Journal of Official Statistics, Vol. 28, No. 3, 2012, pp. 309–334

Calibrated Bayes, an Alternative Inferential Paradigm for Official Statistics

Roderick J. Little[1]

I characterize the prevailing philosophy of official statistics as a design/model compromise (DMC). It is design-based for descriptive inferences from large samples, and model-based for small area estimation, nonsampling errors such as nonresponse or measurement error, and some other subfields like ARIMA modeling of time series. I suggest that DMC involves a form of "inferential schizophrenia", and offer examples of the problems this creates. An alternative philosophy for survey inference is calibrated Bayes (CB), where inferences for a particular data set are Bayesian, but models are chosen to yield inferences that have good design-based properties. I argue that CB resolves DMC conflicts, and capitalizes on the strengths of both frequentist and Bayesian approaches. Features of the CB approach to surveys include the incorporation of survey design information into the model, and models with weak prior distributions that avoid strong parametric assumptions. I describe two applications to U.S. Census Bureau data.

Key words: Bayesian statistics; frequentist statistics; likelihood principle; robust models; model checking; statistical inference.

1. Introduction

The mission of official statistics is to produce relevant, timely and credible statistics about key social and economic phenomena. Statistical agencies face increased demand for data products, and the questions asked by our society are becoming increasingly complex and hard to measure. On the other hand, individuals and organizations are less willing to respond to requests for information, voluntary or not. Surveys and censuses are expensive and challenging to mount. Combining information from a variety of data sources is attractive in principle, but difficult in practice. Disseminating information for small areas is subject to the dangers from disclosure of confidential information from respondents. For these reasons, the standard statistical approach of taking a random sample of the target population and weighting the results up to the population no longer meets our needs. We should see the traditional survey as one of an array of data sources, including administrative records and other information gleaned from cyberspace. Tying this

[1] Professor, Department of Biostatistics, University of Michigan and Associate Director for Research and Methodology and Chief Scientist, Bureau of the Census. University of Michigan, School of Public Health, Dept of Biostatistics, 1420 Washington Heights, Ann Arbor, MI 48109-2029, U.S.A. Email: rlittle@umich.edu
Acknowledgment and Disclaimer: This article was supported as part of an Interagency Personnel Agreement with the U.S. Census Bureau. The views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau. I greatly appreciate the constructive comments of an associate editor and five referees on an earlier draft.

information together to yield cost-effective and reliable estimates requires modern statistical analysis tools.

In response to these challenges, the U.S. Census Bureau has recently formed a new Research and Methodology Directorate. I am its first Associate Director, and I write as the first Bayesian statistician with a senior leadership position at the Census Bureau, and as one who has great respect for the history and statistical traditions of the agency.

One of my responsibilities is to uphold statistical standards, and this role has led me to ponder the prevailing statistical philosophy of the agency, which I believe many other official statistical agencies share. I feel that some of the obstacles faced by official statistics are attributable to the ambivalence of this prevailing philosophy. I suggest that an alternative statistical philosophy, calibrated Bayes, provides a better vehicle for official statistics in the future.

2. The Prevailing Philosophy of Statistical Inference in Official Statistics

Official statistics is largely concerned with censuses and surveys, with a strong emphasis on probability sampling. There are three main competing general philosophies of inference from probability sample surveys (e.g., Little and Rubin 1983; Little 2004): (a) design or randomization-based inference, and (b) model-based inference, in its two main forms of superpopulation inference and Bayesian inference.

2.1. *Design-Based Inference*

The classical randomization or design-based approach to survey inference (e.g., Hansen et al. 1953; Kish 1965; Cochran 1977) has the following main features. For a population with N units, let $Y = (y_1, \ldots, y_N)$ where y_i is the set of survey variables for unit i, and let $I = (I_1, \ldots, I_N)$ 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 (a) 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 of I; and (b) 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 $\hat{q} \pm 1.96\sqrt{\hat{v}}$.

Models can and often do play a role in determining the choice of estimator in this approach. Specifically, regression or ratio estimates are based on implicit models, and model-assisted methods such as generalized regression (Särndal et al. 1992) incorporate model predictions. However, these methods are still fundamentally design-based, since the distribution of I remains the basis for inference.

2.2. *Model-Based Inference*

The model-based approach bases inference on a model for the distribution for Y, perhaps combined with the distribution of I. Initial model formulations did not overtly assign a distribution for I, but modeling both Y and I allows assumptions about the method of selection to be formalized, and clarifies the value of probability sampling. The model is

used to predict the non-sampled values of the population, and hence finite population quantities Q. There are two major variants: superpopulation modeling and Bayesian modeling.

In superpopulation modeling (e.g., Royall 1970; Thompson 1988; Valliant et al. 2000), the population values of Y are assumed to be a random sample from a "superpopulation", and assigned a probability distribution $p(Y|Z,\theta)$ indexed by fixed parameters θ, and conditioned on known design variables Z.

Bayesian survey inference (Ericson 1969, 1988; Basu 1971; Scott 1977; Binder 1982; Rubin 1983, 1987; Ghosh and Meeden 1997; Sedransk 2008; Little 2003, 2004; Fienberg 2011) requires the specification of a prior distribution $p(Y|Z)$ for the population values. Inferences for finite population quantities $Q(Y)$ are then based on the posterior predictive distribution $p(Y_{\text{exc}}|Y_{\text{inc}}, Z, I)$ of the non-sampled values (say Y_{exc}) of Y, given Z and the sampled values Y_{inc}. Probability sampling allows us to "ignore" the distribution of the sample inclusion indicator I in this model, and base inferences on posterior predictive distribution $p(Y_{\text{exc}}|Y_{\text{inc}}, Z)$, simplifying the modeling task. The specification of the prior distribution $p(Y|Z)$ is often achieved via a parametric model $p(Y|Z, \theta)$ indexed by parameters θ, combined with a prior distribution $p(\theta|Z)$ for θ, that is:

$$p(Y|Z) = \int p(Y|Z, \theta)p(\theta|Z)d\theta.$$

The posterior predictive distribution of Y_{exc} is then

$$p(Y_{\text{exc}}|Y_{\text{inc}}, Z) = \int p(Y_{\text{exc}}|Y_{\text{inc}}, Z, \theta)p(\theta|Y_{\text{inc}}, Z)d\theta \tag{1}$$

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

$$p(\theta|Y_{\text{inc}}, Z) = p(\theta|Z)p(Y_{\text{inc}}|Z, \theta)/p(Y_{\text{inc}}|Z),$$

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

Some Bayesians have downplayed the role of randomization, but its importance becomes clear when the model is expanded to the joint distribution of Y and I, as in the above summary. Randomization provides a practical way to assure that the selection or allocation mechanisms are ignorable for inference (Rubin 1978; Sugden and Smith 1984; Gelman et al. 2003, Chapter 7), without making ignorable selection a questionable assumption. On the other hand, making randomization the basis for inference, as with the design-based approach, is restrictive, since it does not provide a framework for handling deviations from randomization, or other non-sampling errors.

The specification of $p(Y|Z,\theta)$ in the Bayesian formulation is the same as in parametric superpopulation modeling, and in large samples, the likelihood based on this distribution dominates the contribution from the prior distribution of θ. As a result, large-sample inferences from the superpopulation modeling and Bayesian approaches are often similar, with the key distinction then being between design-based and model-based inference. Bayes modeling is to my mind superior to superpopulation modeling in small samples,

since the integration over θ in (1) propagates uncertainty in the estimation of θ, yielding better inferences than approaches that fix θ at an estimate.

2.3. The Current Design/Model Compromise

A recent comparative assessment of these approaches is given in Rao (2011).The status quo for statistical inference at the U.S. Census Bureau is a combination of design-based and model-based ideas, which I shall term the "design/model compromise" (DMC); I believe that a similar philosophy pervades other official statistical agencies. DMC applies design-based inference for descriptive statistics like means and totals in large samples, and models are used for small area estimation, to handle survey nonresponse, and in some specialized areas like time series analysis (e.g., Kalton 2002; Rao 2003, 2011). The design-based approach is often *model-assisted*, in that models are used to incorporate auxiliary information. A common form of model assistance is regression calibration, where model predictions are adjusted by adding design-weighted residuals to protect against misspecification (e.g., Cassel et al. 1977; Särndal et al.1992).

Models are used for small area estimation, since direct design-based estimates are too imprecise to be useful. An important early example is Fay and Herriott (1979). Models are used for nonresponse, though sometimes they are implicit, as in hot deck methods. In time series analysis, models are commonly used to smooth and summarize series of estimates collected over time.

Design-based and model-based systems of statistical inference both have strengths and weaknesses, and the key is to combine them in a way that capitalizes on their strengths. For reasons given below, I do not think that DMC is the best way to do this. In the next section, I describe an alternative approach, calibrated Bayes (CB), which avoids "inferential schizophrenia" by assigning distinct roles to models (for the inference) and frequentist methods (for formulating and assessing the model).

3. Calibrated Bayesian (CB) Inference

3.1. Calibrated Bayes Inference for Statistics in General

In CB, *all* inferences are explicitly Bayesian and hence model-based, but models are chosen to yield inferences that are well calibrated in a frequentist sense; specifically, models are sought that yield posterior credibility intervals with (approximately) their nominal frequentist coverage in repeated sampling. Seminal references are Box (1980) and Rubin (1984). Since my arguments in favor of CB have been presented elsewhere (Little 2006, 2011), I summarize them here, specifically in the context of survey sample inference.

Frequentist inference is in essence a set of concepts, like unbiasedness, consistency, confidence coverage, efficiency, and robustness, for assessing properties of inference procedures. It is not a prescriptive system leading to a clear choice of estimator and inferential procedure. Of the many frequentist tools, such as least squares, method of moments, generalized weighting equations or maximum likelihood (ML), asymptotic inferences based on ML seem the closest to being prescriptive, but ML is not satisfactory for small-sample inference. Exact small-sample inferences have been developed for some

problems, but in many others there is no exact frequentist method, in the sense of yielding a confidence interval that has exact nominal confidence coverage for all values of the unknown parameters.

Design-based survey inferences are not only asymptotic, they fail for probability sampling schemes where the number of distinct repeated samples is limited. For example, consider systematic sampling of units with a sampling interval of five, from a random start. The design-based standard error exists, but design-based estimates of standard error are not available, and since there are only five possible repeated samples and hence five possible estimates, design-based 90% or 95% confidence intervals do not exist. Models are needed to create and provide meaning to interval estimates.

Frequentist inference violates the likelihood principle (Birnbaum 1962), and is ambiguous about whether to condition on ancillary or approximately ancillary statistics when performing repeated sampling calculations (Cox 1971; Cox and Hinkley 1974). In the sample survey context, this issue arises in the question of whether the sampling distribution of post-stratified means should condition on post-stratum counts (Holt and Smith 1979; Little 1993).

The Bayesian approach avoids these problems with frequentist inference. Once a model and prior distribution are specified, there is a clear path to inferences based on the posterior distribution, or optimal estimates for a given choice of loss function. Problems of inference under a model become purely computational, and a rich array of Bayesian computational tools are now available, even for complex high-dimensional problems. The likelihood principle is satisfied, issues about conditioning on ancillary statistics do not arise, and uncertainty about nuisance parameters is propagated by integrating them over their posterior distribution, an approach that (with noninformative prior distributions) leads to better small-sample inferences than ML. In the simplest case of a normal model and simple random sampling, integrating out the variance leads to inferences based on the t distribution.

The problem with Bayesian inference in practice is that it generally requires full specification of a likelihood and prior, and we never know the true model (Efron 1986). All models are wrong, and bad models lead to bad answers: under the frequentist paradigm, the search for procedures with good frequentist properties provides a degree of protection against model misspecification, but there seems no such built-in protection under a strict Bayesian paradigm where frequentist properties are not entertained.

We want model-based inferences with good frequentist properties, such as 95% credibility intervals that cover the unknown parameter approximately 95% of the time if the procedure is applied to repeated samples. The Bayesian has some important tools for model development and checking, like Bayes factors and model averaging, but in my view frequentist ideas are essential when it comes to model development and assessment. A natural compromise is thus to use frequentist methods for model development and assessment, and Bayesian methods for inference under a model. This capitalizes on the strengths of both paradigms, and is the essence of calibrated Bayes (CB) (Peers 1965; Welch 1965; Dawid 1982; Box 1980; Rubin 1984; Draper and Krnjajic 2010). Rubin (1984) wrote that

> "The applied statistician should be Bayesian in principle and calibrated to the real world in practice – appropriate frequency calculations help to define such a tie. . . frequency

calculations are useful for making Bayesian statements scientific, scientific in the sense of capable of being shown wrong by empirical test; here the technique is the calibration of Bayesian probabilities to the frequencies of actual events."

3.2. *Calibrated Bayes Inference for Sample Surveys*

What are the implications of CB for sample survey inference? The main features that distinguish survey sampling inference from other areas of statistics are (a) the focus on descriptive finite population quantities (though analytic parameters are also of interest) and (b) the presence of survey design features like stratification, weighting and clustering, which render simple "iid" assumptions invalid.

Bayesian inference is highly suited to finite population quantities; the tool is the posterior predictive distribution. This distribution automatically incorporates finite population corrections – the uncertainty in the posterior predictive distribution goes to zero as the sampling fraction goes to one. The target population quantity does not need to be a parameter of the CB model used for inference; it could be the quantity obtained by applying a "target model" to the full population. CB inference is then based on the posterior predictive distribution of this finite population quantity, for an "analysis model", which captures key features of the sample design, and which may differ from the target model. This point is developed in the context of multiple regression in Section 4.2 below.

Concerning (b), the need for calibration, combined with the appreciation that all models are approximations and hence to some degree misspecified, leads to Bayesian models that incorporate design features like stratification, weighting and clustering. Design features need to be included in the model to protect against the effects of model misspecification.

Specifically, models for cluster samples that assume units within clusters are independent overstate precision when outcomes of units within clusters are correlated. Thus, hierarchical Bayes models that include random effects for clusters, as in the seminal work of Scott and Smith (1969), are needed to model clustering of the sample. Models for stratified unequal probability samples that do not allow distinct parameters across strata make the dubious assumption that strata variables are unrelated to outcomes. Thus, stratified samples require models that include strata indicators as covariates. For probability proportional to size samples, models that misspecify the relationship between the outcome and size are not well calibrated. Robust modeling of this relationship, for example by modeling the outcome as a spline function of size, avoids this problem, and has been shown to yield Bayesian inferences with superior frequentist properties to sample-weighted estimates in simulations (Zheng and Little 2004, 2005; Yuan and Little 2007, 2008; Chen et al. 2010).

Frequentist concepts like design consistency or asymptotic design unbiasedness (Brewer 1979; Isaki and Fuller 1982) are useful in developing CB models, particularly for inference with large samples where asymptotic properties are relevant. Strictly speaking, design consistency of estimates is not a requirement of CB, since a design-inconsistent Bayes estimate for a well-specified model can still achieve good frequentist coverage. However, design consistency plays a role in CB as a useful robustness property that tends to promote good confidence coverage, particularly in large samples; the class of Bayesian

models that yield design consistent estimates is very broad, so design consistency is relatively easy to achieve under the CB paradigm.

Other features of CB models for surveys are that (a) relatively weak prior distributions should be favored so that the evidence in the data dominates the evidence in the prior; and (b) model checks become an important feature of the analysis. The latter point should not be controversial, since any statistical approach, frequentist or Bayesian, needs to evaluate assumptions. Diagnostic approaches include posterior predictive checks (Rubin 1984; Gelman et al. 1996), and cross-validation approaches (Draper 1995; Draper and Krnjajic 2010).

The following simple examples from Little (2003) illustrate these ideas.

Example 1. Stratified Random Sampling. Suppose the population with units $i = 1, \ldots, N$ is divided into H strata and n_h units are randomly selected without replacement from the population of N_h units in stratum h. Define Z as a stratum variable, with $z_i = h$, if unit i is in stratum h. A CB model for an outcome Y that conditions on the stratum variables z_i is

$$\left[y_i | z_i = h, \left\{\theta_h, \sigma_h^2\right\}\right] \sim_{\text{ind}} G\left(\theta_h, \sigma_h^2\right), \tag{2}$$

where $G(a, b)$ denotes the normal (*G*aussian) distribution with mean a, variance b. Suppose first σ_h^2 is known and the stratum mean are assigned a flat prior

$$p(\theta_h | Z) \propto \text{const.}$$

Bayesian calculations lead to the posterior predictive distribution for the population mean $\overline{Y}$:

$$\left[\overline{Y} | Z, \text{ data}, \ \left\{\sigma_h^2\right\}\right] \sim G\left(\overline{y}_{\text{st}}, \sigma_{\text{st}}^2\right)$$

the normal distribution with posterior mean:

$$\overline{y}_{\text{st}} = \sum_{h=1}^{H} P_h \overline{y}_h, P_h = N_h / N, \overline{y}_h = \text{sample mean in stratum } h,$$

and posterior variance:

$$\sigma_{\text{st}}^2 = \sum_{h=1}^{H} P_h^2 (1 - f_h) \sigma_h^2 / n_h, f_h = n_h / N_h.$$

These Bayesian results lead to Bayes posterior credibility intervals that are identical to standard confidence intervals from design-based inference for a stratified random sample. In particular, the posterior mean weights each case by the inverse of its inclusion probability, and the posterior variance equals the design-based variance of the stratified mean.

With unknown variances, the standard design-based approach has weaknesses. Replacing the variances $\{\sigma_h^2\}$ by sample stratum variances leads to normal confidence intervals that fail to achieve nominal coverage when some $\{n_h\}$ are small, since uncertainty in the estimated variances is not incorporated; or pooling the sample stratum variances assumes the variances $\{\sigma_h^2\}$ are equal, leading to confidence intervals with the

wrong width when this assumption is strongly violated. The CB approach addresses these weaknesses, by assigning $\{\log(\sigma_h^2)\}$ uniform prior distributions. The resulting posterior distribution of $\overline{Y}$ is a mixture of t distributions, yielding improved frequentist coverage in small samples because uncertainty in estimating the stratum variances is propagated.

Suppose we ignore stratum effects, that is, we assume $\theta_h = \theta, \sigma_h = \sigma$ in Eq. (2). The posterior mean of $\overline{Y}$ is then the unweighted sample mean, which is potentially very biased if the sampling rates vary across the strata. The problem is that inferences from this model are nonrobust to violations of the assumption of no stratum effects, and we expect stratum effects in most settings. The CB perspective leads to a model like (2) that allows for stratum effects.

Example 2. Two-stage Sampling. Suppose the population is divided into C clusters, based for example on geographical areas. A simple form of two-stage sampling first selects a simple random sample of c clusters, and then selects a simple random sample of n_c of the N_c units in each sampled cluster c. The inclusion mechanism is ignorable conditional on cluster information, but a CB model needs to account for within-cluster correlation in the population. A normal model that does this is:

$$\begin{aligned} &y_{ci} = \text{outcome for unit } i \text{ in cluster } c, \quad i = 1, \ldots, N_c; \quad c = 1, \ldots, C. \\ &\left[y_{ci} | \theta_c, \sigma^2\right] \sim_{\text{ind}} G\left(\theta_c, \sigma^2\right), \\ &[\theta_c | \mu, \phi] \sim_{\text{ind}} G(\mu, \phi). \end{aligned} \tag{3}$$

Unlike the model for stratified sampling in Eq. (2), the cluster means cannot be assigned a flat prior, $p(\theta_c) = \text{const}$, because only a subset of the clusters are sampled; the uniform prior does not allow information from sampled clusters to predict means for non-sampled clusters. The model that assumes no cluster effects, $\phi = 0$ in (3), yields poor confidence coverage in the presence of cluster effects, particularly in highly clustered samples. If the first stage clusters are sampled with probability proportional to size, a CB model needs to include the size variable as a covariate in Eq. (3); see, for example Zheng and Little (2004).

4. DMC and CB Perspectives on Some Analysis Issues

4.1. Design-Based Statistical Standards for Model-Based Analysis

The statistical standards at the U.S. Census Bureau are essentially design-based, whereas many Census Bureau researchers are social scientists targeting substantive journals in disciplines such as economics and demography, where statistical models are the norm. This difference in underlying philosophy leads to confusion and conflict. The statistical standard-bearers play the role of high priests in a religion that many social scientists have not embraced.

If, on the other hand, statistical standards were written from a CB perspective, the inference would always be model-based, greatly reducing the communication gap between social science modelers and standard-setters; the role of design features in the analysis is to find robust and well-specified models. The fact that the inference is Bayesian is admittedly a departure for modelers more versed in superpopulation frequentist modeling. The gap may

not be as large as sometimes suggested – for example, economists act very much like Bayesians, in the sense that prior judgment enters strongly into model specification through variable selection, assumptions about instrumental variables, exclusion restrictions, and so on. The additional information injected by including a diffuse Bayesian prior distribution is usually minor compared to the assumptions required to identify models.

Bayesian inferences have repeated sampling properties, like any other inferential procedure. All modelers interested in obtaining robust inferences should embrace the calibrated part of CB. In the finite population context, estimates for a model fitted to the sample should be close to the estimates that would be obtained if that model were fitted on the entire population. One way of achieving this is to incorporate features of the sample design, such as weighting and clustering, into the model, since ignoring features like the design weights yields inferences that are vulnerable to model misspecification (Kish and Frankel 1974; Holt et al. 1980; Hansen et al. 1983; Pfeffermann and Holmes 1985).

4.2. Role of Sampling Weights in Regression

The conflict between design-based statisticians and modelers arises in the role of sampling weights. A design-based analysis weights units in the regression analysis by the inverse of their selection probability (Horvitz and Thompson 1952), but this form of weighting is seen as unnecessary in many branches of economics, where extrapolation to a population is not the primary aim, and weights, if used at all, model nonconstant variance (Konjin 1962; Brewer and Mellor 1973; Dumouchel and Duncan 1983; Smith 1988; Pfeffermann 1993; Little 2004).

From a CB viewpoint, it is useful to distinguish the case where the variables defining the sampling weights (e.g., the strata indicators in Example 1 above) are or are not included as predictors in the model. If they are, then design weighting is unnecessary if the model is correctly specified. However, from a CB perspective, a comparison of estimates from the weighted and unweighted analysis provides an important specification check, since a serious difference between a design-weighted and unweighted estimate is a strong indicator of misspecification of the regression model. Since specification checks for the hard problem of selection on unobservables are popular in econometrics (e.g., Heckman 1976), we should welcome checks for the much easier problem of selection on observables! Dumouchel and Duncan (1983) propose a test comparing the weighted and unweighted regression coefficients, and extensions of this idea to other complex survey designs would be useful; furthermore, determining what constitutes a "serious" difference between weighted and unweighted estimates is not obvious.

If the variables defining the weights are not included as predictors in the regression model, the design-weighted regression is a simple way of correcting for selection bias in the sample. In fact, the design-weighted estimates have an interpretation as approximate posterior means for a CB model, as in the following example (Little 1991, 2004). This example also illustrates the distinction between a target model and an analysis model mentioned in Section 3.2.

Example 3. Distinct Target and Analysis Models, Leading to a Bayesian Interpretation of Design-weighted Regression Estimates. I noted in Section 3.2 that the target quantity in a CB analysis does not have to be a parameter in a CB model (or its finite population

equivalent). It is useful to distinguish a target model, which determines the target quantity of interest, and the analysis model, the basis for inferences about the target quantity.

Consider first inference about a population mean from a stratified sample, as in Example 1. The target model assumes the outcome y_i for unit i has a mean that does not depend on stratum, and a non-constant variance, namely

Target model:

$$[\,y_i|u_i, z_i = h,\ \{\theta, \sigma_i^2\}] \sim_{\text{ind}} G\big(\theta,\ \ \sigma^2/u_i\big), \tag{4}$$

where u_i is a known constant. The target quantity is the result of applying this model to the whole population with an uninformative prior, namely the precision-weighted mean:

$$\overline{Y}^{(u)} = \left(\sum_{i=1}^{N} u_i y_i\right) \Big/ \left(\sum_{i=1}^{N} u_i\right). \tag{5}$$

This is the finite population mean if $u_i = 1$ for all i, but other choices of $\{u_i\}$ lead to useful target quantities. For example, if $y_i = x_i/u_i$ then Eq. (4) defines the ratio model, and Eq. (5) is the population ratio $\big(\sum_{i=1}^{N} x_i\big)\big/\big(\sum_{i=1}^{N} u_i\big)$.

A standard design-based approach weights cases in stratum j by their sampling weight $w_j = N_j/n_j$, yielding design-unbiased estimates of the numerator and denominator of Eq. (5):

$$\overline{y}^{(w^*)} = \left(\sum_{j=1}^{J} w_j \sum_{i\in s_j} u_i y_i\right) \Big/ \left(\sum_{j=1}^{J} w_j \sum_{i\in s_j} u_i\right) = \left(\sum_{j=1}^{J}\sum_{i\in s_j} w_{ji}^{*} y_i\right) \Big/ \left(\sum_{j=1}^{J}\sum_{i\in s_j} w_{ji}^{*}\right), \tag{6}$$

where $w_{ji}^{*} = w_j u_i$ is the product of the sampling weight and the precision weight. This estimator can also be motivated as an approximate posterior mean under a CB model, as follows:

The target Model (4) ignores the stratified nature of the sample, and for inference purposes it is vulnerable to misspecification if the means of Y and selection rates vary across the strata. Thus for inference about (5), we replace (4) by an analysis model that allows different parameters for the mean and variance in each stratum, that is:

Analysis model:

$$\begin{aligned} &\Big[y_i|z_i = j,\ \Big\{\theta_j, \sigma_j^2\Big\}\Big] \sim_{\text{ind}} G\Big(\theta_j,\ \ \sigma_j^2/u_i\Big), \\ &p\Big(\Big\{\mu_j,\ \log\sigma_j^2\Big\}\Big) = \text{const.} \end{aligned} \tag{7}$$

This model yields a posterior predictive distribution for the nonsampled values, and hence for the target quantity (5). If $\{u_i\}$ are known for all units of the population, a standard Bayesian calculation yields

$$E\big(\overline{Y}^{(u)}|\text{data},\ \{u_i\}\big) = \left(\sum_{j=1}^{J} \overline{y}_j^{(u)} u_{+j}\right) \Big/ \left(\sum_{j=1}^{J} u_{+j}\right),$$

where $\overline{y}_j^{(u)} = \sum_{i\in s_j} u_i y_i / \sum_{i\in s_j} u_i$ is the precision-weighted mean of the sampled units $i \in s_j$ in stratum j, and u_{+j} is the sum of u_i for all units i in stratum j. If $\{u_i\}$ are only known

for sampled units of the population, a model is also needed to predict values $\{u_i\}$ for nonsampled units. A variety of models for $\{u_i\}$ that involve distinct means in each stratum yield a posterior mean of the total in stratum j of the form $E(u_{+j}|\text{data}) \approx w_j \sum_{i\in s_j} u_i$, where $w_j = N_j / n_j$ is the sampling weight for stratum j. Then

$$
\begin{aligned}
E(\overline{Y}^{(u)}|\text{data}) &= E\left[\left(\sum_{j=1}^{J}\overline{y}_j^{(u)}u_{+j}\right)\Big/\left(\sum_{j=1}^{J}u_{+j}\right)|\text{data}\right] \\
&\approx \left(\sum_{j=1}^{J}\overline{y}_j^{(u)}E\{u_{+j}|\text{data}\}\right)\Big/\left(\sum_{j=1}^{J}E\{u_{+j}|\text{data}\}\right) \\
&= \left(\sum_{j=1}^{J}\overline{y}_j^{(u)}w_j\sum_{i\in s_j}u_i\right)\Big/\left(\sum_{j=1}^{J}w_j\sum_{i\in s_j}u_i\right) = \sum_{j=1}^{J}\sum_{i\in s_j}w_{ji}^{*}y_i\Big/\sum_{j=1}^{J}\sum_{i\in s_j}w_{ji}^{*} = \overline{y}^{(w^*)},
\end{aligned}
$$

the design-weighted estimator (6). The approximation in the second line of this expression results from replacing the posterior expectation of a ratio by a ratio of posterior expectations, which ignores terms of order $O(1/n)$. Hence, under this formulation, the CB approach leads to weighting by the product of the sampling weight and precision weight, as in the design-based approach.

An extension of this analysis yields design-weighted estimates for regression coefficients. Consider more generally the target regression model

Target model:

$$
(Y|X,\beta) \sim G(X\beta, U^{-1}\sigma^2), \tag{8}
$$

where Y consists of the population elements as an $(N\times 1)$ vector, X is an $(N\times p)$ matrix of covariates, and U is a $(N\times N)$ diagonal matrix with the value $\{u_i\}$ on the diagonal. The target quantities are the precision-weighted least squares estimates:

$$
B^{(u)} \sim (X^{\mathrm{T}}UX)^{-1}X^{\mathrm{T}}UY. \tag{9}
$$

For inference about (9), we assume an analysis model that allows different stratum regression coefficients, namely

Analysis model:

$$
\begin{aligned}
&\left(Y_j|X_j,\beta_j,\sigma_j^2\right) \sim G\left(X_j\beta_j,\ \ U_j^{-1}\sigma_j^2\right), \\
&p\left(\left\{\beta_j,\ \log\sigma_j^2\right\}\right) = \text{const.}
\end{aligned} \tag{10}
$$

where Y_j, X_j are the components of Y and X in stratum j, with dimension $(N_j\times 1)$ and $(N_j\times p)$ respectively. An approximation to the posterior mean of $B^{(u)}$ under (10) is obtained by writing (9) as a function of sums $B^{(u)} = g(T_1,\ldots,T_L)$, where $\{T_\ell = \sum_{j=1}^{J}\sum_{i=1}^{N_j}u_{ji}h_{\ell ji},\ \ell = 1,\ldots,L\}$, for difference choices of $\{h_{\ell ji}\}$ represent the set of sums, sums of squares, and sums of cross products of the covariates and outcome.

Then

$$E(B^{(u)}|\text{data}) = E\big(g(T_1, \ldots, T_L)|\text{data}\big) = g\big(E(T_1|\text{data}), \ldots, \text{E}(T_L|\text{data})\big) + O(1/n)$$

by a linearization argument similar to that used for design-based inference. Also,

$$E(T_\ell|\text{data}) \approx \sum_{j=1}^{J} \sum_{i \in s_j} w^*_{ji} h_{\ell ji},$$

where $w^*_{ji} = w_j u_{ji}$ and w_j is the sampling rate in stratum j, applying an argument similar to that for the mean model to $\{h_{\ell ji}\}$. Hence the posterior mean is approximated by the design-weighted regression estimates:

$$E(B^{(u)}|\text{data}) \approx \left(X_s^T W^*_s X_s\right)^{-1} X_s^T W^*_s Y_s, \tag{11}$$

where the subscript s denotes sample quantities (Little 2004).

Can sampling weights be ignored when interest lies in "analytic" inference for the parameters β of the target Model (10), rather than in the finite population quantity (9)? I would say no, Eq. (11) should still be used to estimate β. The inference differs only in the omission of finite population corrections, which follows directly from the application of Bayes' theorem. My reason is that the finite population is assumed to be a random sample from the superpopulation under the superpopulation model, so β differs from the finite population quantity $B^{(u)}$ by a (small) quantity of order $O(1/N)$. Since ignoring the sampling weights yields a poor estimate of $B^{(u)}$, it also yields a poor estimate of β.

What is gained by the CB approach if the analysis model (10) merely recovers the design-based estimator? The Bayesian paradigm allows for better small-sample inferences, by propagating error in estimating the variances, and by allowing the possibility of shrinkage of the weights by mixed models.

4.3. DMC and CB for Small Area Estimation

The DMC philosophy suggests that when there are sufficient data to support "direct" estimates that do not borrow strength across subdomains, inferences are design-based, but when the data are too limited then model-based small area estimates are acceptable. This dichotomy implies, for any particular survey, the existence of a tipping point (say n_0), the "point of inferential schizophrenia", such that inferences are design-based when $n > n_0$ and model-based when $n < n_0$. The choice of n_0 is of course rather arbitrary, and it bothers me that one's entire philosophy of statistics, and the nature of the estimator, changes depending on where the sample size falls relative to this value (Fig. 1A). In particular, the (design-based) confidence intervals for the mean for sample sizes slightly more than n_0 will be wider than the (model-based) confidence intervals for the mean for sample sizes slightly less than n_0, even though they are based on more data.

The CB philosophy avoids "inferential schizophrenia", since all inferences are model-based. Hierarchical Bayes models yield estimates close to "direct" estimates when sample sizes are large, and as the sample size decreases, move seamlessly towards predictions from a fixed-effects model (Fig. 1B). Consider, for example, the following simple

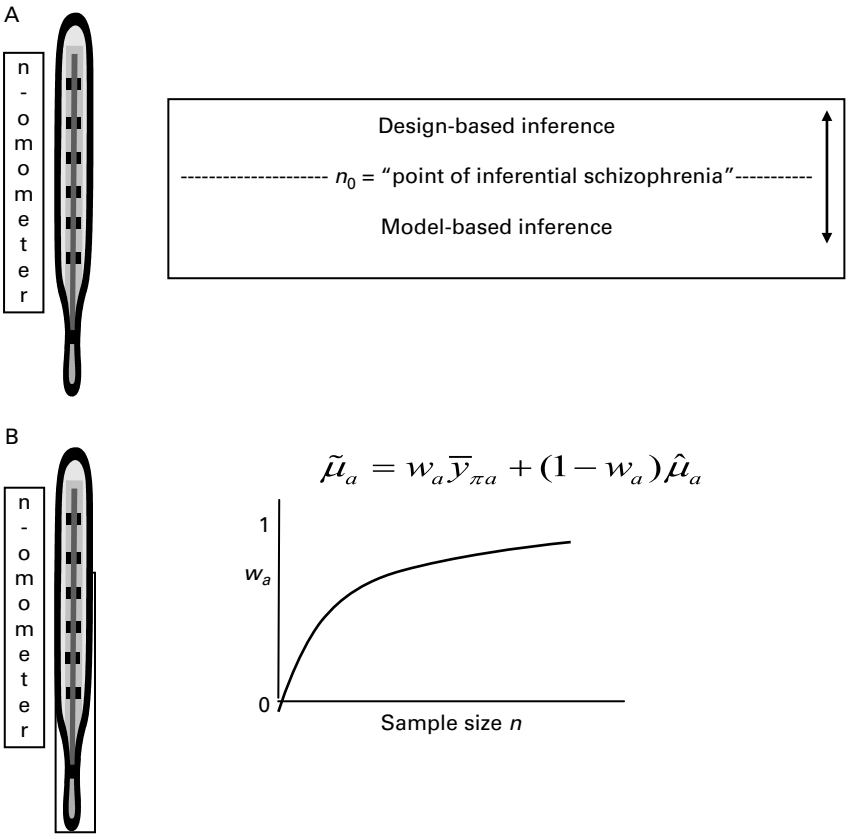

Fig. 1. (A) Discontinuity between design-based and model-based inference in DMC. (B) Hierarchical Bayes estimate for area a is weighted combination of direct estimate $\bar{y}_{\pi a}$ and regression estimate $\hat{\mu}_a$. Weight w_a on the direct estimate increases with the sample size n

hierarchical Bayes model for simple random sampling, relating an outcome Y to a covariate X measured for all units in the population:

$$\begin{aligned} &y_{ai}|\alpha_a \sim N\big(\mu_{ai}, \sigma^2\big), \mu_{ai} = \alpha_a + \beta x_{ai}, \\ &\alpha_a \sim N\big(\alpha, \tau^2\big), \end{aligned} \tag{12}$$

where x_{ai}, y_{ai} are the value of Y and X for unit i in area a, and α_a is a random intercept for area a. (A more complex model would entertain interactions between the areas and covariates). If the sampling fraction in area a is small, the posterior mean of the population mean $\overline{Y}_a$ in area a given (σ^2, τ^2) has the form

$$E\big(\overline{Y}_a|\text{data}\big) = w_a\bar{y}_a + \big(1 - w_a\big)\big(\bar{y} + \hat{\beta}\big(\bar{x}_a - \overline{X}\big)\big), \tag{13}$$

where $\bar{y}_a$, $\bar{x}_a$, n_a are the sample means of Y and X and sample size in area a, $\big(\bar{y} + \hat{\beta}\big(\bar{x}_a - \overline{X}\big)\big)$ is the regression prediction for the mean of Y aggregated over all areas, and $w_a = n_a\sigma^2/(n_a\sigma^2 + \tau^2)$ assigns most of the weight to the sample mean when n_a is large, and most of the weight to the regression prediction over all areas when n_a is small.

The weights here depend on the variances, which in practice need to be estimated. Empirical Bayes approaches replace the variances by point estimates, typically computed by the method of moments or maximum likelihood. When the estimate of τ^2 goes negative, it is replaced by a value 0 on the boundary of the parameter space. Uncertainty in the variance estimates is not reflected in inferences. Fully Bayes methods based on weak priors on the variance components propagate uncertainty and avoid estimates on the boundary of the parameter space, though care is needed with the choice of prior distribution for τ^2 (Gelman 2006).

What precisely is the role of CB in small area estimation? Essentially, that Bayes is preferable to empirical Bayes because it addresses uncertainty in the variance components, and as a result, it tends to be better calibrated, that is, yields credibility intervals with better confidence coverage. Two other related issues raised by the referees are that (a) model-based estimators have a bias that does not necessarily vanish with increasing sample size, and that can be substantial and dominate the MSE if the model fails; and (b) CB for small areas yields estimates that do not necessarily sum to design-based estimates for higher levels of aggregation. My view is that "design consistency", not "design bias", is the important issue, since the essence of shrinkage estimates is that exact unbiasedness is secondary to mean squared error. Estimates for any single CB model are automatically internally consistent, since predictions of quantities at high levels of aggregation are sums of the predictions at lower levels. (Of course, this property no longer holds if different models are applied at different levels of aggregation.) Design-inconsistent estimates from a CB model may be adequately calibrated for small areas, because design bias is not an important component of mean squared error; but design bias from model misspecification becomes an issue when these small area estimates are aggregated to higher levels. Thus, if aggregation to higher levels is important, then I recommend seeking a CB model that yields design-consistent estimates.

4.4. CB for Small Area Inference: Fixing the "Standard Error Error"

Official statistics often presents uncertainty in the form of standard errors or margins of error. In particular, users of the U.S. American Community Survey (ACS) have the ability to generate tables of estimated counts of individuals by race, age and gender, in small areas. Results are reported by an estimate and a margin of error, chosen so that the estimate plus or minus the margin of error is asymptotically a 90% confidence interval. However, in many instances the margin of error is larger than the estimate, yielding intervals containing negative counts of people! The ACS documentation suggests truncating the resulting intervals so that they are bounded below by zero, but the confidence interval based on the margin of error still fails to have the nominal coverage in small samples, since it is based on a large-sample approximation.

This exemplifies a general weakness of design-based inferences – that they are too focused on estimates and standard errors, assuming that we are in the "land of asymptotia" where an estimate plus or minus two standard errors is truly a 95% confidence interval. We learn in elementary statistics that this is false when the sample size is small, as when a t correction is applied to a normal test or confidence interval when the variance is not

known. In simulation studies with realistic sample sizes, design-based confidence intervals often fail to achieve the nominal coverage (e.g., Zheng and Little 2004, 2005; Yuan and Little 2007, 2008; Chen et al. 2010). A comprehensive theory for finite samples should be able to deal with small sample sizes, and (as discussed below) the simplest general way to achieve this is to make the inference Bayesian. The concern is that the introduction of the prior distribution adds subjective information, but Bayes credibility intervals with noninformative priors tend to be more, not less, conservative than design-based confidence intervals.

In particular, it is well known that asymptotic Wald confidence intervals for proportions do not achieve nominal coverage when the sample size is small, particularly for proportions close to zero or one (Brown et al. 2001). Simple fixes such as the Wilson estimate, which for a 95% interval adds 2 to the numerator and 4 to the denominator of the proportion (Agresti and Coull 1998), have a Bayesian interpretation. The Bayesian posterior credibility interval based on a noninformative Jeffreys' prior distribution is constrained to lie between 0 and 1, is appropriately asymmetric when the estimate is close to zero or one, and has better confidence coverage than the asymptotic Wald interval (Brown et al. 2001). Extensions of the Bayesian approach to unequal probability sampling show similar improved frequency properties over design-weighted and model-assisted approaches (Chen et al. 2010).

4.5. Model-Assisted Estimation

The prevailing paradigm of design-based inference is model-assisted, where model predictions are calibrated to yield estimates that are design-consistent (Brewer 1979; Isaki and Fuller 1982) and hence protected from model misspecification – note that this use of the term "calibration" differs from the calibration in CB. This popular approach uses regression models on auxiliary data to increase the efficiency of design-based inferences while retaining the randomization distribution as the basis for inference. A weakness of the method is that, by modifying the prediction estimator to improve its robustness, the resulting estimator can involve parameter estimates from conflicting models held simultaneously. I would rather base inferences on predictions from a model that yields design consistent estimates. Since design consistency is a rather weak property, this is not hard to do in many problems (Firth and Bennett 1998). In short, model-assisted estimators represent for me a rather ad hoc way of making a design-based estimator robust to model misspecification, whereas a more direct approach is simply to choose a more robust model. The following example (Little 2007) illustrates this point.

Example 4. Generalized Regression in An Equal Probability Sample Based on a Regression Model Without An Intercept. Opsomer et al. (2007) applied the model-assisted approach to incorporating auxiliary information into an equal probability sample, where the regression models for prediction did not include the constant term. Let $\bar{y}$ and $\bar{x}$ denote sample means of an outcome Y and a vector of covariates X, $\hat{\beta}_x$ the vector of least squares slopes for the regression of Y on X with no intercept, and $\overline{X}$ the population mean of X. The resulting regression prediction of the mean of Y is $\hat{\beta}_x\overline{X}$, and the average residual for the sampled cases is $\bar{y} - \hat{\beta}_x\bar{x}$, so the generalized regression

estimator has the form

$$\overline{Y}_{\text{GREG}} = \hat{\beta}_x \overline{X} + \{\overline{y} - \hat{\beta}_x \overline{x}\} = \hat{\beta}_0 + \hat{\beta}_x \overline{X}, \text{ where } \hat{\beta}_0 = \overline{y} - \hat{\beta}_x \overline{x}.$$

Observe that the slopes β_x are estimated under the regression model that assumes no intercept, but the inclusion of $\hat{\beta}_0$ in $\overline{Y}_{\text{GREG}}$ implies a model that includes an intercept. If an intercept is needed, it should be included in the model when estimating $\hat{\beta}_x$. Since any linear model with an intercept yields design-consistent predictions under equal probability sampling (Firth and Bennett 1998), there is then no need for calibration at all in this situation. Other examples of model inconsistency in model-assisted estimates from unequal probability samples are given in Little (1983).

Is this a "counter-example"? It depends on the extent to which one cares about the logical consistency of estimators from the viewpoint of the prediction – since the CB perspective views the task of statistics as fundamentally to provide posterior predictive inference for unknowns given the data, it places considerable weight on this aspect. A more pragmatic CB argument against model-assisted approaches is that the resulting confidence intervals do not achieve the nominal coverage, particularly when the sampling weights applied to the residuals are highly variable (Zheng and Little 2004, 2005; Yuan and Little 2007, 2008; Chen et al. 2010).

Another comment about model-assisted estimation is that it is a tool for incorporating auxiliary data, but not effective for small area estimation – hierarchical Bayes models like (4) above that incorporate shrinkage via random effects are more suited to this purpose. For example, in the setting of Model (4) with equal probability sampling, the form of the generalized regression estimator with predictions based on the regression of Y on X is

$$\overline{y}_{\text{GREG},a} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}_a + \overline{y}_a - \hat{\beta}_0 - \hat{\beta}_1 \overline{x}_a = \overline{y}_a + \hat{\beta}_1 (\overline{X}_a - \overline{x}_a),$$

which incorporates information in the auxiliary variable X, but does not incorporate shrinkage to the regression estimate $\overline{y} + \hat{\beta}_1(\overline{x}_a - \overline{X})$ combined over areas, as in the CB estimate (13). This lack of shrinkage also applies to unequal probability samples, where the model-assisted approach calibrates the regression estimate by adding weighted residuals. For discussions of model and design-based approaches to survey weights, see Little (2004, 2008) and Gelman (2007).

4.6. Methods for Propagating Imputation Uncertainty

Single imputation methods lead to confidence intervals that are too narrow (that is, have less than nominal coverage) when imputation uncertainty is not propagated. There are model-based and design-based approaches to correcting this problem. A Bayesian approach is multiple imputation, where multiple data sets are generated with different sets of draws from the predictive distribution of the missing values (Rubin 1987, 1996). A design-based approach is to apply replicate methods such as the jackknife (Rao 1996; Fay 1996), with different imputations in each replicate; these methods are design-based in spirit but "pseudo" randomization-based in fact, since they rely on an assumption that, within classes, nonresponse is in effect a form of random sampling. Multiple imputation does not yield consistent estimates of variance under particular forms of model misspecification (Meng 1994; Rao 1996; Fay 1996; Robins and Wang 2000;

Kim et al. 2006). Modelers accept model misspecification as inevitable, and seek multiple imputation models that capture key features of the population – they also point to simulations suggesting that multiple imputation under plausible models generally yields good or conservative confidence coverage.

I view this as a proxy fight for the more basic underlying philosophical differences. At the Census Bureau it has led to a form of stalemate, where single imputation methods that fail to propagate imputation error continue to be applied, even though both of the alternatives mentioned above are clearly superior to the status quo.

4.7. DMC-Induced Constrictions of the Total Survey Error Paradigm

Total survey error (TSE) centers around a decomposition of mean squared error of a survey estimate into components of sampling error, and nonsampling errors such as frame errors, errors due to nonresponse, response errors, editing and interviewer effects. In a recent review of TSE, Groves and Lyberg (2010) note that the explicit attention to the decomposition of errors in TSE, and the separation of phenomena affecting statistics in various ways, provides a central conceptual basis for the field of survey methodology. At the same time, they point out the following weaknesses of the current TSE paradigm:

i) quantitative measurement of many components is burdensome and lagging;
ii) the TSE paradigm has not led to enriched error measurement in practical surveys;
iii) assumptions required for some estimators of error terms are frequently not true;
iv) there is a mismatch between existing error models and theoretical causal models of the error mechanisms;
v) there is a misplaced focus on descriptive statistics; and
vi) there is a failure to integrate error models developed in other fields.

I believe that a primary source of these weaknesses is the design-based tradition of survey inference, making it difficult to harmonize in a single inference the design-based approach for sampling errors and model-based approach needed for non-sampling errors. An explicitly model-based CB representation of the TSE concept, drawing heavily on Rubin's unified concepts of causal inference and missing data (Rubin 1974), addresses many of the failures in implementing the TSE paradigm.

4.8. Incorporating Information from Multiple Data Sources

The modeling paradigm of CB is particularly relevant to problems of combining data across data sources. The design-based paradigm can incorporate known administrative data, using methods such as post-stratification or raking, and methods from multiple frame probability samples, but a modeling framework like CB is required for combining information from probability samples with information from nonprobability sources, or sources where nonsampling errors need to be modeled. The topic is too large for an extended treatment here, but see Elliott and Little (2005), Schenker and Raghunathan (2007), and Raghunathan et al. (2007) for examples of Bayesian approaches to combining information from different data sources.

5. Two Census Bureau Applications

While DMC is the prevailing philosophy of statistics at the Census Bureau, there is an increasing acceptance of model-based, and even Bayesian methods. In this section I describe two small area estimation topics that are being addressed from a CB perspective.

Example 5. Small Area Income and Poverty Estimates. The U.S. Census Bureau Small Area Income and Poverty Estimates (SAIPE 2011) are intercensal estimates of selected income and poverty statistics for school districts, counties, and states, for the administration of federal programs and the allocation of federal funds to local jurisdictions. Data from administrative records, intercensal population estimates, and the decennial census are combined with direct estimates from the American Community Survey to provide consistent and reliable single-year estimates. Direct survey estimates (from the Current Population Survey, CPS, or more recently from the American Community Survey, ACS) are too unreliable for many areas, and a small area model is applied to integrate survey data with data from administrative records and the previous census long form. The basic form of the model (Fay and Herriott 1979) is

$$y_a|\theta_a, v_a \sim N(\theta_a, v_a)$$
$$\theta_a|\beta, \sigma^2 \sim N\left(x_a'\beta, \sigma^2\right),$$

where y_a is the direct survey estimate of population quantity θ_a for area a, v_a is the sampling variance of y_a, x_a is a vector of regression variables for area a with associated regression parameters β, and σ^2 is the variance of small area random effects. Initially the variances v_a and σ^2 were treated as known, but more recent formulations have included prior distributions as part of a Bayesian formulation.

In particular, for the state poverty rate model for ages 5–17, the direct survey estimates y_i were originally from CPS, but since 2005 are from the ACS; the regression variables in x_i include a constant term and, for each state, pseudo-poverty rate for children from tax return data tax "nonfiler rate", SNAP (food stamp) participation rate, previous census estimated state 5–17 poverty rate, or residuals from regressing previous census estimates on other elements of x_i for the census year. Table 1 presents CPS sample size, direct variance v_a and posterior variance for four states from the State Model for 2004 CPS 5–17 Poverty Rates. For California (CA), the sample size is large, most of the weight (61%) is on the direct estimate, and the posterior variance (0.8) is not much smaller than the direct variance (1.1). For Mississippi (MS), the sample size is small, most of the weight (87%) is on the model prediction, and the posterior variance (3.9) is much smaller than the direct variance (12.0). The other two states lie between these two.

Table 1. Posterior Variances from SAIPE State Model for 2004 CPS 5-17 Poverty Rates. Results for four states

State	n_a	v_a	Var(Y_a|data)	Approx. wt. on Y_a in E(Y_a|data)
CA	5,834	1.1	0.8	.61
NC	1,274	4.6	2.0	.28
IN	904	8.1	2.0	.18
MS	755	12.0	3.9	.13

Example 6: Language Provisions of the Voting Rights Act. The Voting Rights Act determines that certain counties and townships are required to provide language assistance at the polls. Determinations are based in part on there being more than 5 percent of voting age citizens in a political district who are members of a single language minority and are limited English proficient (LEP). The Census Bureau is charged with determining which jurisdictions are covered under the Act, and until now have used direct estimates from Long Form Decennial Census Data. With the replacement of the long form, estimates are henceforward to be based on the smaller ACS, and some districts have small ACS samples and hence have direct estimates with unacceptably high variance. The 2011 determinations use a small area model that combines information from the 2005–2009 ACS and 2010 Census data. To see why a model is needed, let P be the proportion of voting age citizens in a voting district who are members of a single language minority and are LEP. Suppose the ACS was a simple random sample; then a direct estimate of P is the sample proportion m/n, where n is the sample count of voting age citizens in a district, and m is the number of minority voting age citizens in that district who are LEP. For a small District A with $n = 105$, $m = 5$, $m/n < 0.05$, and the 5% provision would not apply, but for a District B with $n = 105$, $m = 6$, $m/n > 0.05$ and the 5% provision would apply. That is, a change in the sample count of just one changes the outcome. A small area model is applied to increase the precision of the estimate, and hence the reliability of the outcome.

The approach to the "more than 5%" provision was to build a district level regression model to predict P based on variables in the ACS, and Census 2010 counts of minority groups. Classify districts into classes with similar predicted P based on the model – predictive mean stratification; and then within classes, apply a hierarchical random-effects model that pulls the direct ACS estimate of P towards the average P for districts in that class; and compare the model estimate with 5% for this aspect of the determination. Comparison of the Bayesian model estimates with the direct ACS estimates indicated large gains in precision, particularly for the small voting districts. The predictive mean stratification is used to reduce dependency on model assumptions, since the regression model is used to group similar jurisdictions rather than to create direct predictions. See Joyce et al. (2012) for more details.

6. Conclusions

I have argued for a paradigm shift in official statistics, away from the current DMC towards Bayesian models that are geared to yield inferences with good frequentist properties. My design-based statistical colleagues raise two principal objections to this viewpoint.

First, the idea of an overtly model-based, even worse Bayesian, approach to probability surveys is not well received, although the calibrated part of CB is welcomed for its focus on good randomization properties. Models are mistrusted, and should be avoided at all costs! My view is simply that classical design-based methods do not provide the comprehensive approach needed for the complex problems that increasingly arise in official statistics: small area estimation, nonresponse and response errors, file linkage and combining information across probabilistic and nonprobabilistic sources. Judicious

choices of well-calibrated models are needed to tackle such problems. Attention to design features and objective priors can yield Bayesian inferences that avoid subjectivity, and modeling assumptions are explicit, and hence capable of criticism and refinement.

The second objection is that Bayesian methods are too complex computationally for the official statistics world, where large numbers of routine statistics need to be computed correctly and created in a timely fashion. It is true that current Bayesian computation may seem forbidding to statisticians familiar with simple weighted statistics and replicate variance methods. Sedransk (2008), in an article strongly supportive of Bayesian approaches, points to the practical computational challenges as an inhibiting feature. I agree that much work remains to meet this objection, but I do not view it as insuperable. Research on Bayesian computation methods has exploded in recent decades, as have our computational capabilities. To take as an example my research area of missing data, methods have evolved from simple imputation methods, to maximum likelihood for general patterns of missing data via iterative algorithms like EM, to Bayesian multiple imputation methods for increasingly complex models based on Gibbs' sampling, now widely available in standard software (Little and Rubin 2002; Little 2011). Bayesian models have been fitted to very large and complex problems, in some cases much more complex than those faced in the official statistics world.

Part of the problem here is a lack of familiarity with modeling and Bayesian methods among government statisticians, since unfamiliar tasks are often easier than they seem. Clearly government statisticians need to be skilled in statistical computation, a better marriage is needed between computer science and statistics, and infrastructure is needed to bring more sophisticated analysis methods into production environments. These are challenging problems, but I do not see them as insuperable, if there is recognition that they are worth tackling.

The move to a more overt modeling approach means that government agencies need to recruit and train statisticians who are adept in modeling (and yes, Bayesian) methods, as well as being familiar with survey sampling design. Survey sampling needs to be considered a part of mainstream statistics, in which Bayesian models that incorporate complex design features play a central role. A CB philosophy would improve statistical output, and provide a common philosophy for statisticians and researchers in substantive disciplines such as economics and demography. A strong research program within government statistical agencies, including cooperative ties with statistics departments in academic institutions, would also foster examination and development of the viewpoints advanced in this article (Lehtonen et al. 2002, Lehtonen and Särndal 2009).

Change is also needed before statisticians are recruited into government agencies. Currently Bayesian statistics is absent or "optional" in many programs for training MS statisticians, and even Ph.D. statisticians are often trained with very little exposure to Bayesian ideas, beyond a few lectures in a theory sequence dominated by frequentist ideas. This is clearly incompatible with the rising prominence of Bayes in science, as evidenced by the strong representation of modern-day Bayesians in science citations (Science Watch 2002).

The examples in Section 5 are for me an encouraging sign that the Census Bureau is more open to the CB approach I favor, at least in the context of small area estimation. I would like to see it applied more generally to other problems, such as the treatment of

missing data, and applications that require combining across data sources, which are becoming more urgent with the attempts to incorporate administrative record data into Census Bureau products. Aside from the statistical benefits of modeling, direct substitution of administrative records may be problematic because of privacy and legal issues, but using the administrative records as predictors in a model to impute missing records is often more acceptable (Zanutto and Zaslavsky 2001).

When it comes to consumers of statistics, Bayes is not a part of most introductory statistics courses, so most think of frequentist statistics as all of statistics, and are not aware that Bayesian inference exists. Defenders of the status quo claim that Bayesian inference is too difficult to teach to students with limited mathematical ability, but my view is that these difficulties are overrated. The basic idea of Bayes' Theorem can be conveyed without calculus, and Bayesian methods seem to me quite teachable if the emphasis is placed on interpretation of models and results, rather than on the inner workings of Bayesian calculations. Indeed, Bayesian posterior credibility intervals have a much more direct interpretation than confidence intervals, as noted above. Frequentist hypothesis testing is no picnic to teach to consumers of statistics, for that matter.

Formulating useful statistical models for real problems is not simple, and students need more instruction on how to fit models to complicated data sets. We need to elucidate the subtleties of model development. Issues include the following: (a) models with better fits can yield worse predictions; (b) all model assumptions are not equal, for example in regression lack of normality of errors is secondary to misspecification of the error variance, which is in turn secondary to misspecification of the mean structure; (c) if inferences are to be Bayesian, more attention needs to be paid to the difficulties of picking priors in high-dimensional complex models, objective or subjective.

Models are imperfect idealizations, and hence need careful checking; this is where frequentist methods have an important role. These methods include Fisherian significance tests of null models, diagnostics that check the model in directions that are important for the target inferences, and model-checking devices like posterior predictive checking and cross-validation. Such diagnostics are well known for regression, but perhaps less developed and taught for other models, particularly when complex survey designs are involved.

7. References

Agresti, A. and Coull, B.A. (1998). Approximate is Better than 'Exact' for Interval Estimation of Binomial Proportions. The American Statistician, 52, 119–126.

Basu, D. (1971). An Essay on the Logical Foundations of Survey Sampling, Part 1. In Foundations of Statistical Inference, V. Godambe and D. Sprott (eds). Toronto: Holt, Rinehart and Winston, 203–242.

Binder, D.A. (1982). Non-parametric Bayesian Models for Samples from Finite Populations. Journal of the Royal Statistical Society, 44, 388–393.

Birnbaum, A. (1962). On the Foundations of Statistical Inference (with discussion). Journal of the American Statistical Association, 57, 269–326.

Box, G.E.P. (1980). Sampling and Bayes Inference in Scientific Modelling and Robustness (with discussion). Journal of the Royal Statistical Society, Series A, 143, 383–430.

Brewer, K.R.W. (1979). A Class of Robust Sampling Designs for Large-Scale Surveys. Journal of the American Statistical Association, 74, 911–915.

Brewer, K.R.W. and Mellor, R.W. (1973). The Effect of Sample Structure on Analytical Surveys. Australian Journal of Statistics, 15, 145–152.

Brown, L.D., Cai, T.T., and DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. Statistical Science, 16, 101–133.

Cassel, C.-M., Särndal, C.-E., and Wretman, J.H. (1977). Foundations of Inference in Survey Sampling. New York: Wiley.

Chen, Q., Elliott, M.R., and Little, R.J.A. (2010). Bayesian Penalized Spline Model-Based Inference for a Finite Population Proportion in Unequal Probability Sampling. Survey Methodology, 36, 23–34.

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

Cox, D.R. (1971). The Choice between Alternative Ancillary Statistics. Journal of the Royal Statistical Society, Series B, 33, 251–255.

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

Dawid, A.P. (1982). The Well-Calibrated Bayesian. Journal of the American Statistical Association, 77, 605–610.

Draper, D. (1995). Assessment and Propagation of Model Uncertainty (with discussion). Journal of the Royal Statistical Society, Series B, 57, 45–97.

Draper, D. and Krnjajic, M. (2010). Calibration Results for Bayesian Model Specification. Bayesian Analysis, 1, 1–43.

Dumouchel, W.H. and Duncan, G.J. (1983). Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples. Journal of the American Statistical Association, 78, 535–543.

Efron, B. (1986). Why Isn't Everyone a Bayesian? The American Statistician, 40, 1–11, (with discussion and rejoinder).

Elliott, M. and Little, R.J. (2005). A Bayesian Approach to Census 2000 Evaluation Using A.C.E. Survey Data and Demographic Analysis. Journal of the American Statistical Association, 100, 380–388.

Ericson, W.A. (1969). Subjective Bayesian Models in Sampling Finite Populations. Journal of the Royal Statistical Society, Series B, 31, 195–234.

Ericson, W.A. (1988). Bayesian Inference in Finite Populations. In Handbook of Statistics 6, P.R. Krishnaiah and C.R. Rao (eds). Amsterdam: North-Holland, 213–246.

Fay, R.E. (1996). Alternative Paradigms for the Analysis of Imputed Survey Data. Journal of the American Statistical Association, 91, 490–498.

Fay, R. and Herriot, R. (1979). Estimates of Income for Small Places: an Application of James-Stein Procedures to Census Data. Journal of the American Statistical Association, 366, 269–277.

Fienberg, S.E. (2011). Bayesian Models and Methods in Public Policy and Government Settings. Statistical Science, 26, 212–226.

Firth, D. and Bennett, K.E. (1998). Robust Models in Probability Sampling. Journal of the Royal Statistical Society, Series B, 60, 3–21.

Gelman, A. (2006). Prior Distributions for Variance Parameters in Hierarchical Models. Bayesian Analysis, 1, 515–533.

Gelman, A. (2007). Struggles with Survey Weighting and Regression Modeling. Statistical Science, 22, 153–164 (with discussion and rejoinder).

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

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies (with discussion). Statistica Sinica, 6, 733–807.

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

Groves, R.M. and Lyberg, L. (2010). Total Survey Error: Past. Present, and Future. Public Opinion Quarterly, 74, 934–955.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sampling Survey Methods and Theory, 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. Journal of the American Statistical Association, 78, 776–793 (with discussion).

Heckman, J. (1976). The Common Structure of Statistical Models of Truncation. Sample Selection and Limited Dependent Variables, and a Simple Estimator for such Models. Annals of Economic and Social Measurement, 5, 475–492.

Holt, D. and Smith, T.M.F. (1979). Poststratification. Journal of the Royal Statistical Society, Series A, 142, 33–46.

Holt, D., Smith, T.M.F., and Winter, P.D. (1980). Regression Analysis of Data from Complex Surveys. Journal of the Royal Statistical Society, Series A, 143, 474–487.

Horvitz, D.G. and Thompson, D.J. (1952). A Generalization of Sampling without Replacement from a Finite Universe. Journal of the American Statistical Association, 47, 663–685.

Isaki, C.T. and Fuller, W.A. (1982). Survey Design under the Regression Superpopulation Model. Journal of the American Statistical Association, 77, 89–96.

Joyce, P.M., Malec, D., Little, R.J., and Gilary, A. (2012). Statistical Modeling Methodology for the Voting Rights Act Section 203 Language Assistance Determinations. Center for Statistical Research and Methodology, Research and Methodology Directorate Research Report Series (Statistics \#2012-02). U.S. Census Bureau. Available online at http://www.census.gov/srd/papers/pdf/rrs2012-02.pdf.

Kalton, G. (2002). Models in the Practice of Survey Sampling (Revisited). Journal of Official Statistics, 18, 129–154.

Kim, J.K., Brick, J.M., Fuller, W.A., and Kalton, G. (2006). On the Bias of the Multiple-Imputation Variance Estimator in Survey Sampling. Journal of the Royal Statistical Society, Series B, 68, 509–521.

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

Kish, L. and Frankel, M.R. (1974). Inferences from Complex Samples (with discussion). Journal of the Royal Statistical Society, Series B, 36, 1–37.

Konijn, H.S. (1962). Regression Analysis in Sample Surveys. Journal of the American Statistical Association, 57, 590–606.

Lehtonen, R., Pahkinen, E., and Särndal, C.-E. (2002). Research and Development in Official Statistics and Scientific Co-operation with Universities: An Empirical Investigation. Journal of Official Statistics, 18, 87–110.

Lehtonen, R. and Särndal, C.-E. (2009). Research and Development in Official Statistics and Scientific Co-operation with Universities: A Follow-Up Study. Journal of Official Statistics, 25, 467–482.

Little, R.J. (1983). Estimating a Finite Population Mean from Unequal Probability Samples. Journal of the American Statistical Association, 78, 596–604.

Little, R.J. (1991). Inference with Survey Weights. Journal of Official Statistics, 7, 405–424.

Little, R.J. (1993). Post-Stratification: a Modeler's Perspective. Journal of the American Statistical Association, 88, 1001–1012.

Little, R.J. (2003). The Bayesian Approach to Sample Survey Inference. In Analysis of Survey Data, R.L. Chambers and C.J. Skinner (eds). New York: Wiley, 49–57.

Little, R.J. (2004). To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling. Journal of the American Statistical Association, 99, 546–556.

Little, R.J. (2006). Calibrated Bayes: A Bayes / Frequentist Roadmap. The American Statistician, 60, 213–223.

Little, R.J. (2007). Comment on "Model-assisted Estimation of Forest Resources with Generalized Additive Models" by Jean D. Opsomer, F. Jay Breidt, Gretchen G. Moisen, and Goran Kauerman. Journal of the American Statistical Association, 102, 412–414.

Little, R.J. (2008). Weighting and Prediction in Sample Surveys. Diamond Jubilee volume. Calcutta Statistical Association Bulletin, 60, 1–47 (with discussion and rejoinder).

Little, R.J. (2011). Calibrated Bayes, for Statistics in General, and Missing Data in Particular (with Discussion and Rejoinder). Statistical Science, 26, 162–186.

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

Meng, X.-L. (1994). Multiple Imputation Inferences with Uncongenial Sources of Input (with discussion). Statistical Science, 9, 538–573.

Opsomer, J.D., Breidt, F.J., Moisen, G.G., and Kauerman, G. (2007). Model-Assisted Estimation of Forest Resources with Generalized Additive Models. Journal of the American Statistical Association, 102, 400–416 (with discussion).

Peers, H.W. (1965). On Confidence Points and Bayesian Probability Points in the Case of Several Parameters. Journal of the Royal Statistical Society, Series B, 27, 9–16.

Pfeffermann, D. (1993). The Role of Sampling Weights when Modeling Survey Data. International Statistical Review, 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. Journal of the Royal Statistical Society, Series A, 148, 268–278.

Raghunathan, T.E., Xie, D., Schenker, N., Parsons, V., Davis, W., Rancourt, E., and Dodd, K. (2007). Combining Information from Multiple Surveys for Small Area Estimation: A Bayesian Approach. Journal of American Statistical Association, 102, 474–486.

Rao, J.N.K. (1996). On Variance Estimation with Imputed Data. Journal of the American Statistical Association, 91, 499–506.

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

Rao, J.N.K. (2011). Impact of Frequentist and Bayesian Methods on Survey Sampling Practice: A Selective Appraisal. Statistical Science, 26, 240–256.

Robins, J.M. and Wang, N. (2000). Inference for Imputation Estimators. Biometrika, 87, 113–124.

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

Rubin, D.B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. Journal of Educational Psychology, 66, 688–701.

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. Journal of the American Statistical Association, 78, 803–805.

Rubin, D.B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. Annals of Statistics, 12, 1151–1172.

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

Rubin, D.B. (1996). Multiple Imputation After 18+ Years. Journal of the American Statistical Association, 91, 473–489.

SAIPE (2011). U.S. Census Bureau Small Area Income and Poverty Estimates Program. Available at http://www.census.gov//did/www/saipe/

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

Schenker, N. and Raghunathan, T.E. (2007). Combining Information from Multiple Surveys to Improve Measures of Health. Statistics in Medicine, 26, 1802–1811.

Science Watch (2002). Vital Statistics on the Numbers Game: Highly Cited Authors in Mathematics, 1991-2001. Science Watch, 13(3), 2.

Scott, A.J. (1977). Large-Sample Posterior Distributions for Finite Populations. Annals of Mathematical Statistics, 42, 1113–1117.

Scott, A.J. and Smith, T.M.F. (1969). Estimation in Multistage Samples. Journal of the American Statistical Association, 64, 830–840.

Sedransk, J. (2008). Assessing the Value of Bayesian Methods for Inference about Finite Population Quantities. Journal of Official Statistics, 24, 495–506.

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: Oxford University Press, 437–451.

Sugden, R.A. and Smith, T.M.F. (1984). Ignorable and Informative Designs in Survey Sampling Inference. Biometrika, 71, 495–506.

Thompson, M.E. (1988). Superpopulation Models. Encyclopedia of Statistical Sciences, 1, 93–99.

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

Welch, B.L. (1965). On Comparisons between Confidence Point Procedures in the Case of a Single Parameter. Journal of the Royal Statistical Society, Series B, 27, 1–8.

Yuan, Y. and Little, R.J. (2007). Model-Based Estimates of the Finite Population Mean for Two-Stage Cluster Samples with Unit Nonresponse. Journal of the Royal Statistical Society, Series C, 56, 79–97.

Yuan, Y. and Little, R.J. (2008). Model-Based Estimates of the Finite Population Mean for Two-Stage Cluster Samples with Item Nonresponse. Journal of Official Statistics, 24, 193–211.

Zanutto, E. and Zaslavsky, A. (2001). Using Administrative Records to Impute for Nonresponse. In Survey Nonresponse, R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: Wiley.

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, 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. Journal of Official Statistics, 21, 1–20.

Received October 2011
Revised May 2012