Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Bayes, buttressed by design-based ideas, is the best overarching paradigm for sample survey inference

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

by Roderick J. Little

Release date: December 15, 2022

Statistics Canada Statistique Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2022
Vol. 48, No. 2, pp. 257-281
Statistics Canada, Catalogue No. 12-001-X

Bayes, buttressed by design-based ideas, is the best overarching paradigm for sample survey inference

Roderick J. Little[1]

Abstract

Conceptual arguments and examples are presented suggesting that the Bayesian approach to survey inference can address the many and varied challenges of survey analysis. Bayesian models that incorporate features of the complex design can yield inferences that are relevant for the specific data set obtained, but also have good repeated-sampling properties. Examples focus on the role of auxiliary variables and sampling weights, and methods for handling nonresponse. The article offers ten top reasons for favoring the Bayesian approach to survey inference.

Key Words: Calibrated Bayes inference; Design-based inference; Penalized splines; Post-stratification; Probability proportional to size sampling; Proxy pattern-mixture models; Response propensity; Super-population models; Survey weighting.

1. Introduction

Bayesian inference is in my view the best overarching inferential paradigm for statistical inference from surveys, whether from probability or non-probability samples. See for example Ericson (1969), Binder (1982), Rubin (1987), Ghosh and Meeden (1997), Little (2003ab, 2004, 2012, 2015), Sedransk (2008) and Fienberg (2011). However, design-based properties of Bayesian inferences are important, because "all models are wrong", and broad acceptance of results requires inferences that have good operating characteristics in repeated sampling. In particular, Bayesian models need to incorporate complex design features to yield inferences that are approximately calibrated, in the sense that credible intervals have close to nominal levels when treated as confidence intervals in repeated sampling (Rubin, 1984, 2019; Little, 2006). In large samples, flexible working models can avoid strong parametric assumptions that lead to potentially biased estimates.

To focus discussion, consider the problem of deriving a point estimate q of a finite population quantity $Q,$ and a 95% interval estimate $I_{0.95} = (l, u)$ that captures uncertainty in $q,$ the interval may have a frequentist interpretation as a 95% confidence interval, or a Bayesian interpretation as a 95% posterior credible interval for $Q.$ I think scientists who are not statisticians generally interpret the interval I in a Bayesian way, as a fixed interval capturing the uncertainty about $Q.$ However, I do not focus unduly on the difference in interpretation of $I_{0.95}$ under the two paradigms. The 95% nominal value is by convention and other levels could be chosen.

An appealing feature of finite population survey sampling is that it deals with real (though unknown) quantities. For "analytic" survey inference, where the focus is on parameters of idealized models of the population, such as regression coefficients in a multiple regression model, define the finite population

[1] Roderick J. Little, Department of Biostatistics, University of Michigan. E-mail: rlittle@umich.edu.

quantity Q as the estimate of the parameter of interest if the model was fitted to data for the whole population, according to some agreed fitting method such as least squares or maximum likelihood (ML). A useful feature of this construction is that Q is then a real quantity, rather than a feature of a simplified hypothetical model of the population (e.g., Little, 2004).

Under the Bayesian paradigm, inference for Q is based on its posterior predictive distribution given the data, for judicious choices of model and prior distribution for unknown parameters. Thus q may be the posterior predictive mean of Q, and $I_{0.95}$ the 2.5th to 97.5 percentile of the posterior predictive distribution, or the limits of the range of values of Q with the highest posterior density, assuming the posterior predictive distribution is unimodal. A useful feature of the Bayesian approach is that "finite population corrections" are automatically incorporated in the posterior predictive distribution of finite population quantities – as the sample converges to the finite population, the posterior variance tends to zero.

The focus is on developing suitable models and prior distributions. Computation used to be a major challenge and is still a practical consideration, though less so now with the advent of Markov Chain Monte Carlo methods and rapid advances in Bayesian computation. Thus, the complaint that Bayes is conceptually appealing but simply too difficult to implement is harder to sustain than it was, say, thirty years ago.

The remainder of the paper is organized as follows. In Section 2, I introduce some notation and describe formally the models and prior distributions required by the Bayesian approach to survey data, with and without nonresponse. In Section 3, I describe generally desirable features of an inference about Q, and discuss why I believe the Bayesian paradigm for suitably-chosen models can be more successful at achieving these features than the design-based approach. In Section 4 I present a variety of examples, intended to illustrate the points in Section 3. I conclude in Section 5 by proposing ten reasons to be Bayesian in the survey sampling setting.

2. Notation, and a seminal paper

In this section, I introduce some notation and a seminal paper that underlies much of the thinking in this paper. Let $Y = (y_1, \ldots, y_N)$ and $S = (S_1, \ldots, S_N)$ where $N < \infty$ is the number of units in the population, y_i is the set of survey variables and S_i is the selection indicator for the i^{th} unit, with value 1 when the i^{th} unit is selected and 0 otherwise. Let Z represent design information such as stratum or cluster indicators, and z_i the value of Z for unit i. Consider inference about a finite population quantity $Q(Y, Z)$, for example the population total $Q(Y, Z) = \sum_{i=1}^{N} y_i$, where $Y = (y_1, \ldots, y_N)$. A general model-based approach treats both S and Y as random variables, with joint distribution given Z:

$$f_{S,Y|Z}(S, Y | Z, \theta, \psi) = f_{Y|Z}(Y | Z, \theta) f_{S|Y,Z}(S | Z, Y, \psi), \tag{2.1}$$

where $f_{Y|Z}$ represents the density of survey variables Y indexed by unknown parameters $\theta,$ and $f_{S|Y,Z}$ represents the model for inclusion indexed by unknown parameters $\psi.$ For a probability sample with no nonresponse, the sampling distribution is known and does not depend on $Y,$ that is,

$$f_{S|Y,Z}\left(S\,|\,Z,Y,\psi\right) = f_{S|Z}\left(S\,|\,Z\right); \tag{2.2}$$

design-based methods base inferences on the distribution of statistics in repeated sampling from this distribution.

For a survey with unit nonresponse, inclusion occurs when a unit is selected, and then responds given selection. Accordingly, let $R_i = 1$ if selected unit i responds and $R_i = 0$ otherwise. The model-based approach models the joint distribution of $S,$ R and Y given Z as

$$f_{S,R,Y|Z}(S,R,Y\,|\,Z,\theta,\psi) = f_{Y|Z}\left(Y\,|\,Z,\theta\right) f_{S|Y,Z}\left(S\,|\,Z,Y,\psi\right) f_{R|S,Y,Z}\left(R\,|\,Z,Y,S,\phi\right), \tag{2.3}$$

adding to equation (2.1) a model for unit nonresponse with density $f_{R|S,Y,Z}.$ Item nonresponse can also be treated by modeling indicators for the patterns of item missingness (e.g., Little, 2003b).

Treating $S,$ R and Y as random variables is a key feature of Rubin (1978), which I regard as one of the landmark statistics papers in the history of statistics. The paper provides conditions under which the missingness and selection mechanisms are ignorable, that is, do not need to be modeled for likelihood-based inference, extending definitions of ignorability for missing data in Rubin (1976), while providing a framework for inference when selection and/or missingness is non-ignorable. The significance of the paper for survey sampling is easily missed, because its main focus is on the role of the treatment assignment mechanism in the context of inference about causal effects. The assignment mechanism is ignorable under random treatment assignment, as in randomized clinical trials. The paper thus lays a general framework for causal inferences comparing treatments, and it is for this feature that the paper is best known. However, the paper also provides a Bayesian justification for random sampling, as a means of avoiding the need for a model for selection.

In frequentist superpopulation modeling (e.g., Valliant, Dorfman and Royall, 2000), the parameters in models are treated as fixed; in Bayesian survey modeling, these parameters are assigned a prior distribution, and inferences for $Q(Y)$ are based on its posterior predictive distribution given the data. In large samples, the prior distribution plays a minor role, and the two approaches yield similar answers for comparable models; in particular the ML estimate of a parameter is essentially the mode of the posterior distribution under a uniform prior, and as such has a Bayesian interpretation. In small samples, uncertainty about the model parameters is propagated when they are integrated out of the posterior distribution. This approach to propagating error in parameters allows Bayesian inferences for judiciously chosen models and priors to be better calibrated than inferences from superpopulation modeling inferences, in a sense of having better frequentist properties in repeated sampling (Rubin, 1978). So, in my opinion "superpopulation modeling is super, but Bayes is better".

3. Design-based versus model-based inference

The survey sampling literature features many lively controversies (e.g., Smith, 1976, 1994; Kish, 1995; Brewer, 2013; Little, 2014) between "design-based" inference, where inference is based on the sampling distribution (2.2) and "model-based" inference, where inference is based on model distribution $f_{Y|Z}(Y|Z, \theta)$ if selection is ignorable, or on the full model distribution (2.1) or (2.3) if selection is nonignorable or there is nonresponse. I seek inferences that are both design-based and model-based, in that they are based on Bayesian models but have good design-based properties.

Rubin (1984) distinguished between *statistical inference for a particular data set*, and the *properties* of that inference – consistency, confidence coverage – in repeated sampling. To be broadly credible, the inference should also have good repeated-sampling properties. This goal of the "design-based" approach should also be a goal of Bayesian models for surveys – the model and prior should be chosen to yield inferences with good design-based properties. To achieve this, features of the complex probability sample design need to be part of the model – stratification and weighting incorporated via covariates, multistage sampling incorporated via hierarchical models. The inclusion of a prior distribution in Bayesian modeling, decried by some as yet another assumption to be added to the model, for me provides an additional tool over superpopulation modeling. It provides more flexibility than superpopulation modeling, which effectively restricts the choice to uniform priors.

In addition to having good frequentist properties, the inference based on q and I needs to be appropriate – Rubin (2019) uses the term "relevant" – for the realized data set. Let D denote the data that are the basis for the inference, and $\tilde{D}$ the particular realization of D, the sample and respondent values actually obtained. Whether q and I are derived from a formal Bayesian model, an estimating equation, or some algorithmic procedure, they should provide good inference for the data $\tilde{D}$, not other data sets D that may have been obtained. Bayesian methods tend to have this property, because the posterior distribution conditions on $\tilde{D}$; but a confidence interval should also be approximately valid when viewed as a credible interval that conditions on $\tilde{D}$, if only because this is how a non-statistician tends to interpret it. Design-based confidence intervals can be lacking from this perspective, as illustrated in examples 2 and 4 below.

To summarize, a common goal of design-based and model-based inference is to arrive at a value of q that makes efficient use of the data, has some property like design consistency which implies that it is not too far from Q, and an interval I that is as narrow as the information in the data allows, while including Q with a probability close to the nominal 95% value. Rubin (2019) associates these properties with a "sage" statistician in his Waksberg lecture.

Design-based methods are often rationalized as avoiding the need for a model, because properties like design consistency are not based on a model for the data. However, the performance of design-based methods often depends on an implicit model, and modifying the estimate based on a more realistic model can improve the inference, from a design-based or model-based perspective. This point is illustrated in examples 3-5 below.

The question of the appropriate reference set for repeated sampling properties like confidence intervals is fraught with difficulties, specifically on whether to condition on ancillary statistics or on statistics that are close to ancillary (e.g., Birnbaum, 1962; Berger and Wolpert, 1988; Ghosh, Reid and Fraser, 2010). These questions also arise when assessing the repeated-sampling properties of a Bayesian inference, but do not apply to the inference itself because the posterior distribution conditions on $\tilde{D}$.

The design-based approach to sample survey inference is too limited in scope, failing to address adequately many of the problems of sample surveys in practice. Limitations include the following:

1. Design-based inference is asymptotic, and does not provide valid inferences in small samples. Consider the following simple example.

   Example 1. Inference about a population mean from a simple random sample. Consider inference about a population mean of a variable Y from a simple random sample of size n from a population of size N. The standard design-based 95% confidence interval takes the form

   $$\bar{y} \pm 1.96 s \sqrt{(1/n - 1/N)}, \tag{3.1}$$

   where $\bar{y}$ is the sample mean and s is the sample standard deviation. This interval is asymptotic and does not provide valid small sample inferences. In particular, if Y is continuous, better inference is usually obtained by replacing 1.96, the 97.5th percentile of the normal distribution, by the 97.5th percentile of a distribution that reflects uncertainty about estimating the variance, such as the t distribution with $n-1$ degrees of freedom. However, that procedure assumes a normal distribution for Y and hence is not design-based. If Y is binary with values 0 and 1, then $\bar{y}$ is the sample proportion, $s = \sqrt{\bar{y}(1-\bar{y})}$ and (3.1) is the asymptotic Wald interval, which performs very poorly in small samples, particularly when the true proportion is near to 0 or 1. The Bayesian credible interval for a Jeffreys or uniform prior has much better frequentist properties. See Dean and Pagano (2015) and Franco, Little, Louis and Slud (2019) for comparisons of Wald intervals with alternatives for complex designs. Design-based inference is often a poor option for small samples, and in particular for small area estimation, where a model for Y is invoked to "borrow strength" across areas.

2. Design-based inference does not handle survey unit or item nonresponse or response errors, because these problems require models to yield generally satisfactory results.

3. Design-based inference is not prescriptive, in a sense of prescribing the appropriate choice of inference method for the data at hand. The appropriate choice of estimator effectively requires an implicit model, as in "model-assisted" estimation (e.g., Särndal, Swensson and Wretman, 1992). For example, the regression or ratio estimator for incorporating auxiliary information, or the Horvitz-Thompson or Hájek estimator for incorporating survey weights, are all based on implicit models, and if that model is far from realistic these methods may be severely suboptimal – Basu's (1971) elephants being an extreme and satirical example. Bayesian inference based on more flexible models tend to do better, as discussed in example 5 below.
4. Design-based inference does not address how to provide inferences for non-probability samples, which are increasingly prevalent given the expense and difficulty of obtaining true random samples.

Point 4 does not rule out the use of design-based methods for deriving q and I, because we can always pretend that we have a probability sample, by assuming a model for the selection indicators S that describe inclusion in the sample, and estimating the unknown parameters in that model (e.g., Elliott and Valliant, 2017). Statisticians who use design-based methods for inference from random samples tend to favor this "quasi-randomization" approach. However, it shares the limitations of the design-based approach for probability samples, namely the inability to handle small samples, missing data or response errors; the Bayesian toolkit for these problems is much more extensive.

4. Examples

A variety of examples are offered, admittedly somewhat slanted towards my own work with colleagues. I avoid topics such as small area or time series estimation, because the need for modeling is well established there.

Example 1 continued. Normal random sample. A common critique of model-based methods is "if I base an inference on a model and the model is wrong, the inference must be wrong. Because (to paraphrase George Box), all models are wrong, therefore all model-based inference is wrong. So I prefer design-based inference, which does not require modeling assumptions". The reasoning is plausible but it's not that simple. The validity of model-based methods depends both on the design and on the degree and nature of model misspecification; the fact that design-based methods do not overtly depend on models does not mean they are necessarily superior to methods that do.

Suppose a simple random sample of size n is taken from a continuous distribution with unknown mean μ and standard deviation $\sigma.$ Consider three interval estimates for the population mean:

(i) Interval A is the standard design-based 95% confidence interval, equation (3.1).

(ii) Interval B replaces the 97.5th normal percentile in equation (3.1), namely 1.96, by the 97.5th percentile of the t distribution with $n-1$ df.

(iii) Interval C is the 95% posterior credible interval based on a normal model with the Jeffreys' prior distribution $p(\mu, \sigma) \propto 1/\sigma$.

From a frequentist perspective, which of these intervals is the best? Interval A makes no distributional assumption on Y, and B makes a normal assumption – does that mean that A is superior? If n is large then the intervals are essentially the same, and if n is small then B is arguably better than A even if the data are not normal, because it is reflecting uncertainty about the variance.

In an informal survey, two thirds of a recent class of our well-trained Ph.D. students preferred B to C, on the grounds that it avoids the choice of prior distribution and hence makes fewer assumptions. But B and C are equally good or bad, because they are the same procedure! That this interval is an exact 95% confidence interval under normality, and is a 95% credible interval for the stated choice of prior, are just two properties of the procedure. Judging a method by its overt assumptions is an over-simplification.

Example 2. Simple random sample with a lower bound on the variance (Little, 2006). Suppose in the previous example that $n = 7$, $\bar{y} = 1$, $s = 1$. The standard t interval (ignoring finite population corrections) is

$$\bar{y} \pm t_{6,975}\, s\big/\sqrt{n} \;=\; 1 \pm 2.447\big/\sqrt{7} = 1 \pm 0.925 \qquad (4.1)$$

if in fact we know that $\sigma =$ 1.5, a better interval is

$$\bar{y} \pm z_{0.975}\, \sigma\big/\sqrt{n} \;=\; 1 \pm 1.96 \times 1.5\big/\sqrt{7} = 1 \pm 1.11, \qquad (4.2)$$

the wider interval reflecting the fact that σ is greater than the particular value of s for this sample. The t interval (4.1) has exact confidence coverage, but, given what we know about σ, it is the wrong inference for this specific data set: we should not pick it over (4.2) because it is narrower!

Now suppose we know that $\sigma >$ 1.5, because of some unaccounted source of additional variation. The Bayesian approach then incorporates this information into the prior distribution for σ, resulting in a credible interval that is wider than (4.2). What is the frequentist answer? The t confidence interval still has exactly nominal coverage in repeated sampling, but it is clearly too narrow as an inference for the observed dataset, because it fails to reflect what is known about σ. The interval (4.2) is anticonservative, despite the fact that it is wider than (4.1) for the realized data set – a property that can happen for confidence intervals but cannot happen for credible intervals under a specific model. Asymptotic frequentist methods are no help here, so what is the alternative to Bayes in this example?

A related example arises in one group random-effects analysis of variance, when the least squares estimate of the between-group variance is negative – a Bayesian analysis addresses this with a prior distribution on the between-group variance that does not allow negative values. Random-effects and mixed-effects models are important to handle clustering in surveys, and Bayesian methods are better than ML in this setting.

Example 3. Post-stratification on a categorical covariate. Prediction (as in modeling) is a more reliable general approach to inference than weighting (as in design-based inference). I illustrate this general statement with the simple example of post-stratification on a single variable Z. A more complex example – Example 7 – is given later.

Consider an equal probability sample with a single categorical post-stratifying variable Z, for which known population counts N_h are available for each post-stratum h, $h = 1, 2, \ldots, H$. Let $\bar{y}_h$ be the sample mean in post-stratum h, based on sample size n_h, and $n = \sum_{h=1}^{H} n_h$, $N = \sum_{h=1}^{H} N_h$. The standard estimate of the population mean is the post-stratified mean:

$$\bar{y}_{\text{PS}} = \sum_{h=1}^{H} P_h \, \bar{y}_h, \tag{4.3}$$

where $P_h = N_h / N$ and N is the population size. This can be viewed as a weighted mean

$$\bar{y}_{\text{PS}} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_i \, y_i,$$

where $w_i = N_h / (N n_h)$ is the post-stratification weight for sampled units in poststratum h. These weights can be very large in post-strata with small sample sizes n_h, which, unlike stratified sampling based on Z, are not under the control of the sampler. These large weights can lead to excessive variability in $\bar{y}_{\text{PS}}$. In fact, strictly speaking, $\bar{y}_{\text{PS}}$ does not have a distribution in repeated sampling, because with positive probability the sample sizes in some post-strata may be zero. This remains true if the post-strata are modified to ensure that the post-strata sample counts are all positive for the observed sample, for example by pooling adjacent strata.

The standard design-based approach to excessive variability of $\bar{y}_{\text{PS}}$ is to modify the weights $\{w_i\}$, for example by trimming the large ones. However, from a prediction perspective, this is misguided. The problem is not the weights – the population proportions $\{P_h\}$ in each post-stratum are known, after all – the problem is that sparsity of sample in some post-strata renders the estimates $\{\bar{y}_h\}$ unreliable. It is the estimates in sparse post-strata that need to be modified, not the weights $\{w_i\}$ attached to sampled units. The principled way to modify $\{\bar{y}_h\}$ is to assume a model relating Y and Z. The design-based approach, by avoiding such a model, leads to the wrong principle – modifying the weights rather than the predictions of non-sampled values.

A related point: a common design-based approximation (Kish, 1992) measures the proportionate increase in variance from weighting as $1 + \text{cv}(w_i)$, where $\text{cv}(w_i)$ is the coefficient of variation of the weights; trimming the weights reduces $\text{cv}(w_i)$ and hence this proportionate increase. However, this rule of thumb is only valid when Y and Z are unrelated, in which case post-stratification is useless. If Y and Z are related and the sample size is not too small, the variance of the post-stratified mean is *smaller*, not *larger*, than the variance of the unweighted sample mean (Holt and Smith, 1979; Little and Vartivarian, 2005). The rule of thumb fails because the relationship between Y and Z is not modeled.

What is the Bayesian approach to excessive variability of $\bar{y}_{\text{PS}}$? The latter is the posterior mean for the stratified normal model

$$\left(y_{hi} \,|\, \mu_h, \sigma^2\right) \overset{\text{iid}}{\sim} G\left(\mu_h, \sigma^2\right), \tag{4.4}$$

$$p\left(\mu_h, \sigma\right) \propto 1/\sigma, \tag{4.5}$$

where $G\left(a, b^2\right)$ denotes the normal (Gaussian) distribution with mean $a,$ variance $b^2,$ and equation (4.5) is the Jeffreys' prior distribution on the mean and standard deviation in each post-stratum. Given a sparse sample in some post-strata, the prior distribution needs to be modified to allow borrowing of strength from other strata. One approach is to assume the normal random-effects model

$$p\left(\mu_h \,|\, \mu, \tau, \sigma\right) \overset{\text{iid}}{\sim} G\left(\mu, \tau^2\right),\, p\left(\mu, \sigma, \tau\right) \propto \sigma^{-1}, \tag{4.6}$$

which treats the stratum means as random effects. The variances in each post-stratum might also be treated as distinct random effects and assigned a prior distribution, rather than pooled. The posterior mean of $\bar{Y}$ for the prior distribution (4.6) moves the weight w_i of sampled units in post-stratum h towards one, with a degree of shrinkage that depends on the relative size of estimates of σ and τ (Lazzeroni and Little, 1998).

The prior distribution (4.6) makes the non-trivial assumption that the post-stratum means are exchangeable. It can be relaxed by restricting the random effects model to a subset of post-strata with small sample counts; or the constant mean μ in (4.6) might be replaced by a regression on known post-stratum characteristics $C_h,$ as in:

$$p\left(\mu_h \,|\, \beta_0, \beta_1, \tau, \sigma, c_h\right) \overset{\text{ind}}{\sim} G\left(\beta_0 + \beta_1 c_h, \tau^2\right),\, p\left(\beta_0, \beta_1, \sigma, \tau\right) \propto \sigma^{-1}, \tag{4.7}$$

which limits the exchangeability assumption to the errors in the regression of μ_h on $c_h.$ For extensive generalizations of this basic example, see Gelman and Little (1997), Elliott and Little (2000), Elliott (2007), Gelman (2007) and Si, Trangucci, Gabry and Gelman (2020).

Example 4. Regression estimator of the mean, given a population auxiliary variable. If the auxiliary variable Z in the previous example is continuous, a common way to incorporate it in the inference is via the regression estimate of the mean:

$$q = \bar{y}_{\text{REG}} = \bar{y} + \hat{\beta}_1\left(\bar{Z} - \bar{z}\right),$$

where $\hat{\beta}_1$ is the least squares estimate of the slope of Y on Z in the sample, and $\bar{z}$ and $\bar{Z}$ are respectively the sample and population mean of Z. In a simulation study of five real populations, Royall and Cumberland (1981, 1985) assess inferences centered at q, with (a) the standard design-based standard error based on simple random sampling, namely:

$$I_{0.95D} = \bar{y}_{\text{reg}} \pm 1.96\,\mathrm{s\hat{e}}_D\left(\hat{\beta}\right),\ \mathrm{s\hat{e}}_D\left(\hat{\beta}\right) = \sqrt{(1-f)\,s_{Y.Z}/n},$$

where $s_{Y.Z}^2$ is the sample residual variance and $f = n/N$ is the sampling fraction; and (b) the prediction standard error based on the normal linear regression model with constant variance, namely:

$$\begin{aligned} I_{0.95M} &= \bar{y}_{\text{reg}} \pm t_{0.975,\,n-2} 1.96\,\mathrm{s\hat{e}}_M\left(\hat{\beta}\right), \\ \mathrm{s\hat{e}}_M\left(\hat{\beta}\right) &= \mathrm{s\hat{e}}_D\left(\hat{\beta}\right)\sqrt{\left(1+\left(\bar{z}-\bar{Z}\right)^2\right)\Big/(1-f)\left(\textstyle\sum_{i=1}^{n}\left(z_i-\bar{z}\right)^2/n\right)}. \end{aligned}$$

The design-based confidence intervals $I_{0.95D}$ exhibit very poor conditional confidence coverage when the observed $\bar{z}$ deviates substantially from $\bar{Z}$. The model-based confidence interval $I_{0.95M}$ takes into account this lack of balance with respect to $\bar{z}$, but is vulnerable to model misspecification, specifically lack of linearity in the relationship between Y and Z or non-constant residual variance. Robust estimates of standard error, based on the sandwich estimator or the jackknife, yield intervals with better conditional coverage properties, although still sometimes deviating from nominal coverage levels. An alternative approach is Bayesian inference based on a more flexible model relating Y to Z, such as the penalized spline model:

$$\begin{aligned} &\left(y_i \mid z_i, \beta, \sigma^2\right) \overset{\text{ind}}{\sim} G\left(\text{spline}\left(z_i, \beta\right), \sigma^2 z_i^{\alpha}\right), \\ &\text{spline}\left(z_i, \beta\right) = \beta_0 + \sum_{j=1}^{p} \beta_j z_i^j + \sum_{\ell=1}^{m} \beta_{\ell+p}\left(z_i - \kappa_\ell\right)_+^p, \\ &\left(\beta_{l+p} \mid \tau\right) \overset{\text{iid}}{\sim} N\left(0, \tau^2\right), l = 1, \ldots, m;\ p\left(\beta_0, \ldots, \beta_p, \alpha, \sigma, \tau\right) \propto 1/\sigma, 0 < \alpha < 2, \end{aligned} \qquad (4.8)$$

where the constants $\kappa_1 < \ldots < \kappa_m$ are selected fixed knots, and $(u)_+^p = u^p$ if $u > 0$ and 0, otherwise (see, for example Ruppert, Wand and Carroll, 2003). The parameter α allows for a variety of common forms of heteroskedasticity. The Bayesian standard errors then reflect imbalance in distribution of Z in the sample and population, and the flexibility of the model limits bias from model misspecification.

Suppose that the target quantity is not the population mean of Y, but the least squares slope of Y on Z in the population. A robust approach is to impute the non-sampled values of Y using the model (4.8), and then estimate the slope of Y on Z as the least squares slope estimated on the filled-in population data. Uncertainty can be propagated by multiple imputation (Rubin, 1987), a method founded on Bayesian ideas. In this context, Little (2004) distinguishes between the "target model" that determines the target population quantity of interest, here the linear regression of Y on X, and the "working model" (4.8) that is the basis for inference, and is used to predict survey variables for the non-sampled and nonresponding units in the population. Distinguishing between these two models provides for a robust form of Bayesian survey inference.

Szpiro, Rice and Lumley (2010) apply a similar idea in a superpopulation regression setting, and Little (2019) argues that this is more straightforward than changing the interpretation of the estimand, the approach adopted by Buja, Berk, Brown, George, Pitkin, Zhan and Zhang (2019). Szpiro et al. (2010) show that the approach provides a Bayesian interpretation of the sandwich estimator of variance in regression, which is asymptotically equivalent to sample reuse estimates of variance like the bootstrap or jackknife, which are commonly applied in sample survey settings.

Example 5. Inference for samples with unequal probabilities of selection. For designs with unequal selection probabilities, classic papers critiquing the modeling approach to surveys (Kish and Frankel, 1974; Hansen, Madow and Tepping, 1983) do not include the selection probabilities in the model, yielding inferences that are vulnerable to model misspecification. The selection probabilities play an important role in robust model-based inference, but as model covariates rather than as sampling weights.

Consider, for example, inference about a population mean $\bar{Y}$. If y_i is the value of a survey variable Y and π_i is the selection probability for unit i, the usual design-based estimator of the mean of Y weights sampled units (say units $i = 1, \ldots, n$) by the inverse of π_i resulting in the Horvitz-Thompson estimate (Horvitz and Thompson, 1952)

$$\bar{y}_{\mathrm{HT}} = N^{-1} \sum_{i=1}^{n} y_i / \pi_i, \tag{4.9}$$

if the population size N is known, or the Hájek (1971) estimate

$$\bar{y}_{\mathrm{HK}} = \left(\sum_{i=1}^{n} y_i / \pi_i \right) \Bigg/ \left(\sum_{i=1}^{n} 1 / \pi_i \right), \tag{4.10}$$

if N is estimated. Weighting is a “one size fits all” approach, with sampled units receiving the same weight irrespective of the relationship between π_i and y_i. This is potentially inefficient – if the relationship is weak, weighting simply reduces the precision of the estimate without a compensating reduction in bias.

The modeling approach incorporates the selection probabilities by regressing y_i on π_i. The strength of relationship between y_i and π_i then moderates how the selection probability affects the estimator – if the relationship is weak, the regression coefficient of π_i is small and the sampling weight has little influence. This results in more efficient estimates.

A linear regression of y_i on π_i is vulnerable to bias if the linearity is misspecified, but the impact of misspecification can be reduced by choosing a model that results in a design-consistent estimate. Many models satisfy this requirement; see for example Firth and Bennett (1998).

In stratified sampling with stratifying variable Z, the natural regression model includes covariates that are dummy variables for the strata. The resulting estimate of the population mean of a continuous survey variable Y is the stratified mean, which has the same form as equation (4.3) but with Z forming strata rather than post-strata. Weighting by the inverse of the selection probability in each stratum and dummy variable regression both yield the estimator (4.3) in this situation. The Bayesian approaches to reducing variance described in example 3 tend to have less pay-off in the case of stratified sampling than for post-stratification, because the sampler has control over the sample sizes in each stratum.

In probability proportional to size (PPS) sampling, the covariate Z is the size of the unit and $\pi_i = \min\left(cz_i, 1\right)$. Now Z is a continuous variable, and weighting and regression may yield different answers. The approaches can be unified by considering models that yield the design-weighted estimates when used to predict the non-sampled units. In particular, ignoring finite population corrections, the HT estimate (4.9) is the posterior mean for the "Horvitz Thompson model":

$$\left(y_i \,|\, z_i, \beta, \sigma^2\right) \overset{\text{ind}}{\sim} G\left(\beta z_i, \sigma^2 z_i^2\right),\ p\left(\beta, \sigma\right) \propto 1/\sigma, \tag{4.11}$$

and the Hájek estimate (4.10) is the posterior mean for the "Hájek model":

$$\left(y_i \,|\, z_i, \beta, \sigma^2\right) \overset{\text{ind}}{\sim} G\left(\beta, \sigma^2 z_i\right),\ p\left(\beta, \sigma\right) \propto 1/\sigma. \tag{4.12}$$

These underlying models describe situations where the corresponding design-based estimates are optimal. However, they involve strong parametric assumptions. A robust Bayesian modeling approach embeds these models within a larger model, such as the penalized spline model (4.8), as proposed by Zheng and Little (2003). Zheng and Little (2005) and Chen, Elliott, Haziza, Yang, Ghosh, Little, Sedransk and Thompson (2017) provide simulation studies suggesting that this modeling approach can yield substantial gains over HT or Hájek estimation, both in terms of efficiency and closer to nominal (frequentist) confidence coverage in moderate samples – remember, design-based results are asymptotic. The model (4.8) is readily expanded to include other auxiliary variables measured for all the population units, and the flexibility of small-sample inferences increased by including proper prior distributions for the model parameters.

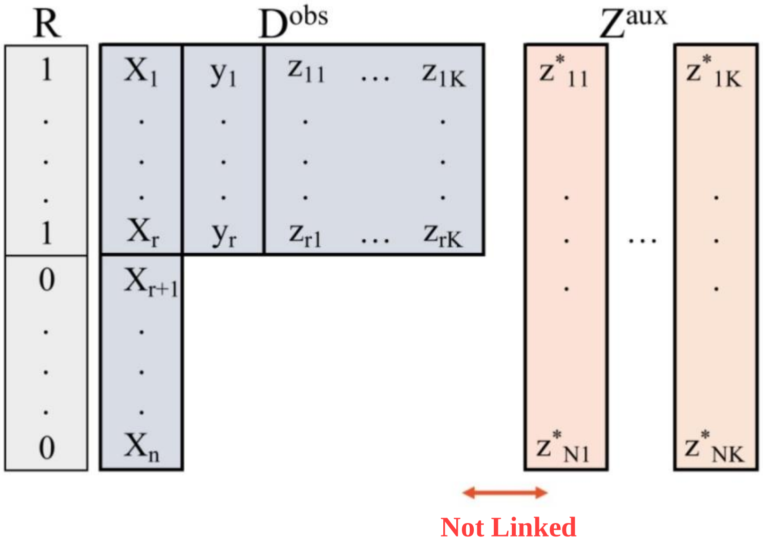
Leon-Novelo and Savitsky (2019) consider Bayesian models for the joint distribution of y_i and π_i, calling models that fix π_i a "plug-in" approach. However, assigning a distribution to π_i seems to me both unnatural and unnecessary in this setting; if π_i is recorded for all population units, it can be conditioned in the model, as in standard regression approaches that treat covariates as fixed. Even if values of π_i for non-sampled units are not provided to analysts, their values can be predicted via Bayes Theorem, as in Zangeneh and Little (2015).

Example 6. Penalized spline of response propensity models. Consider unit nonresponse on a survey variable Y, when a set of variables say $X_1, \ldots, X_p$ are observed for respondents and nonrespondents in the sample. The response propensity for unit i, $\theta_i = \Pr\left(R_i = 1 \,|\, x_{i1}, \ldots, x_{ip}, \psi\right)$, where ψ represents unknown parameters, plays an important role in both weighting and prediction approaches to survey nonresponse. In nonresponse weighting the sampling weight is multiplied by the nonresponse weight, as in

$$\begin{aligned} w_i &= 1/\Pr(\text{unit } i \text{ is selected and responds}) \\ &= (1/\Pr(i \text{ selected})) \times (1/\Pr(i \text{ responds} \mid \text{selected})) \\ &\quad \text{sampling weight} \times \text{response weight.} \end{aligned}$$

Unlike the sampling weight, the response weight is unknown, and the definition needs to be clear on what the probability is conditioned. Little (2022) argues that it should condition on auxiliary and survey variables, but not on other variables that might affect it. As with sampling weights, weighting by the inverse of the estimated response propensity is a “one size fits all” approach that does not take into account the strength of relationship between R and Y. Penalized Spline of Propensity Prediction (PSPP, Zhang and Little, 2009) regresses Y on a penalized spline of the estimated response propensity, with other variables that are observed for respondents and nonrespondents entering parametrically. The spline models a flexible relationship between the response propensity and Y, providing robustness to model misspecification. The balancing property of the propensity score also provides a double robustness property, in that the parametric form of the regression on other predictors can be misspecified without bias, provided the penalized spline captures the relationship between the propensity and Y. This method performs favorably with weighting methods in simulations (Zhang and Little, 2009; Yang and Little, 2015), and the fully Bayes version with prior distributions on the parameters propagates error in estimating the propensities.

Figure 4.1 Data structure for example 7.



Example 7. Unit nonresponse with poststratification. Another example where Bayesian and design-based approaches to inference differ concerns unit nonresponse with post-stratification. Consider the setting of Figure 4.1, where Y is a survey variable subject to nonresponse, X is a variable observed for all units in the sample, and $Z^{\text{aux}} = (Z_1, \ldots, Z_K)$ consists of a set of categorical auxiliary variables. The joint distribution of Z^{aux} is observed for survey respondents, and the marginal distributions of Z_k, $k = 1, \ldots, K$, are also observed for the population or sample from external sources. A distinctive feature is that the units in the auxiliary data are not linked with the units in the survey. This scenario occurs frequently in settings where post-stratification is used for nonresponse adjustment. Define R_i as the response indicator for (y_i, z_i) for sample unit i, and suppose that

$$\Pr(R_i = 1 | x_i, z_i, y_i, \psi) = \Pr(R_i = 1 | x_i, z_i, \psi), \tag{4.13}$$

so that this probability does not depend on y_i. Note that if this probability depends on z_i, the response mechanism is missing not at random (MNAR) according to Rubin's (1976) definition, because values of z_i are missing for survey nonrespondents. Zangeneh and Little (2022) consider Bayes and ML estimation for data with this pattern. Considering for simplicity models that are i.i.d. over the units i, the joint distribution of X, Z, Y and R is factored as

$$\begin{aligned} f_{X,Z,Y,R}(x_i, z_i, y_i, r_i | \theta, \phi) &= f_{Y|X,Z,R}(y_i | x_i, z_i, \theta, R_i = r_i)\, f_{X,Z,R}(x_i, z_i, r_i | \phi) \\ &\overset{\text{Under Eq. (4.13)}}{=} f_{Y|X,Z,R}(y_i | x_i, z_i, \theta)\, f_{X,Z,R}(x_i, z_i, r_i | \phi), \end{aligned}$$

where θ and ϕ are distinct parameters (Little and Rubin, 2019, Chapter 6). The parameters θ in this factorization can thus be estimated from the respondent survey data with $R_i = 1$, and parameters of the joint distribution of X and Z are estimated by combining the respondent and auxiliary data on those variables.

Two special cases considered by Zangeneh and Little (2022) are as follows:

(1) No covariates X and a single post-stratifier Z. The missingness assumption in equation (4.13) then reduces to $\Pr(R_i = 1 | Z_i, Y_i, \psi) = \Pr(R_i = 1 | Z_i, \psi)$. The parameters of the conditional distribution of Y given Z are estimated from the survey respondents, and the parameters of the marginal distribution of Z are estimated from the auxiliary data on Z. In particular, if Z is multinomial with $\Pr(z_i = j | \phi) = \phi_j$ the ML estimate of ϕ_j is simply the proportion of the auxiliary counts in post-stratum j. If, in addition, Y given $Z = j$ is assumed normal with mean μ_j and variance σ_j^2, the resulting ML estimate of the population mean of Y is simply the post-stratified mean given in equation (4.3). This simple example is interesting theoretically, because the response mechanism is MNAR but ignorable for likelihood inference; MAR is a sufficient but not always necessary condition for ignorability.

(2) X is a single categorical variable and Z is a single categorical post-stratifier. Now the joint distribution of X, Z, and R is not identified under a saturated model, and requires additional

assumptions to identify the model. In particular, a constrained MNAR "RAKE" model assumes that the marginal distributions of Z_1 and Z_2 are different for respondents and nonrespondents, but the odds ratios of Z_1 and Z_2 are the same for respondents and nonrespondents. This yields a just-identified model. Raking the table of respondent counts of Z_1 and Z_2 to the auxiliary margins of Z_1 and Z_2 yields ML estimates of ϕ under this RAKE model. (Little and Wu, 1991; Little, 1993). The post-stratified estimator of the mean of Y is then

$$\bar{Y}_{\text{rake}} = \sum_{j=1}^{J} \sum_{k=1}^{K} P_{jk}\left(\hat{\phi}\right) \bar{Y}_{jk}\left(\hat{\theta}\right),$$

where $P_{jk}\left(\hat{\phi}\right)$ is the proportion of the population with $X = j, Z = k$ from raking the respondent counts to the margins, and $\bar{Y}_{jk}\left(\hat{\theta}\right)$ is the estimated mean of Y given $X = j, Z = k$ from the model for Y given X and Z.

Some comments on this approach are as follows:

(1) Note again this is a MNAR model, and it is weaker than the MAR model that assumes that response depends on X but not on Z. As a result, the method tends to be less biased than alternative methods that assume MAR for consistent estimates.

(2) As was the case in example 4, the ML approach to this model does not involve modifying the weights applied to the data in cell $X = j, Z = k$ using arbitrary distance functions – the ML estimates are fully efficient under the assumed model.

(3) Sparse data in the cells can be addressed by adding proper prior distributions for the parameters and applying Bayesian methods. For example, for inference about the parameters θ of the distribution of Y given $X = j, Z = k$, one might assume a flat prior on the main effects of X and Z but a normal prior on the interactions, shrinking them towards zero. This achieves shrinkage of the cell mean $\bar{y}_{jk}$ towards the fitted means from an additive model. In frequentist terminology, this is a mixed ANOVA model with fixed main effects and random interactions.

(4) The idea of raking to the X and Z margins is familiar, but note that if X is a stratifier and Z is a post-stratifier, the standard approach is to perform one iteration of raking, first matching the X margin and then matching the Z margin: under the assumed model, raking iteratively is the correct procedure. Raking is ML here, but Bayesian forms of raking also can be used to propagate parameter uncertainty.

Example 8. Proxy pattern-mixture analysis. Proxy pattern-mixture analysis is a method for assessing nonresponse bias for the mean of a survey variable Y subject to nonresponse, when there is a set of covariates observed for nonrespondents and respondents. Historically the amount of missing data, as measured by the response rate, has been the most often-used metric for evaluating survey quality. However, response rates ignore the information contained in auxiliary covariates observed for nonrespondents. Methods based on the estimated probability of nonresponse such as the R indicator (Schouten, Cobben and Bethlehem, 2009) and q2 measure (Särndal and Lundström, 2010) do not take into account the strength of association of the survey variable of interest and the probability of response, which arguably should be factored into the assessment of bias. These measures assume MAR, but if the auxiliary variables are available for nonresponse adjustment, it is deviations from MAR that lead to bias.

Andridge and Little (2011) propose proxy pattern-mixture analysis (PPMA), a method based on a pattern-mixture model for nonresponse that combines in a simple and intuitive way the key features of nonresponse adjustment. Let Y_i denote the value of a continuous survey outcome and $Z_i = \left(Z_{i1}, Z_{i2}, \ldots, Z_{ip}\right)$ denote the values of p covariates for unit i in the sample. Only r of the n sampled units respond, so observed data consist of $\left(Y_i, Z_i\right)$ for $i = 1, \ldots, r$ and Z_i for $i = r+1, \ldots, n.$ Let R denote the response indicator, such that for unit i $R_i = 1$ if Y_i is observed and $R_i = 0$ if Y_i is missing. To reduce dimensionality, we replace Z by a single proxy variable X that has the highest correlation with Y in the respondent sample. This proxy variable can be estimated by regressing Y on Z using the respondent data, including important predictors of $Y,$ as well as interactions and nonlinear terms where appropriate. Specifically, we assume the regression model $E\left(Y \,|\, Z, R=1\right) = \alpha_0 + \alpha Z,$ and let $X = \alpha Z.$ The joint distribution of $X,$ Y and R using the following proxy pattern-mixture model, similar in form to that discussed in Little (1994):

$$\left(X, Y \,|\, R = r\right) \sim G_2\left((\mu_x^{(r)}, \mu_y^{(r)}), \Sigma^{(r)}\right)$$

$$R \sim \text{Bernoulli}(\pi)$$

$$\Sigma^{(r)} = \begin{pmatrix} \sigma_{xx}^{(r)} & \rho^{(r)}\sqrt{\sigma_{xx}^{(r)}\sigma_{yy}^{(r)}} \\ \rho^{(r)}\sqrt{\sigma_{xx}^{(r)}\sigma_{yy}^{(r)}} & \sigma_{yy}^{(r)} \end{pmatrix}$$

where N_2 denotes the bivariate normal distribution. Our interest is bias in the marginal mean of $Y,$ which can be written as $\mu_y = \pi\mu_y^{(1)} + \left(1-\pi\right)\mu_y^{(0)}.$ Andridge and Little (2011) assume that

$$\Pr\left(R = 1 | Y, X\right) \; = \; f\left(X^* + \lambda Y\right),$$

for some unspecified function f and known constant $\lambda.$ Here $X^* = X\sqrt{\sigma_{yy}^{(1)} / \sigma_{xx}^{(1)}}$ is the proxy X scaled to have the same variance as $Y,$ which aids the interpretation of λ by putting X and Y on the same scale. This mechanism is MAR when $\lambda = 0,$ and deviates increasingly from MAR as λ increases. With this assumption, the parameters are just identified and the ML estimate of the mean of $Y,$ averaging over patterns, is

$$\hat{\mu}_y \; = \; \bar{y}_1 + \left(\frac{n-r}{n}\right)\left(\frac{\lambda + \hat{\rho}}{\lambda\hat{\rho} + 1}\right)\left(\bar{x}_0 - \bar{x}_1\right),$$

where $\bar{x}_1, \bar{y}_1$ are the respondent means of X and Y and $\bar{x}_0$ is the mean of X for nonrespondents. The index of bias is then the adjustment of the sample mean $\bar{y}_1$ implied by this estimate, namely

$$\left(\frac{n-r}{n}\right)\left(\frac{\lambda+\hat{\rho}}{\lambda\hat{\rho}+1}\right)\left(\bar{x}_0-\bar{x}_1\right), \tag{4.14}$$

where $\hat{\rho}$ is the respondent sample correlation. This adjustment incorporates three key factors that affect the potential bias in a simple and intuitive manner – the nonresponse rate $(n-r)/n$, the correlation $\hat{\rho}$ between X and Y, and the deviation of $\bar{x}_0$ from $\bar{x}_1$. In particular, the index increases with the nonresponse rate and the deviation of the mean of X for respondents and nonrespondents.

There is no information about the parameter λ in the data – this is generally the case for methods that model deviations from MAR. Following Little (1994), Andridge and Little (2011) propose a sensitivity analysis, where estimates are generated for a range of values of λ between 0 and infinity, specifically 0, 1 and infinity; the choice of 0 corresponds to MAR, the intermediate choice of 1 implies the bias of Y is the same as the bias of the proxy variable X^*, and the choice of infinity is the most extreme deviation from MAR; estimates for this case have the highest variance. As λ varies between 0 and infinity, the middle factor $(\lambda+\hat{\rho})/(\lambda\hat{\rho}+1)$ varies between $\hat{\rho}$ (when $\hat{\mu}_y$ is the standard regression estimator of the mean) and $1/\hat{\rho}$ (when $\hat{\mu}_y$ is the inverse regression estimator proposed by Brown (1990). The sensitivity of the estimate to the choice of λ is small when $\hat{\rho}$ is close to 1, that is we have a strong proxy variable, and large when $\hat{\rho}$ is close to 0, that is we have a weak proxy variable. So having auxiliary variables that are good predictors of the survey outcomes is crucial.

Bayesian versions of the proxy pattern-mixture model are readily developed, allowing for propagation of error in the model parameters. Also, the model can be applied to create multiple imputations of nonrespondent values, allowing the incorporation of complex sample design elements into the index.

More recently, the model has been used to develop indices of departure from random sampling, by applying it to model selection rather than nonresponse (Little, West, Boonstra and Hu, 2020; Boonstra, Little, West, Andridge and Alvaredo-Leiton, 2021). Refinements in this work are (a) to replace the parameter λ by $\phi = \lambda / (1+\lambda)$, a better parametrization because ϕ ranges from 0 (MAR) to 1, and (b) (with some additional assumptions) to allow missingness also to depend on auxiliary variables orthogonal to X. The method has also been extended to handle binary outcomes (Andridge, West, Little, Boonstra and Alvarado-Leiton, 2019) and indices of potential bias in regression coefficients (West, Little, Andridge, Boonstra, Ware, Pandit and Alvarado-Leiton, 2021).

5. Conclusion: Ten reasons to be Bayesian for survey inference

My examples are intended to give some idea of the richness of Bayesian modeling possible for survey data, but they are far from exhaustive. Bayes is also useful in areas I have not touched on, including multistage sampling, time series, latent class and factor analysis models, measurement error, the combination of data from multiple sources, the creation of synthetic data for disclosure avoidance, and so on. I conclude by summarizing my reasons for advocating the Bayesian approach to survey inference:

1. The design-based approach is asymptotic, and too limited to handle the varied problems of inference from surveys, whether probability or non-probability based.
2. The Bayesian approach is both unified and flexible enough to handle the various problems encountered in surveys, and it includes superpopulation modeling as a form of large-sample inference.
3. Carefully-chosen Bayesian models can yield credible intervals that have good design-based properties in repeated sampling. In particular, weighting, stratification and post-stratification can be modeled via covariates, and clustering incorporated via Bayesian hierarchical models. Flexible models that incorporate design features render hybrid approaches such as model-assisted estimation (e.g., Särndal, Swensson and Wretman, 1992) unnecessary.
4. Early critiques of the modeling approach concern models that do not incorporate design-features, and hence are vulnerable to model misspecification. Such models can and should be avoided.
5. The Bayesian calculus of integrating out nuisance parameters provides inferences that have good frequentist properties in small as well as large samples.
6. The approach of being design-based for some problems and model-based for others leads to logical inconsistencies (see, for example, Little, 2012, Section 4.3); the Bayesian approach yields inferences that are unified and logically consistent.
7. The specification of prior distributions in the Bayesian approach is a strength, not a weakness, because it provides additional modeling flexibility. For some problems, weak "objective" priors yield results that parallel standard frequentist solutions. For other problems, stronger "subjective" priors provide useful answers for models that are not identified, as in MNAR nonresponse.
8. Computational challenges in the Bayesian approach have been greatly reduced by recent methodological advances and expanded computing power.
9. Modeling puts survey research in the mainstream of statistical modeling for other types of data. The particular features of survey research – complex sampling design, and the focus on finite population quantities, are well handled by the Bayesian paradigm.
10. The Bayesian approach does not negate the utility of probability sampling for design, which is enormously valuable for achieving robust inferences that limit the need for debatable assumptions concerning representativeness of the sample.

Acknowledgements

I thank the Waksberg Award committee for providing me the opportunity to present this work, and Yajuan Si and a referee for helpful comments.

References

Andridge, R.H., and Little, R.J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27, 2, 153-180.

Andridge, R.R., West, B.T., Little, R.J.A., Boonstra, P.S. and Alvarado-Leiton, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, 68, 5, 1465-1483.

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part I. In *Foundations of Statistical Inference*, (Eds., V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston, 203-242.

Berger, J.O., and Wolpert, R.L. (1988). The likelihood principle. *Institute of Mathematical Statistics Lecture Notes-Monograph Series*, 6, 1-199.

Binder, D.A. (1982). Non-parametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical Society, Series B*, 44,3, 388-393.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, 57, 269-326.

Boonstra, P.S., Little, R.J., West, B.T., Andridge, R.R. and Alvaredo-Leiton, F. (2021). A simulation study of diagnostics for bias in non-probability samples. *Journal of Official Statistics*, 37, 3, 751-769.

Brewer, K. (2013). Three controversies in the history of survey sampling. *Survey Methodology*, 39, 2, 249-262. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2013002/article/11883-eng.pdf.

Brown, C.H. (1990). Protecting against nonrandomly missing data in longitudinal studies. *Biometrics*, 46, 143-155.

Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Zhan, L. and Zhang, K. (2019). Models as approximations 1: Consequences illustrated with linear regression. *Statistical Science*, 34, 4, 580-583.

Chen, Q., Elliott, M.R., Haziza, D., Yang, Y., Ghosh, M., Little, R.J., Sedransk, J. and Thompson, M. (2017). Approaches to improving survey-weighted estimates. *Statistical Science*, 32, 2, 227-248.

Dean, N., and Pagano, M. (2015). Evaluating confidence interval methods for binomial proportions in clustered surveys. *Journal of Survey Statistics and Methodology*, 3, 484-503.

Elliott, M.R. (2007). Bayesian weight trimming for generalized linear regression models. *Survey Methodology*, 33, 1, 23-34. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2007001/article/9849-eng.pdf.

Elliott, M.R., and Little, R.J. (2000). Model-based approaches to weight trimming. *Journal of Official Statistics*, 16, 191-210.

Elliott, M.R., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32, 2, 249-264.

Ericson, W.A. (1969). Subjective Bayesian models in sampling finite populations (with discussion). *Journal of the Royal Statistical Society, Series B*, 31, 195-233.

Fienberg, S.E. (2011). Bayesian models and methods in public policy and government settings. *Statistical Science*, 26, 2, 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.

Franco, C., Little, R.J., Louis, T.A. and Slud, E.V. (2019). Comparative study of confidence intervals for proportions in complex sample surveys. *Journal of Survey Statistics and Methodology*, 7, 3, 334-364.

Gelman, A. (2007). Struggles with survey weighting and regression modeling (with discussion). *Statistical Science*, 22, 2, 153-188.

Gelman, A., and Little, T.C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 23, 2, 127-135. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997002/article/3616-eng.pdf.

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

Ghosh, M., Reid, N. and Fraser, D.A.S. (2010). Ancillary statistics: A review. *Statistica Sinica*, 20, 1309-1332.

Hájek, J. (1971). Comment on a paper by D. Basu. In *Foundations of Statistical Inference*, (Eds., V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart and Winston, 236.

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys (with discussion). *Journal of the American Statistical Association*, 78, 776-793.

Holt, D., and Smith, T.M.F. (1979). Post stratification. *Journal of the Royal Statistical Society, Series A*, 142, 1, 33-46.

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.

Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics*, 8, 183-200.

Kish, L. (1995). The hundred years' wars of survey sampling. *Statistics in Transition*, 2, 813-830. Reproduced as Chapter 1 of *Leslie Kish: Selected Papers,* (2003, Eds., G. Kalton and S. Heeringa), New York: John Wiley & Sons, Inc.

Kish, L., and Frankel, M.R. (1974). Inferences from complex samples (with discussion). *Journal of the Royal Statistical Society, Series B*, 36, 1-37.

Lazzeroni, L.C., and Little, R.J. (1998). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics*, 14, 61-78.

Leon-Novelo, L.G., and Savitsky, T.D. (2019). Fully Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 13, 1608-1645.

Little, R.J. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

Little, R.J. (1994). A class of pattern-mixture models for normal missing data. *Biometrika*, 81, 3, 471-483.

Little, R.J. (2003a). The Bayesian approach to sample survey inference. In *Analysis of Survey Data*, (Eds., R.L. Chambers and C.J. Skinner), New York: John Wiley & Sons, Inc., 49-57.

Little, R.J. (2003b). Bayesian methods for unit and item nonresponse. In *Analysis of Survey Data*, (Eds., R.L. Chambers and C.J. Skinner), New York: John Wiley & Sons, Inc., 289-306.

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, 3, 213-223.

Little, R.J. (2012). Calibrated Bayes: An alternative inferential paradigm for official statistics (with discussion). *Journal of Official Statistics,* 28, 3, 309-372.

Little, R.J. (2014). Survey sampling: Past controversies, current orthodoxies, and future paradigms. In *Past, Present and Future of Statistical Science*, COPSS 50th Anniversary Volume, (Eds., X. Lin, D.L. Banks, C. Genest, G. Molenberghs, D.W. Scott and J.-L. Wang), CRC Press.

Little, R.J. (2015). Calibrated Bayes, an inferential paradigm for official statistics in the era of big data. *Statistical Journal of the International Association of Survey Statisticians*,31, 4, 555-563.

Little, R.J. (2019). Comment on “Models as approximations 1: Consequences illustrated with linear regression” by A. Buja et al. *Statistical Science*, 34, 4, 580-583.

Little, R.J. (2022). A note about the definition of propensity weights. To appear in *Journal of Survey Statistics and Methodology*.

Little, R.J., and Rubin, D.B. (2019). *Statistical Analysis with Missing Data*, 3rd edition. New York: John Wiley & Sons, Inc.

Little, R.J., and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31, 2, 161-168. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9046-eng.pdf.

Little, R.J., and Wu, M.M. (1991). Models for contingency tables with known margins when target and sampled populations differ. *Journal of the American Statistical Association*, 86, 87-95.

Little, R.J., West, B.T., Boonstra, P.S. and Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of Survey Statistics and Methodology*, 8, 5, 932-964.

Royall, R.M., and Cumberland, W.G. (1981). The finite population linear regression estimator and estimator of its variance-an empirical study. *Journal of the American Statistical Association*, 76, 924-930.

Royall, R.M., and Cumberland, W.G. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, 80, 355-359.

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

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6, 1, 34-58.

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: John Wiley & Sons, Inc.

Rubin, D.B. (2019). Conditional calibration and the sage statistician. *Survey Methodology*, 45, 2, 187-198. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2019002/article/00010-eng.pdf.

Ruppert, D., Wand, M.P and Carroll, R.J. (2003). *Semiparametric Regression*, Cambridge University Press.

Särndal, C.-E., and Lundström, S. (2010). Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, 36, 2, 131-144. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2010002/article/11376-eng.pdf.

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

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, 1, 101-113. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10887-eng.pdf.

Sedransk, J. (2008). Assessing the value of Bayesian methods for inference about finite population quantities. *Journal of Official Statistics*, 24, 495-506.

Si, Y, Trangucci, R., Gabry, J.S. and Gelman, A. (2020). Bayesian hierarchical weighting adjustment and survey inference. *Survey Methodology*, 46, 2, 181-214. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2020002/article/00003-eng.pdf.

Smith, T.M.F. (1976). The foundations of survey sampling: A review (with discussion). *Journal of the Royal Statistical Society, Series A*, 139, 183-204.

Smith, T.M.F. (1994). Sample surveys 1975-1990: An age of reconciliation? (with discussion). *International Statistical Review*, 62, 5-34.

Szpiro, A.A., Rice, K.M. and Lumley, T. (2010). Model-robust regression and a Bayesian "sandwich" estimator. *Annals of Applied Statistics*, 4, 4, 2099-2113.

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

West, B.T., Little, R.J., Andridge, R.R., Boonstra, P.S., Ware, E.B., Pandit, A. and Alvarado-Leiton, F. (2021). Measures of selection bias in regression coefficients estimated from non-probability samples. *The Annals of Applied Statistics*, 15(3), 1556-1581.

Yang, Y., and Little, R.J. (2015). A comparison of doubly robust estimators of the mean with missing data. *Journal of Statistical Computation and Simulation*, 85, 16, 3383-3403.

Zangeneh, S.Z., and Little, R.J. (2015). Bayesian inference for the finite population total in heteroscedastic probability proportional to size samples. *Journal of Survey Statistics and Methodology*, 3, 162-192.

Zangeneh, S.Z., and Little, R.J. (2022). Likelihood based estimation of the finite population mean with post-stratification information under nonignorable nonresponse. To appear in *International Statistical Review*.

Zhang, G., and Little, R.J. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65, 3, 911-918.

Zheng, H., and Little, R.J. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics,* 19, 2, 99-117.

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.