportional to size measures x_i that are approximately proportional to the values y_i of a variable of interest and independently proposed the now well-known Horvitz-Thompson (HT) estimator of the total Y based on the design weights $d_i = \pi_i^{-1}$ (Horvitz and Thompson 1952) for this particular design. Unfortunately, Narain's paper has been overlooked outside India and I have earlier suggested calling this estimator as NHT (Rao 1999) to also recognize the seminal 1951 paper of Narain.

In his discussion of Hansen et al (1983), RJL classified survey samplers as D, E or I according to whether models are used in design, estimation or inference. In their response, Hansen et al noted that the class D is essentially a null set and that no serious practicing sampling statistician belongs in class D. As noted by them, most of the samplers advocating design-based inference are DE modelers, either implicitly or explicitly, taking account of practical considerations, and they make repeated sampling inferences in the spirit of Neyman (1934), in particular for variance estimation and confidence intervals for large samples based on the normal approximation. The currently popular model-assisted approach to inference uses explicit "working" models for the choice of efficient design-consistent estimators and then the repeated sampling set up for asymptotically valid inferences regardless of the validity of the assumed working model. On the other hand, RJL seems to be a DEI modeler and likes parametric Bayesian inferences based on models that are possibly calibrated in the sense of having good design-based properties in repeating sampling under the specified design. RJL gives several examples assuming normality to illustrate his ideas but it is not clear to me how one ensures that the posterior inferences are well calibrated in complex situations without explicitly introducing the design effect associated with the estimator, in particular the joint inclusion probabilities π_{ij}. Formal asymptotic design-based properties for the proposed posterior inferences are not provided to validate the claims, although some simulation studies with cleverly constructed re-sampling variance estimators seem to perform well in repeated sampling under a probability proportional to size (PPS) single stage design with small sampling fraction (Zheng and Little 2005). Proposed methods will be attractive to users if they indeed lead to more efficient design-valid large sample inferences than customary design-based methods, say those based on the model-assisted approach. For small samples, RJL claims that he can provide exact Bayesian credible intervals conditional on the data, but such inferences may be sensitive to distributional assumptions and possibly to the choice of prior.

RJL proposes to handle the problem of sample selection bias in model-based inference by including all the design variables among the predictor variables in the model. However, this goal may not be of easy to achieve in practice because not all design variables used in the sample selection may be known or accessible to the user. As noted by Pfeffermann and Sverchkov (2007), adding the survey weights to the model as surrogates for the design variables may not summarize the design variables adequately and also not operational if the weights are not available for the non-sampled units. Further, the analyst may be interested in making inferences on model parameters of a specified population model, such as regression parameters. In this case, the parameters of the expanded model that includes the design variables may not be interest to the analyst. On the other hand, design-based approach can be applied in a routine manner for inference on the parameters of the original model that is of interest to the analyst.

Godambe's (1955) famous result that the best estimator of the population mean in a general class of design-unbiased estimators does not exist attracted a lot of attention from theoreticians and also created the misconception that samplers insist on design unbiasedness. For example, Basu (1971) commented that "surveyors got mixed up with the idea of unequal probability sampling" to make the estimator based on the mean of ratios y_i/x_i look good by eliminating its design bias. On the contrary, design unbiased estimation is not insisted upon in practice because it "often results in much larger MSE than necessary" (Hansen et al, 1983). Instead, design consistency is deemed necessary in large samples and attention is paid to reducing the MSE. Hansen et al (1983) discuss the role of probability proportional to size (PPS) sampling in multi-stage cluster sampling and explain why it is widely used in practice. I am glad that RJL recognizes the importance of design consistency in large samples for his methods and also advocates the use of a calibrated Bayes approach.

Example 1 (stratified sampling)

This example deals with the estimation of a population mean from a stratified random sample. RJL first considered the normal model (2) with a common mean to show that the posterior mean does not reduce to the design-consistent stratified mean. But it appears to me that this example is not appropriate because model (2) does not hold for the sample under the stratified design and hence the posterior mean under (2) is not valid. RJL then introduced model (5) with separate strata means and the design is not informative in this case so that the posterior mean which agrees with the stratified mean is valid. This example is fine when there is no auxiliary information. But suppose that we have an auxiliary variable x observed on all the population units and the number of strata is large and the sample sizes within strata are small, as in the example of Hansen et al. (1983). In this case, the use of models with separate strata parameters (e.g. a ratio model with different slopes) leads to separate ratio or regression estimators which are not design consistent unless the within strata sample sizes are large. On the other hand, a model-assisted approach assuming a working model with common parameters across strata but using design weights to make the estimator design consistent leads to widely used combined ratio or regression estimators (also called GREG estimators) which perform well even with small strata sample sizes as long as the over all sample size is large (e.g. when the number of strata is large as in many business surveys). Note that the design is informative with respect to the working model but still the model-assisted approach provides asymptotically valid design inferences. It is possible to construct model-assisted "optimal" combined regression estimators which are asymptotically more efficient than the GREG estimators and also lead to more appealing conditional design-based inferences where the reference set is a set of samples under the design that are relevant to the sample at hand (see e.g. Rao 1999, section 3.4). The conditional design-based approach addresses the criticism of modelers that unconditional design-based inferences are scientifically less relevant than model-based inferences conditional on the observed sample.

Example 2 (Basu's circus elephants)

Godambe's result on the non-existence of best linear unbiased estimator prompted some researchers to advance other criteria for the choice of estimator. One such criterion is admissibility and many papers on admissibility appeared in prestigious journals, but unfortunately it is not sufficiently selective. As a result, the so-called hyper admissibility criterion was advanced to show that the NHT estimator is "optimal" under *any* design according to this criterion. This led to the famous Basu's circus elephants example in which a "bad" design with π_i unrelated to y_i was constructed to demonstrate the absurdity of the NHT estimator under that design which in turn prompted the well-known Bayesian Dennis Lindley to conclude that this counter example "destroys frequentist sample survey theory" (Lindley 1996). RJL is very fond of citing this example which is discussed in example 2 of the paper and he says that "Design-based statisticians groan when modelers bring up Basu's example". In a 1969 Technical Report of the Indian Statistical Institute (Rao and Singh 1969) I showed why the hyper-admissibility criterion, which requires admissibility for all possible subpopulations (domains) including those with only one member, makes no practical sense. Essentially, the NHT estimator is the best linear unbiased estimator in domains of size one and hence all other candidates become inadmissible in those domains and thus eliminated from competition. In practice, we are seldom interested in all subpopulations, and certainly not in subpopulations of size 1. Basu (1971) also made a similar observation.

I encountered practical situations with π_i unrelated to y_i long before Basu (1971) appeared, but my work (Rao 1966) on how to get around the difficulty with the NHT estimator in such situations has been overlooked by Basu, RJL and other Bayesian modelers. In my

1966 paper I showed that the NHT estimator is highly inefficient when compared to $N\bar{y}$ when the size measure π_i is unrelated to y_i and recommended using $N\bar{y}$ as the estimator of total in such cases, where $\bar{y}$ is the un-weighted sample mean. Interestingly, the un-weighted estimator $N\bar{y}$ is also design-unbiased when π_i is not related to y_i. Scott and Smith (1968) showed that $N\bar{y}$ is in fact the best estimator in the wider class of design-model unbiased estimators (that includes NHT estimator and $N\bar{y}$), assuming the model used in Rao (1966) that reflects the knowledge that y_i and π_i are unrelated. PPS sampling in multi-purpose surveys is typically designed to ensure that the size measure is strongly related to main variables of interest but it is possible that the same size measure is unrelated or weakly related to some other variables of interest. For example, in the Iowa Farm Survey that led to my 1966 paper, the farm size was strongly related to area under corn but it was also unrelated to poultry count. The solution I proposed uses varying weights across variables of interest, but often the user is interested in using a common weight. The approach of RJL also has a similar limitation. Beaumont (2008) studied this problem and developed a common smoothed weight under the design-based framework and moderate size samples, and the resulting estimator performed well across variables that are strongly related or moderately related or weakly related to the size measure. I hope the above comments will convince RJL that design-based samplers were aware that "no single estimator is optimal in all situations, and weighted estimators can do very badly ..." long before Basu (1971) appeared. This is also evident from the following observations of Hansen et al (1983) : "Unless reasonably good measures are available to determine the varying probabilities, substantial variance increases rather than decreases may result from their use ..."

Examples 3 and 4 (post-stratification)

RJL notes some difficulties with the standard design-based post-stratified estimator in the context of simple random sampling. Because of random sample sizes n_j within post strata, the n_j in one or more post-strata may become very small or even zero when the overall sample size n is small. This problem can also occur in larger samples if the number of post-strata is also large, as in the cross-classification of two or more post-stratification variables. Design-based samplers have approached the latter problem through calibration only to marginal post-strata population counts (Deville and Sarndal 1992); in many practical situations, only marginal population counts may be accurately known from demographic projections of census counts. Moreover, the calibration approach can handle complex sampling designs, and avoids the difficulty of searching for suitable design-specific models using the RJL approach. It is now widely used in the production of official statistics because it uses a common weight across variables and it ensures calibration to user-specified population totals of auxiliary variables. However, it should be noted that calibration estimation is not necessarily model-assisted and it can lead to poor coverage performance of confidence intervals even in moderate size samples for highly skewed auxiliary variables when the model implied by the calibration constraints provides a poor fit to the data (Rao et al 2003). On the other hand, the model-assisted approach with a working model that accounts for major model misspecifications performs well in terms of coverage performance; for example, when the true model is clearly quadratic in x and the model implied by the calibration to the population size N and the total of x is linear. Convergence to normality depends on the skewness of the residuals from the assumed model and the residuals from the linear regression remain highly skewed because of the omitted quadratic term unlike the residuals under the quadratic model.

RJL recommends the use of random effect models to borrow strength across post strata and thus obtain more efficient estimators for post-strata with small or zero sample counts. Lazzeroni and Little (1998) used such models in the case of ordinal post strata, and conducted a simulation study to compare the design efficiency of their model-based estimator to an estimator based on an ad hoc collapsed post-strata approach. Alternative design-based estimators are also available (Rao 1985) but not included in this study. Simulation study showed that modest efficiency gains may be achieved by using the proposed approach when estimating the population mean. However, it may not be easy to formulate a realistic random effects model in the absence of auxiliary information about the post-strata; assuming exchangeable random effects model may not be realistic in such cases because the post-strata means are known to be not homogeneous. In the context of two-way stratification designs with the total sample size smaller than the total number of strata, I have suggested the use of random effects models when feasible (Rao 1985).

Scott and Smith (1969) were the first authors to propose random effect models in the context of two-stage cluster sampling. They obtained a Bayes predictor of the population mean as a weighted average of the customary ratio estimator (appropriate under random sampling of the clusters) and the mean of ratios estimator (appropriate under PPS sampling of the clusters). Bellhouse and Rao (1986) conducted a simulation study on the efficiency of the Bayes predictor and found that the gain in efficiency over the customary strategies is minimal if any. On the other hand, big gains in efficiency can occur by borrowing strength using random effects models when the parameters of interest are the clusters (small areas) themselves or the post strata in the RJL example. In such cases, traditional design-based approach is inefficient or not feasible (as in the case of non-sampled clusters or post strata with zero sample counts). Small area estimation using random effects models has attracted a lot of attention in recent years due to growing demand for reliable small area statistics. Rao (2003) gives a detailed account of model-based small area methods.

Example 5 (PPS sampling and spline regression)

In example 5 RJL gives a brief account of his important work on using flexible penalized spline regression models in conjunction with PPS sampling. The NHT estimator is efficient under the regression through the origin model (13) with error variance proportional to x^2, called the HT model, but it can perform poorly when this model is seriously violated. Spline regression models make minimal assumptions on the regression function (assuming it accounts for all the relevant auxiliary variables) and Zheng and Little (2005) showed through simulations that predictive inferences based on the spline model perform better than the design inferences using the NHT estimator when the HT model is violated, while sacrificing little in terms of efficiency when the HT model holds, even when the prediction estimator may not be design consistent. It appears that the spline model approach may hold some promise for practical applications. In his concluding remarks, RJL alludes to the possibility of using spline models for "routine applications to standard sampling designs", although he warns against black box type applications. Note that the spline model requires the specification of number of knots and location of the knots in addition to the choice of predictor variables.

Breidt et al (2005) studied model-assisted design inference under a spline model and stratified random sampling. Their simulation results indicated that the model-assisted estimator can be considerably more efficient than the prediction estimator under stratified sampling. A possible reason for the inefficiency of the prediction estimator here could be that it is not design consistent under the spline model of Breidt et al while the model-assisted estimator remains design consistent. Hence, it appears that the use of a prediction estimator under a spline model may require specification of a model that ensures design consistency. In the Breidt et al simulation study strata indicators are not included in the model so that the design is informative. As noted earlier, model-assisted estimator is design consistent even when the design is informative with respect to the working model.

Example 6 (unit and item non-response)

Finally, I would like to make a few comments on the seminal work of RJL on making inference in the presence of unit and item non-response. RJL makes an important point on adjustment cell weighting under unit non-response. He notes that the formation of adjustment cells based on a good predictor of the outcome is more important than using a good predictor of unit non-response if the latter predictor is not related to the outcome. A drawback of the adjustment based on a predictor of the outcome is that it leads to varying weights across the variables in a multi-purpose survey, unlike the adjustment based on a predictor of unit non-response. However, adjustment based on predictor of outcome might be practical if one variable is the most important variable and the resulting cells are used for all the variables, leading to a common weight across variables. In fact, using both the predictor of outcome and the predictor of unit non-response could lead to reduced bias and variance (Smith et al 2004). Vartivarian and Little (2002) reported favorable results from cross-classifying on the predictor of non-response and the predictor of outcome to form adjustment cells.

Imputation for item non-response has attracted a lot of attention and RJL advocates the use of Rubin's multiple imputation (MI) variance estimator, based on multiple imputed data sets. Although the MI approach may be attractive as a "black box" approach to estimation and analysis of survey data from public-use completed data sets, there are a number of theoretical difficulties with this approach especially in the context of complex survey data involving dependent data structures or low response rates (see e.g. Fay 1996; Wang and Robins 1998), some times leading to inefficient inferences or asymptotically not valid inferences, especially with small number of imputed data sets (say 2 to

5). Rubin (2003) defended his methods by saying that even complete data survey practice can also some times go wrong and referring to my paper (Rao et al 2003) where I have shown, as mentioned before, that a standard linear regression estimator can perform poorly in terms of coverage rates when the working model is strongly quadratic and the predictor variable is highly skewed. However, I have also shown in that paper that a model-assisted approach with a working model that accounts for major misspecifications performs well in terms of coverage performance. Similarly, it is possible to develop alternative approaches (even under commonly used single imputation) that are asymptotically valid under general sampling designs, at least for inference on descriptive parameters such as population totals and domain totals which are of primary interest to statistical agencies. In a recent paper, Kim and Rao (2009) have developed a unified approach to linearization variance estimation for population totals and domain totals that leads to asymptotically valid inferences under a missing at random (MAR) imputation model and different imputation methods including Rubin's MI.

Finally, I congratulate RJL on his outstanding contributions to survey sampling theory and practice, especially to the important topic of inference from missing data, and for his thought provoking and stimulating paper on calibrated Bayes prediction approach to inference from survey data.

ADDITIONAL REFERENCES

Beaumont, J. F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika* **95** 539 - 553.

Bellhouse, D. R. and Rao, J. N. K. (1986). On the efficiency of prediction estimators in two stage sampling. *J. Statist. Plann. and Infe.* **13** 269 - 281.

Breidt, F. J., Claeskens, G. and Opsomer, J. D. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika* **92** 831 - 846.

Deville, J. C. and Sarndal, C. E. (1992). Calibration estimation in survey sampling. *J. American Statist. Assoc.* **87** 376 - 382.

Godambe, V. P. (1955). A unified theory of sampling from finite populations. *J. Royal Statist. Soc.* Series B **17** 269 - 278.

Kim, J. K. and Rao, J. N. K. (2009). Unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika* **96** (in press).

Lindley, D. V. (1966). Letter to the editor. *American Statistician* **50** 197.

Narain, R. D. (1951). On sampling without replacement with varying probabilities. *J. Indian Soc. Agri. Statist.* **3** 169 - 174.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J. Royal Statist. Soc.* **97** 558 - 606.

Rao, J. N. K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhya* Series A **28** 47 - 60.

Rao, J. N. K. and Singh, M. P. (1969). On the criterion of 'hyper-admissibility' and 'necessary bestness' in survey sampling. Technical Report No. 40, Research and Training School, Indian Statistical Institute.

Rao, J. N. K. (1985). Conditional inference in survey sampling. *Survey Methodology* **11** 15 - 31.

Rao, J. N. K. (1999). Some current trends in sample survey theory and methods. *Sankhya* Series B **61** 1 - 57.

Rao, J. N. K. , Jocelyn, W. and Hidiroglou, M. A. (2003). Confidence interval coverage performance for regression estimators in uniphase and two-phase sampling. *J. Official Statist.* **19** 17 - 30.

Rubin, D. B. (2003). Discussion on multiple imputation. *International Statist. Review* **71** 619 - 625.

Smith, P. J., Hoaglin, D. C., Rao, J. N. K., Battaglia, M. P. and Daniels, D. (2004). Evaluation of adjustments for partial nonresponse bias in the US National Immunization Survey. *J. Royal Statist. Soc.* Series A **167** 141 - 156.

Vartivarian, S. L. and Little, R. J. L. (2002). On the formation of weighting adjustment cells for unit nonresponse. In Proceedings A, *American Statist. Assoc.* pp. 3353 - 33358.

Wang, N. and Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85** 935 - 948.

DISCUSSION

D. B. RUBIN
Department of Statistics,
Harvard University, USA

I congratulate Roderick Little on contributing this eloquently written review of perspectives on using weights when drawing inferences in sample surveys. Because Rod and I have worked together for over three decades, our views on many topics are compatible. There are, however, a few points that I would add in this brief discussion. These points are, I believe, really ones of emphasis and are not in conflict with views expressed in the target article.

1. Exchangeability of the Bayesian Model on the Data

The first point concerns the exchangeability of units in Bayesian inference for sample surveys. Ericson (1969) proposed that the exchangeability of units, that is, the row exchangeability of the model for the N (units) by d (variables) population data matrix, Y, is somehow related to the choice to use simple random sampling : "I believe that the notion of exchangeability and exchangeable prior distributions very closely approximates the real opinions of thoughtful 'classical' practitioners in many situations where they deem simple random sampling to be appropriate." (Ericson, 1969, p. 198).

In contrast, in Rubin (1976, 1978, 1979, 1983), I distinguished between the model for the population data, $P(Y)$, and the model for the process that creates observed and missing values in Y, which in the sampling survey context is the sampling mechanism, $P(I \mid Y)$, where I is the $N \times d$ indicator matrix for which values in Y are included in the drawn sample (here, I ignore complications such as unit and item nonresponse). As stated in Rubin (1978, 1983, 1987), the row (unit) exchangeability of $P(Y)$ follows from putting all possibly relevant information in (Y, I), so that the labeling of the rows of (Y, I) is a random permutation of $1, \ldots, N$, thereby implying that $P(Y)$ must be row exchangeable, no matter what sampling mechanism is used. Simple random sampling implies $P(I \mid Y) = P(I)$, but it implies nothing about the exchangeability of units. I believe it is often important to distinguish between those modeling assumptions that are justified by physical actions such as random sampling, stratification, cluster sampling, etc. and those that are a consequence of mathematical formulations.

2. The Complementary Roles of Bayesian Inference and Repeated-Sampling Evaluations

The second point concerns the complementary roles for the Bayesian approach, which derives inferences directly, and the repeated-sampling approach, which evaluates procedures over the randomization distribution induced by the assignment mechanism (e.g., bias of point estimates, confidence coverage of interval estimates). Such evaluations can be extremely illuminating, especially when they are restricted to distributions for Y that are plausible in particular applications. These evaluations of the operating characteristics of procedures can be used with any procedure, no matter how created (e.g., created using the Bayesian paradigm, via an asymptotic frequentist argument, from a dream I had), and they are important to apply to Bayesianly-derived procedures because essentially all models on Y are only approximations to reality.

For example, the Bayesianly-derived interval estimate for a population mean under normality has excellent repeated-sampling coverage in many situations, and it is the standard in the design-based approach, although it can also be motivated from a robust Bayesian perspective (e.g., see Pratt, 1965). Also see Rod's first example for the stratified random sample case. For another example, one of six chapters in Rubin (1987, 2004) is devoted to randomization-based evaluations of Bayesianly-derived multiple imputation procedures. This text also provides other examples of procedures that are usually justified from their repeated sampling evaluations, but the text derives them from the Bayesian perspective; see examples 2.3 and 2.4, and problems 2.11, 2.12. and 2.13 for ratio, pps, and regression estimators, respectively. The evaluations of such Bayesianly-derived procedures are the path to being a calibrated Bayesian, as Rod states.

However, contorting these design-based evaluations to be principles for creating procedures (e.g., unbiased estimation) just does not work in any generality, as illustrated in Rod's second example, Basu's elephant.

For creating statistical procedures, one really needs the Bayesian approach, which supplements the model for the assignment mechanism, which is all that is essential for the evaluations, with a model on the population data, thereby allowing the direct prediction of the missing values in Y from the observed values in Y and the observed values of I.

3. An Analogy with "Word Problems" and the Calculus of Algebra

An informal analogy that I have often used to dramatize the complementary roles of the design-based and Bayesian approaches involves an intelligent child solving a word problem such as the following one. Anne is 22 years younger than her mother; her mother is 36 years old; how old is Anne? The intelligent child will answer this question without trouble: Anne is 14 years old. This problem is roughly analogous to estimating the population mean from a simple random sample. But now consider the following word problem : Anne is 22 years younger than her mother; Anne's father is 2 years older than her mother; her mother's age plus her father's age is ten years less than six times Anne's age; how old is Anne? This problem is roughly analogous to handling nonresponse in an multistage survey. The intelligent child will probably struggle with this question until learning the calculus of translating the word problem into algebra (i.e., symbols) and correctly manipulating the resulting equations. Now to make sure Anne's age is correctly found, it is always wise to "plug" the purported answer back into the separate sentences of the word problem to ensure its correctness.

The calculus of algebra is analogous to the Bayesian approach, which creates answers in both easy and difficult problems based on explicit assumptions and principles of inference, whereas plugging answers into equations to evaluate their correctness is analogous to the frequentist's repeated-sampling evaluations. Both activities, algebra and plugging in, are useful for correctly solving word problems. And both statistical approaches, Bayesian derivations and frequentist evaluations, are needed for obtaining good answers when confronting complex survey problems.

4. The Future and the Calculus of Modern Bayesian Computation

The range of procedures currently available to the Bayesian survey methodologist goes far beyond the set of procedures that are available using closed-form algebra. Rod's work using penalized spline methods for predicting non-sampled values (Zheng and Little, 2003, 2004, 2005) nicely illustrates this approach. The EM algorithm (Dempster, Laird, and Rubin, 1977) and its relatives, including the fully Bayesian Data Augmentation (Tanner and Wong, 1987) and other Markov Chain Monte Carlo algorithms (e.g., see Gelman et al. 2003, Carlin and Lewis, 2000), or other computationally intensive techniques (e.g., SIR, Rubin, 1987) have allowed the use of complicated explicit and implicit models that were difficult to imagine using in practice a few decades ago. Having such extremely flexible modeling tools available is a tremendous advantage because the collection of Bayesian models that are appropriate in complex survey settings and that have closed-form answers are really so limited that they often could produce answers that could easily be viewed as too non-robust for general survey practice.

This past limitation of the Bayesian approach is the primary reason, I believe, for the traditional dominance of the design-based approach even when deriving procedures in survey practice. In the future, there should be a more complementary and balanced use of both perspectives in survey practice, which is nicely reflected in this target article by Roderick Little.

ADDITIONAL REFERENCES

Carlin, B. P. and Louis, T. A. (2000), Bayes and Empirical Bayes Methods for Data Analysis, Second Edition, London : Chapman & Hall.

Dempster, A. P., Laird, N., Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *J. Royal Statist. Soc.* Series B : **39(1)** 138. With discussion and reply.

Ericson, W. A. (1969). Subjective Bayesian Models in Sampling Finite Populations. *J. Royal Statist. Soc.* Series B : **31(2)** 195 - 233.

Gelman, A. Carlin, J., Rubin, D. B. and Stern, H. (2003). Bayesian Data Analysis, 2nd ed. New York : CRC Press.

Little, R. J. A. (2009). (commented article) Weighting and prediction in sample surveys.

Pratt, J. W. (1965). Bayesian Interpretation of Standard Inference Statements. *J. Royal Statist. Soc.* Series B : **27** 169 - 203.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika* **63** 581 - 592. With discussion and reply

Rubin, D. B. (1978). Bayesian Inference for Causal Effects : The Role of Randomization. *Annals Statist.* **6** 34 - 58.

Rubin, D. B.,(1983). Conceptual Issues in the Presence of Nonresponse. Incomplete Data in Sample Surveys (Volume 2): Theory and Bibliography. New York : Academic Press, Inc., pp. 123 - 142.

Rubin, D. B. (1987a). Multiple Imputation for Nonresponse in Surveys. New York : John Wiley and Sons.

Rubin, D. B. (2004). Multiple Imputation for Nonresponse in Surveys. Reprinted with appendices as a "Wiley Classic." New York : John Wiley and Sons. Appendix 1 : (1977) "The Design of a General and Flexible System for Handling Non Response in Sample Surveys". Appendix 2 : (1983) "Progress Report on Project For Multiple Imputation of 1980 Codes". Manuscript distributed to the U.S. Bureau of the Census, the U.S. National Science Foundation and the Social Science Research Foundation.

Rubin, D. B. (1987b). A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information are Modest: The SIR Algorithm." Discussion of "The Calculation of Posterior Distributions by Data Augmentation" by Tanner and Wong. *J. American Statist. Assoc.* **82** 543 - 546.

Tanner M. A. and Wong W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *J. American Statist. Assoc.* **82** 543 - 546.

Zheng, H. and Little, R. J. (2003). Penalized Spline Model-Based Estimation of the Finite Population Total from Probability-Proportional-To-Size Samples. *J. Official Statist.* **19** 2, 99 - 117.

Zheng, H. and Little, R. J. (2004). Penalized Spline Nonparametric Mixed Models for Inference about a Finite Population Mean from Two-Stage Samples. *Survey Methodology* **30** 2, 209 - 218.

Zheng, H. and Little, R. J. (2005). Inference for the Population Total from Probability- Proportional-to-Size Samples Based on Predictions from a Penalized Spline Nonparametric Model. *J. Official Statist.* **21** 1 - 20.

REJOINDER

RODERICK J. LITTLE
University of Michigan, USA

I am honoured to have such a distinguished set of discussants, and appreciate their kind remarks. Since I have learnt so much from Don Rubin over the years, it is perhaps not surprising that I generally agree with his points. The article by Ericson (1969) was, I think, very important because it emphasized that Bayes provides a solution to the finite population sampling problem, so no new principles of inference were needed. However that article focused on simple random sampling, for which the distribution of I and Y are independent, and hence did not consider complex designs. I think Rubin's explicit formulation in terms of the joint distribution of I and Y was important in allowing design information to be included, and it makes the argument for randomization more explicit, since other forms of sampling may create hidden dependences between I and Y that are assumed away at one's peril. As a young statistician struggling with the complexities of survey inference debated in the 1980's, Rubin's formulation clarified for me the appropriate Bayesian treatment of inference for complex sampling designs. I found his Bayesian approach refreshingly simple - see for example, Little and Rubin (1983).

I appreciate Don's comments about exchangeability. One criticism of Bayes is that full probability modelling is too much work, and including "all the possibly relevant information in Y" to achieve exchangeability greatly expands the modelling task. I think one advantage of random sampling is that it allows us to get by with simplified models that do not include all the possibly relevant information in Y. For example, suppose a simple random sample of students is sampled from a population of students, and some sampled students are in the same school. The model that ignores school information and assume the students are exchangeable violates the fact that characteristics of students in the same school are correlated. Nevertheless, inference assuming exchangeability of students is still reasonable, since the random sampling renders the inference insensitive to violation of that assumption. Random sampling of students allows us to get away with ignoring the school information, despite the clustering within schools. This would clearly not be the case if a two-stage sample was selected with the schools as primary sampling units.

I find little to argue with in Andrew Gelman's discussion. In the context of pps sampling, Gelman mentions the case of two-stage cluster sampling, in which clusters are sampled pps and then a sample is drawn from each cluster with probability inversely proportional to size. In this case, the classical unit weights are all equal, and Gelman questions whether the model-based spline approach improves much on the unweighted mean. Simulations in Zheng and Little (2004) suggest that considerable gains are in fact possible in this setting too. I admit to some surprise, since like Gelman my intuition suggested that gains would be minor.

Gelman's analogy between weighting and multiple imputation is intriguing. Various model-based estimation approaches (including least squares regression) can be considered as weighting with weights that deviate from the inverse of selection probabilities. However, I suspect Gelman would agree that prediction of unknowns is the more general and compelling principle.

My friend Danny Pfeffermann appears receptive to the Bayesian viewpoint I share with Gelman and Rubin, but his work is more expansive, and includes "conditional likelihood" approaches that do not lie strictly within the Bayesian paradigm. I think the approaches developed by Pfeffermann and colleagues are interesting, particularly from a frequentist perspective, but they run counter to the unified Bayesian approach espoused in my article.

Pfeffermann's comments on different forms of conditioning illustrate his essentially frequentist orientation. An attraction of the Bayesian paradigm is that conditioning is transparent and unambiguous - posterior distributions condition on all the data. In particular, the choice between design and variance weights in my initial example in *not* for me a "matter of conditioning", but rather a matter of the choice of model. In that example, consider the (ill-advised) model

$$(Y_i \mid Z_i, I_i) \sim N(\mu, \sigma^2/u_i); I_i \mid Z_i \sim pps(Z_i), \quad (1)$$

where Z_i denotes size for unit $i, pps(Z_i)$ denotes some form of pps sampling, and u_i are known constants modelling heteroscedasticity. If the sampling fraction is small, Bayes inference under (1) leads to the variance-weighted mean $\sum_{i\in s} Y_i u_i / \sum_{i\in s} u_i$, not the mean that weights by the product weight $Z_i u_i$ as in Pfeffermann and Sverchkov's conditional likelihood approach. The problem with this Bayesian analysis is that the model (1) assumes Y does not depend on Z, and inferences are not robust to violations of this assumption. Little (2004, Example 11) considers Bayesian inference for a modification of (1) that leads to the product weights $\{Z_i u_i\}$, but it is the model that changes, not the nature of the conditioning or estimation principle.

Pfeffermann discusses the case of pps sampling when non-sampled values of the size Z are not available to the analyst. He states that the likelihood is not "operational" in this case; I disagree. The Bayesian approach simply requires an additional model for Z, since Z is not known for the non-sample cases. Little and Zheng (2007) present a simple Bayesian approach based on a Bayesian bootstrap model for Z. Even if there are many design variables, it is only necessary to model the single size variable Z that determind the pps sample, so this approach is not as computationally complex as Danny implies.

Concerning Pfeffermann's comments on robust modelling, I feel his concerns are addressed under the calibrated Bayes approach, as articulated in Rubin's work, including his discussion of this article; see also Little (2006). The key is to distinguish between the *inference* for the unknown quantity under a model M, and the *operating characteristics* of that inference, which are allowed to influence the choice of M. The inference under a given model M is purely Bayesian - it conditions on the observed data, including the inclusion indicators I. The repeated sampling properties with respect to I (for example, the confidence coverage of Bayesian credibility intervals computed under M) are invoked when considering the "operating characteristics" of M, and these can modify the choice of M. For example, the model (1) should in many cases be rejected since the Bayesian credibility intervals under M are subject to poor confidence coverage (are *poorly calibrated*) if the assumption of independence of Y and Z is violated, as seems likely in many applications.

Concerning variance estimation, Danny is right that computing a Bayesian estimator with a design-based estimate of the variance is not calibrated Bayes, since the inference is not entirely model-based. I confess that I have used bootstrap or jackknife variance estimation as a practical expedient in my work, since the extra work of carefully modelling the variance structure seems unlikely to be worth any gains in the quality of the inference. This is a deviation from the true calibrated Bayes path, though perhaps not a major one.

I appreciate Jon Rao's extensive and thoughtful discussion, which clearly reflects a deep appreciation of the issues. A proper rejoinder would probably require another full-length paper, so I limit myself to a few random comments. In principle I am a DEI modeler, although (as noted above) I do lapse sometimes by using a replication-based variance estimator to avoid full modelling of the variance structure. Jon noted potential sensitivity to the choice of prior distribution for small sample inferences - this clearly exists to some extent, but I note that the classical design-based approach is asymptotic and not guaranteed to yield good confidence coverage in small samples. Indeed, in the simulations I have conducted with my collaborators, the confidence coverage of the "model-assisted" approaches that Jon espouses is in fact inferior to the Bayesian approach. I have not seen evidence that the calibration step improves confidence coverage when the model yields design-consistent prediction estimators on its own. (The simulations of Breidt et al (2005) do not consider confidence coverage). Jon's discussion of the Hansen, Madow and Tepping (HMT) example is topical for me since I am currently working on the case HMT considers, and hope to report some results in the summer. I am not sure why he states "the separate ratio estimator is not design-consistent unless the within strata sample sizes are large", since my understanding of design consistency would let these sample sizes increase, given a finite number of strata. I agree that the separate ratio estimator can be unstable if the stratum sample size are small, but there are Bayesian fixes for that problem. I'll reserve comments on the other examples for future work.

Finally I'd like to thank Jon for augmenting my list of distinguished Indian statisticians who have contributed to this important topic. It gives me the opportunity to again congratulate the Calcutta Statistical Association on this jubilee, and to predict numerous additional contributions by its members in the future.

ADDITIONAL REFERENCES

Little, R. J. A. and Rubin, D. B. (1983). Discussion of "Six approaches to enumerative sampling". by K. R. W. Brewer and C. E. Sarndal. In *Incomplete Data in Sample Surveys, Vol. 3 : Proceedings of the Symposium.* W. G. Madow and I. Olkin, eds. New York : Academic Press

Little, R. J. and Zheng, H. (2007). The Bayesian Approach to the Analysis of Finite Population Surveys. *Bayesian Statistics* 8 J. M. Bernardo, M. J. Bayarri, J. O.l Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West (Eds.), 283 - 302. (with discussion and rejoinder), Oxford University Press.