Statistical Science
2017, Vol. 32, No. 2, 227–248
DOI: 10.1214/17-STS609

Approaches to Improving Survey-Weighted Estimates

Qixuan Chen, Michael R. Elliott, David Haziza, Ye Yang, Malay Ghosh, Roderick J. A. Little, Joseph Sedransk and Mary Thompson

Abstract. In sample surveys, the sample units are typically chosen using a complex design. This may lead to a selection effect and, if uncorrected in the analysis, may lead to biased inferences. To mitigate the effect on inferences of deviations from a simple random sample a common technique is to use survey weights in the analysis. This article reviews approaches to address possible inefficiency in estimation resulting from such weighting.

To improve inferences we emphasize modifications of the basic design-based weight, that is, the inverse of a unit's inclusion probability. These techniques include weight trimming, weight modelling and incorporating weights via models for survey variables. We start with an introduction to survey weighting, including methods derived from both the design and model-based perspectives. Then we present the rationale and a taxonomy of methods for modifying the weights. We next describe an extensive numerical study to compare these methods. Using as the criteria relative bias, relative mean square error, confidence or credible interval width and coverage probability, we compare the alternative methods and summarize our findings. To supplement this numerical study we use Texas school data to compare the distributions of the weights for several methods. We also make general recommendations, describe limitations of our numerical study and make suggestions for further investigation.

Key words and phrases: Design-based survey weights, finite population survey sampling, inclusion probability, weight modeling, weight trimming.

1. INTRODUCTION

The setting for this review paper is the analysis of data from sample surveys involving complex probability designs, potentially with auxiliary information. Using such a design may lead to a selection effect and, if uncorrected in the analysis, may lead to biased inferences. To mitigate the effect on inferences of deviations from a simple random sample, a common analytic technique is to use survey weights. Since weighting can lead to inefficient estimation, this article reviews approaches to address such inefficiency.

Qixuan Chen is Assistant Professor, Department of Biostatistics, Columbia University, New York, New York 10032, USA (e-mail: qc2138@cumc.columbia.edu). Michael R. Elliott is Professor, Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA (e-mail: mrelliot@umich.edu). David Haziza is Professor, Department of Mathematics and Statistics, Université de Montréal, Montreal, Quebec, Canada H3T 1J4, (e-mail: haziza@dms.umontreal.ca). Ye Yang is Graduate Student, University of Michigan, Ann Arbor, Michigan 48105, USA (e-mail: yeya@umich.edu). Malay Ghosh is Professor, Department of Statistics, University of Florida, Gainesville, Florida 32611, USA (e-mail: ghoshm@stat.ufl.edu). Roderick J. A. Little is Professor, Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA (e-mail: rlittle@umich.edu). Joseph Sedransk is Professor, Joint Program in Survey Methodology, University of Maryland, College Park, Maryland 20742, USA (e-mail: jxs123@cwru.edu). Mary Thompson is Professor, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1, (e-mail: methomps@uwaterloo.ca).

Here, inferences are required for finite population quantities such as means or totals in the entire population or subpopulations. A key requirement of a probability sample is that the inclusion probability of each of the samples that could be drawn is known, and each unit in the population has a nonzero chance of being included. In a landmark paper, Neyman (1934) considered stratified random sampling. For estimation, Neyman weighted sampled cases by the inverse of the stratum sampling rate; more generally, Horvitz and Thompson (1952) assigned unit i in the sample, with inclusion probability π_i, the design or sampling weight

$$d_i = 1/\pi_i. \tag{1.1}$$

In data sets from population surveys, weights attached to the units can include adjustments for unit nonresponse, and post-stratification to match the distributions of auxiliary variables with known distributions in the population. Thus, a more general form of weight is $w_i = d_i \times w_{in} \times w_{ip}$ where w_{in} is a unit nonresponse adjustment, and w_{ip} is a post-stratification adjustment. Weighting units is a convenient way to allow for the effects of differential inclusion into the sample, but resulting estimates can be very inefficient. This article reviews methods that attempt to ameliorate this inefficiency, either by modifying the weights or by model-based approaches that treat weights as covariates.

Most surveys are multipurpose surveys in the sense that information is collected on a possibly large number of characteristics of interest. In this context, the weights are generally constructed so that they may be applied to any characteristic of interest. Such weights are often referred to as multipurpose weights, discussed in detail in Haziza and Beaumont (2017). In the current paper, both multipurpose and single-purpose weights are considered. For the latter, the weights are variable specific so the resulting weights are, in general, not applicable to all characteristics of interest.

1.1 The Design-Based Perspective

Sampling weights are a key feature of the *randomization* or *design-based* approach to descriptive survey inference (e.g., Hansen, Hurwitz and Madow, 1953; Kish, 1995; Cochran, 1977), which has the following main elements. For a population U with N units, let y_i, $i = 1, \dots, N$, be the value of the survey (or outcome) variable of the ith unit. From the population U, a sample s, of size n, is selected according to a given sampling design. Let I_i, $i = 1, \dots, N$, be the inclusion indicator variable of the ith unit, with value 1 if the unit is included in the sample and 0 otherwise. Let Z represent design information, such as stratum or cluster indicators. We consider "descriptive" inference about a finite population quantity $Q(Y, Z)$, for example, the population total

$$Q(Y, Z) = t_y = \sum_{i=1}^{N} y_i,$$

where $Y = (y_1, \dots, y_N)$. In the design-based or randomization approach, inferences are based on the distribution of $I = (I_1, \dots, I_N)$, and the outcome variables $y_1, \dots, y_N$ are treated as fixed quantities. Inference involves (a) the choice of an estimator for Q, $\widehat{q} = \widehat{q}(Y_{\text{inc}}, I, Z)$, where Y_{inc} is the included part of Y; and (b) the choice of a variance estimator $\widehat{v} = \widehat{v}(Y_{\text{inc}}, I, Z)$ that is unbiased or approximately unbiased for the variance of $\widehat{q}$ 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 $\widehat{q} \pm 1.96\sqrt{\widehat{v}}$, where 1.96 is the 97.5th percentile of the standard normal distribution. For samples that are not large, and quantities Q that are functions of population totals, resampling-based confidence intervals are available for some designs (Rao and Wu, 1988; Rao, Wu and Yue, 1992).

Estimators $\widehat{q}$ are chosen to have good design-based properties, such as *design unbiasedness*: $E(\widehat{q}|Y) = Q$, or *design consistency*: $\widehat{q}/Q \to 1$, that is, $\widehat{q}$ converges to Q in probability under the sampling design and a suitable asymptotic framework (Brewer, 1979; Isaki and Fuller, 1982; Fuller, 2009).

Two weighted estimators play a central role in design-based inference, in the absence of unit nonresponse. Weighting sampled cases by the design weight, the inverse of the inclusion probability, yields the Horvitz–Thompson (Horvitz and Thompson, 1952) estimator

$$\widehat{t}_{\text{HT}} = \sum_{i=1}^{N} d_i I_i y_i, \tag{1.2}$$

which is design-unbiased for t_y. That is, $E(\widehat{t}_{\text{HT}}|Y) = t_y$ for all Y. An alternative to $\widehat{t}_{\text{HT}}$ is the Hájek estimator (Hájek, 1971):

$$\widehat{t}_{\text{HA}} = \frac{\widehat{t}_{\text{HT}}}{\widehat{N}_{\text{HT}}} N, \tag{1.3}$$

with $\widehat{N}_{\text{HT}} = \sum_{i=1}^{N} d_i I_i$, which is design-consistent for t_y. The corresponding estimators of the population mean $\overline{Y} = t_y/N$ are $\bar{y}_{\text{HT}} = \widehat{t}_{\text{HT}}/N$ and $\bar{y}_{\text{HA}} = \widehat{t}_{\text{HT}}/\widehat{N}_{\text{HT}}$.

EXAMPLE 1.1 (Stratified sampling). Assume that the finite population U is partitioned into H strata and let N_h be the size of stratum h, $h = 1, \dots, H$. Suppose a simple random sample of size n_h is selected from stratum h,without replacement. The inclusion probability of unit i in stratum h is thus $\pi_{hi} = n_h/N_h$. For this design, the HT and Hájek estimators of the population mean are identical and are equal to

$$(1.4)\quad \overline{y}_{\text{HT}} = N^{-1} \sum_{h=1}^{H} \sum_{i=1}^{N_h} (N_h/n_h) I_{hi} y_{hi} = \sum_{h=1}^{H} P_h \overline{y}_h,$$

where $P_h = N_h/N$ and $\overline{y}_h$ is the sample mean in stratum h. The standard estimator of variance is $\widehat{v} = \sum_{h=1}^{H} P_h^2 (1 - (n_h/N_h)) s_h^2/n_h$, where s_h^2 is the sample variance of y in stratum h. A 95% confidence interval for $\overline{Y}$ is $\overline{y}_{\text{HT}} \pm 1.96\sqrt{\widehat{v}}$.

EXAMPLE 1.2 (Estimating a population total from a PPS sample). In applications such as establishment surveys or auditing, a measure of size x is available for all units in the population. The larger units often contribute more to summaries of interest, and it is efficient to sample them with higher probability. In particular, for probability proportional to size (PPS) sampling, unit i with size x_i is sampled with probability cx_i, where c is chosen to yield the desired sample size; units included with certainty are removed from the pool before sampling. The HT estimator (1.2), which weights sampled units by the inverses of their inclusion probabilities, is the standard estimator of the population total in this setting.

The design-based approach does not require a statistical model for the survey outcomes, but the performance of design-based estimators varies with the validity of the assumptions about y implied by the form of the estimator; a useful guide is to assess an estimator's implied predictions of nonsampled values, and check whether they are sensible. For example, if $y_i = \beta\pi_i$ for all i, $\widehat{t}_{\text{HT}} = t_y$ for any sample s and, so $V(\widehat{t}_{\text{HT}}|Y) = 0$. More generally, consider the model:

$$(1.5)\qquad y_i \overset{\text{ind}}{\sim} N(\beta\pi_i, \sigma^2\pi_i^2),$$

where $N(\mu, \tau^2)$ denotes the normal distribution with mean μ and variance τ^2. Using (1.6) leads to predictions $\widehat{\beta}\pi_i$, where $\widehat{\beta} = n^{-1} \sum_{i=1}^{N} I_i y_i/\pi_i$. Hence, $\widehat{t}_{\text{HT}} = \sum_{i=1}^{N} \widehat{\beta}\pi_i$ is the result of using this model to predict the sampled and nonsampled values. In view of this property, we call (1.5) the HT model. Similarly, the Hájek estimator (1.3) of the total results from predictions from the model

$$(1.6)\qquad y_i \overset{\text{ind}}{\sim} N(\beta, \sigma^2\pi_i),$$

and accordingly we call (1.6) the Hájek model.

EXAMPLE 1.3 (Calibration estimators). Suppose that an L-vector of auxiliary variables $\mathbf{x} = (x_1, \dots, x_L)^\top$ is available for all the sample units and that the vector of population totals (benchmarks)

$$\mathbf{t_x} = (t_{x_1}, \dots, t_{x_L})^\top$$

is known, where $t_{x_l} = \sum_{i \in U} x_{li}$, $l = 1, \dots, L$, with U denoting the population of units. We seek calibrated weights w_{Ci} as close as possible to the design weights d_i such that the calibration constraints

$$\sum_{i \in s} w_{Ci} x_{li} = t_{x_l}$$

are satisfied; see Deville and Särndal (1992) and Särndal (2007) for detailed discussions about calibration. The resulting calibrated weights are given by $w_{Ci} = d_i g_i$ where g_i is the calibration adjustment associated with unit i. The resulting calibration estimator is

$$(1.7)\qquad \widehat{t}_C = \sum_{i \in s} w_{Ci} y_i,$$

which is design consistent for the t_y. If $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ for a vector of constants $\boldsymbol{\beta}$, then $\widehat{t}_C = t_y$ for any sample s and $\text{MSE}(\widehat{t}_C|Y) = 0$. For the simple case where $x_i = 1$ for all i, the calibration estimator $\widehat{t}_C$ in (1.7) reduces to the Hájek estimator given by (1.3). The commonly-used Generalized Regression (GREG) estimator is also a special case of (1.7), where

$$(1.8)\qquad \begin{aligned} g_i &= 1 + q_i \Bigl(\mathbf{t_x} - \sum_{j \in s} d_j \mathbf{x}_j\Bigr)^\top \\ &\quad \cdot \Bigl(\sum_{j \in s} d_j q_j \mathbf{x}_j \mathbf{x}_j^\top\Bigr)^{-1} \mathbf{x}_i, \end{aligned}$$

q_i^{-1} being a known positive weight unrelated to d_i. See, for example, Särndal, Swensson and Wretman (1992).

Estimator (1.7) is an example of a model-assisted estimator (e.g., Särndal, Swensson and Wretman, 1992), and is constructed to be design-consistent but is also approximately model-unbiased under an assumed model. The estimator is design-consistent whether or not the model is correctly specified, and as such foreshadows "doubly-robust" estimators in the mainstream statistics literature. Kang and Schafer (2007) discuss connections between survey weights and inverse propensity weights in that literature.

1.2 The Model-Based Perspective

An alternative to randomization inference is the model-based approach which, traditionally, bases inferences directly on an assumed model for the population outcomes. A model is assumed for the population outcomes y with underlying parameters θ, and this model is used to predict the nonsampled values in the population, and hence to predict finite population quantities. More generally, inferences are based on the joint distribution of Y and I (Rubin, 1978, 1983; Peng, Little and Raghunathan, 2004). Rubin (1976) and Sugden and Smith (1984) show that under probability sampling, inferences can be based on the distribution of Y alone, provided the design variables (say Z) are conditioned in the model, and the distribution of I given Y and Z is independent of the distribution of Y given Z. In other words, random sampling is justified since it makes the sampling mechanism ignorable, simplifying model-based inferences. Since the sampling weights can be viewed as design variables, this perspective argues against ignoring the weights for model-based inference.

There are two major variants: superpopulation modelling and Bayesian modelling. In superpopulation modelling (e.g., Royall, 1970; Godambe and Thompson, 1986; Thompson, 1997; Valliant, Dorfman and Royall, 2000), the population values 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. Bayesian survey inference (Ericson, 1969, 1988; Basu, 1971; Scott, 1977; Binder, 1982; Rubin, 1984, 1987; Ghosh and Meeden, 1997; Little, 2003; Peng, Little and Raghunathan, 2004) requires the specification of a prior distribution $p(Y|Z)$ for the population values. Inferences for finite population quantities are then based on the posterior predictive distribution of the nonsampled values (say Y_{exc}) of Y, given the sampled values. The prior distribution is often specified via a parametric model $p(Y|Z,\theta)$ indexed by parameters θ, combined with a prior distribution $p(\theta|Z)$ for θ, that is,

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

where the sampling mechanism is assumed to be ignorable.

The posterior predictive distribution of the nonsampled values Y_{exc} is then

$$\begin{aligned}&p(Y_{\text{exc}}|Y_{\text{inc}},Z)\\ &\quad=\int p(Y_{\text{exc}}|Y_{\text{inc}},Z,\theta)p(\theta|Y_{\text{inc}},Z)\,d\theta,\end{aligned}\tag{1.9}$$

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 $p(Y_{\text{inc}}|Z)$ is a normalizing constant. This posterior distribution induces a posterior distribution for finite population quantities Q.

The specification of $p(Y|Z,\theta)$ in the Bayesian formulation is the same as in parametric superpopulation modelling, and in large samples, the likelihood based on this distribution dominates the contribution from the prior distribution of θ. As a result, large-sample inferences from the superpopulation modelling and Bayesian approaches are often similar, with the key distinction then being between design-based and model-based inference. Bayes modelling has advantages over some superpopulation model techniques in small samples, since the integration over the parameters θ in (1.9) propagates uncertainty in the estimation of θ, yielding better inferences than approaches that fix θ at an estimate $\widehat{\theta}$.

Many authors have considered design-based properties of model-based inferences (e.g., Isaki and Fuller, 1982). Advocates of calibrated Bayes inference (Box, 1980; Rubin, 1984; Little, 2006, 2012) argue that inferences should be Bayesian, but under models that yield inferences with good design-based properties; in other words, Bayesian credible intervals when assessed as confidence intervals in repeated sampling should have close to nominal coverage. For surveys, good calibration requires that Bayes models should incorporate sample design features such as weighting, stratification and clustering. For example, clustering is captured by Bayesian hierarchical models, with clusters as random effects. Prior distributions are generally weakly informative, so that the likelihood dominates the posterior distribution.

What is the role of weights in model-based inference about finite population quantities? A hybrid approach is to apply the sampling weight to the contribution to the model from a sampled case (Molina and Skinner, 1992). For simple random samples, the basis of inference for a parametric model $p(Y|Z,\theta)$ is the log-likelihood

$$l(\theta|Y_{\text{inc}},Z)=\sum_{i=1}^{N}I_i\log p(y_i|z_i,\theta);$$

in particular, maximum likelihood (ML) estimates of θ maximize this function. With survey weights, an analogous "pseudo-" or "weighted" log-likelihood can be defined as

$$(1.10)\qquad l_w(\theta|Y_{\text{inc}}, Z) = \sum_{i=1}^{N} I_i w_i \log p(y_i|z_i, \theta),$$

and θ estimated by maximizing this function (1.10). From a strict modelling perspective, however, the primary focus for descriptive inference is on prediction of nonsampled or nonresponding values, and the weights enter among the design variables Z as covariates in the prediction model (e.g., Gelman, 2007).

EXAMPLE 1.1 CONTINUED. For a stratified random sample, the design variables Z consist of the stratum indicators, and conditioning on Z suggests that models need to have distinct stratum parameters. In particular, consider the normal model $y_{hi}|\mu_h, \sigma_h^2 \overset{\text{ind}}{\sim} N(\mu_h, \sigma_h^2)$, with prior $p(\mu_h, \log\sigma_h^2) = \text{const}$. This model yields the stratified mean as the posterior mean which, as described above, weights sampled cases in stratum h by their design weight N_h/n_h. The posterior variance for known σ_h^2 is the stratified variance. When $\{\sigma_h^2\}$ are unknown, assigning them a flat prior leads to a posterior distribution that is a mixture of t distributions. Many variants of this basic normal model are possible. □

EXAMPLE 1.2 CONTINUED. Estimating a population total from a PPS sample. The posterior mean from the HT model (1.5) differs from the HT estimator by a quantity that tends to zero with the sampling fraction n/N. Zheng and Little (2003, 2005) relax the linearity assumption of the mean structure, modelling the mean of y given size x as a penalized spline, and the variance of y given x as proportional to a power of the x-variable. Simulations suggest that this model yields estimates of the total that have smaller mean square error than the HT estimator when the HT model is a misspecification of the mean function. Further, confidence intervals from the expanded model using jackknife standard errors have better confidence coverage.

Simple weighted estimates like $\bar{y}_{\text{HT}}$ or $\bar{y}_{\text{HA}}$ work well when their underlying models are reasonable, but when they are not, the estimators can have unacceptably high variance—a comical extreme case is Basu's famous (Basu, 1971) elephant example—and associated confidence intervals with poor confidence coverage. Alternatives to these weighted estimators that address these weaknesses are reviewed in Section 2 while considerations when estimating for domains are in Section 3. A large simulation study is described in Section 4 together with summaries of the results. Section 5 has an analysis of a single data set, emphasizing distributions of the weights corresponding to several methods. Further discussion and conclusions are in Section 6.

2. RATIONALE AND TAXONOMY OF METHODS FOR MODIFYING WEIGHTS

2.1 Overview

Large variability in the survey weights is often associated with unstable estimators, although variable weights do not necessarily lead to an increase in the variance of point estimators. For example, suppose that $y_i \propto \pi_i$ for all i. In this case, the Horvitz–Thompson estimator, $\widehat{t}_{\text{HT}}$, provides a perfect estimate of the population total t_y since $V(\widehat{t}_{\text{HT}}|Y) = 0$. This is true even if the π_i's are highly dispersed. However, if the survey weights are highly dispersed *and* exhibit a low correlation with the study variables, the resulting estimators tend to be unstable.

A number of approaches have been developed for modifying the weights to improve survey estimates by reducing mean square error or improving inferential properties like confidence coverage: (a) weight trimming or truncation based on some measure of influence, discussed in Section 2.2; (b) weight modelling, either marginally or conditionally on survey variables, discussed in Section 2.3; and (c) weight modification arising as predictions from model-based estimates, discussed in Section 2.4.

2.2 Weight Trimming

A number of trimming techniques have been proposed in the literature, all sharing the same goal, that is, to modify the survey weights so that the resulting estimators have a lower mean square error than that of the usual estimators (e.g., the Horvitz–Thompson estimator). Weight trimming consists of modifying the weight of units that are identified as "influential". The concept of influential unit is not always clearly defined in practice. In some cases, a unit is identified as influential if it exhibits a large weight; this is discussed in Section 2.2.1. In other cases, a unit showing a large weighted value (i.e., $w_i y_i$) is labeled influential; see Section 2.2.2. More recently, Moreno-Rebollo, Muñoz-Reyes and Muñoz-Pichardo (1999); Moreno-Rebollo et al. (2002) and Beaumont, Haziza and Ruiz-Gazen (2013) used the concept of conditional bias of a unit as a measure of influence in the context of finite population sampling. This is discussed in Section 2.2.3.

Regardless of the weight trimming method used, one must determine an appropriate cut-off point. The weights of units that are above the cut-off point are trimmed. The choice of the tuning constant is important as a bad choice may lead to trimmed estimators with a mean square error larger than that of the nontrimmed estimators. A method frequently encountered in statistical agencies consists of reducing to one the weight of units identified as influential, whereas the outstanding weight is redistributed among the remaining units. However, this approach tends to induce large biases, which in turn, leads to large mean square errors. Estimators with a smaller mean square error are generally obtained through a compromise between the original weight and a weight of one. Different criteria may be used for determining the cut-off point: (i) It may be set in an *ad hoc* fashion; for example, the weights above $b\bar{w}$ may be trimmed, where b is constant and $\bar{w}$ is the mean of the weights. (ii) Alternatively, the cut-off point may be selected so that the estimated mean square error of the resulting trimmed estimator is minimized; for example, Hulliger (1995); Kokic and Bell (1994) and Rivest and Hurtubise (1995). (iii) Beaumont, Haziza and Ruiz-Gazen (2013) proposed a method that consists of determining the value of the cut-off point that minimizes the maximum absolute estimated conditional bias of the trimmed estimator. Unlike (i), the approaches (ii) and (iii) lead to y-specific weights in the sense that each y-variable requires its own cut-off point. As such, the resulting weights are not multipurpose weights.

2.2.1 *Trimming large weights*. The most common form of weight trimming, the *ad hoc* cutpoint method, redistributes sampling weights by picking a cutpoint w_0, forcing weights above this cutpoint to this value, and then multiplying weights below this value by the constant that allows the sum of the trimmed weights to equal the sum of the untrimmed weights:

$$\widetilde{w}_i = \begin{cases} w_0, & \text{if } w_i \geq w_0, \\ \gamma w_i, & \text{if } w_i < w_0, \end{cases}$$

where $\gamma = (n - \sum \kappa_i w_o)/\sum(1-\kappa_i)w_i$, and κ_i is an indicator variable for whether or not $w_i \geq w_0$. Often the adjustment is ignored and γ is simply set to 1. The trimmed estimator of t_y is given by

$$\widehat{t}_{\mathrm{TR}} = \sum_{i \in s} \widetilde{w}_i y_i.$$

Potter (1988, 1990) provided one of the first formal treatments of weight trimming outside of US Census Bureau internal documentation. His weight distribution method assumes that the reciprocals of the scaled survey weights follow a beta distribution, leading to

$$f(w_i) = \frac{n\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \bigl(1/(nw_i)\bigr)^{\alpha+1} \bigl(1 - 1/(nw_i)\bigr)^{\beta-1}.$$

The parameters of the beta distribution are estimated by method-of-moment estimators $\hat{\alpha} = 2 + [\overline{w}(n\overline{w} - 1)/ns_w^2]$ and $\hat{\beta} = (n\overline{w} - 1)[1 + \overline{w}(n\overline{w} - 1)/ns_w^2]$, where $\overline{w}$ and s_w^2 are the sample mean and variance of the weights. The weights from the upper tail of the distribution, say where $1 - F(w_i) < 0.01$, are trimmed to w_0 such that $1 - F(w_0) = 0.01$. This process is repeated a number of times (Potter suggests 10), using the newly trimmed weights from the previous iteration. In this fashion only the "unlikely" weights will be trimmed. Another approach, termed the "NAEP procedure" because of its use in the National Assessment of Educational Progress (Benrud et al., 1978), trims all weights $w_i > w_0 = \sqrt{(c/n)\sum w_i^2}$ for a fixed c, then iterates the procedure until all weights are below some factor c of the square root of the mean of the squared sum. Potter suggests that the value of c be chosen by considering the distribution of $\sqrt{nw_i/\sum w_i^2}$ before trimming.

Another option for trimming is to set the cutpoint to $w_0 = c\bar{w}$, where c must be determined. When c is very large, the trimmed estimator essentially reduces to the nontrimmed estimator (e.g., the Horvitz–Thompson estimator or the Hájek estimator). As c decreases, we expect the proportion of trimmed values to increase, which may or may not translate to a reduction of the mean square error. When $c = 1$ the trimmed estimator is equivalent to an unweighted estimator. In other words, large values of c tend to preserve all the information contained in the π_i's, whereas small values of c tend to get rid of this information. Whether or not the trimmed estimator would be efficient with respect to the untrimmed estimator depends on the relationship between y_i and π_i as well as the form of the nontrimmed estimator. Suppose that we are using a Horvitz–Thompson estimator and that the relationship between y_i and π_i is linear through the origin and the coefficient of correlation between y_i and π_i is strong. In this case, the Horvitz–Thompson estimator is expected to be very efficient. Here, the information contained in the π_i's about the y-variable is highly relevant and it is not advisable to get rid of the information contained in the π_i's through trimming. On the other hand, if there is no relationship between y_i and π_i, the Horvitz–Thompson estimator tends to be very inefficient. In this case, trimming may help in terms of mean square error because it results in a smaller variability of the weights. In fact, if c is set to 1, the resulting estimator is known to have good properties (Rao, 1966; Scott and Smith, 1969) when there is no relationship between y_i and π_i.

Without considering the relationship between the outcome of interest and the probability of inclusion, the (at least implicit) bias-variance tradeoff cannot be made correctly, since bias can only be estimated in the context of a particular outcome. Potter (1988) described one of the first attempts to do this, which explicitly focused on the bias-variance tradeoff using a "minimum mean square error" approach to weight trimming. The weight trimming cutpoint is based on an unbiased estimator of mean square error (MSE) for a trimmed estimator $\hat{t}_{\mathrm{TR}_c}$:

$$\widehat{\mathrm{MSE}}(\widehat{t}_{\mathrm{TR}_c}) = (\widehat{t}_{\mathrm{TR}_c} - \widehat{t}_{\mathrm{HT}})^2 - \mathrm{var}(\widehat{t}_{\mathrm{HT}}) + 2\,\mathrm{cov}(\widehat{t}_{\mathrm{HT}}, \widehat{t}_{\mathrm{TR}_c}),$$

where $\hat{t}_{\mathrm{TR}_c}$ has cutpoint c. The weight trimming cutpoint is given by choosing w_0 to equal the value of c that minimizes $\widehat{\mathrm{MSE}}(\widehat{t}_{\mathrm{TR}_c})$. One drawback of this approach is that the resulting weights are y-specific. The same holds true for the methods presented in Sections 2.2.2 and 2.2.3.

2.2.2 *Trimming large weighted values.* Another approach is winsorization, which reduces the weight of units with a large weighted y-value. Define

$$\tilde{y}_i = \begin{cases} y_i, & \text{if } w_i y_i \le c, \\ \dfrac{c}{w_i}, & \text{if } w_i y_i > c, \end{cases} \tag{2.1}$$

the y-value attached to unit i after winsorization, where $c > 0$ is a cut-off point to be determined. The standard winsorized estimator of t_y is then $\hat{t}_{\mathrm{WIN}} = \sum_{i\in s} w_i \tilde{y}_i$. The latter can also be written as

$$\widehat{t}_{\mathrm{WIN}} = \sum_{i\in s} \tilde{w}_i y_i, \tag{2.2}$$

where

$$\tilde{w}_i = w_i \frac{\min(y_i, \frac{c}{w_i})}{y_i}.$$

The weight $\tilde{w}_i$ can be viewed as a trimmed weight for unit i. If $\min(y_i, \frac{c}{w_i}) = y_i$, we have $\tilde{w}_i = w_i$. In other words, the weight of a noninfluential unit is not modified. For influential units, that is, those for which $w_i y_i > c$, the trimmed weight $\tilde{w}_i$ is smaller than the original weight w_i.

One drawback of the standard winsorization procedure is that some units may receive a weight smaller than 1. To overcome this problem, Dalén (1986) and Tambay (1988) considered a modified winsorization procedure under which the $\tilde{y}$-values are defined as

$$\tilde{y}_i = \begin{cases} y_i, & \text{if } w_i y_i \le c, \\ \dfrac{c}{w_i} + \dfrac{1}{w_i}\Big(y_i - \dfrac{c}{w_i}\Big), & \text{if } w_i y_i > c. \end{cases} \tag{2.3}$$

The resulting estimator can be written as a weighted sum of the original y-values with trimmed weights

$$\tilde{w}_i = 1 + (w_i - 1)\frac{\min(y_i, \frac{c}{w_i})}{y_i}, \tag{2.4}$$

which cannot be smaller than one.

In the context of stratified simple random sampling, Kokic and Bell (1994) and Rivest and Hurtubise (1995) suggested determining the cut-off point c that minimizes the mean square error of the winsorized estimator. Favre-Martinoz, Haziza and Beaumont (2015) suggested determining the cut-off point c that minimizes the absolute estimated conditional bias with respect to the winsorized estimator. This method is discussed in more detail in the next section.

2.2.3 *Trimming weights of cases with large conditional biases.* Beaumont, Haziza and Ruiz-Gazen (2013) proposed reducing the influence of units that exhibit a large influence, assessed through their conditional biases.

The design-based conditional bias associated with the sample unit i with respect to an estimator $\widehat{\theta}$ of a finite population parameter θ is defined as

$$B_i(I_i = 1) = E(\widehat{\theta}|Y, I_i = 1) - \theta; \tag{2.5}$$

see Moreno-Rebollo, Muñoz-Reyes and Muñoz-Pichardo (1999); Moreno-Rebollo et al. (2002) and Beaumont, Haziza and Ruiz-Gazen (2013). If $\theta = t_y$ and $\widehat{\theta} = \widehat{t}_{\mathrm{HT}}$, the conditional bias associated with the sample unit i with respect to $\widehat{t}_{\mathrm{HT}}$ is given by

$$\begin{aligned} B_i^{\mathrm{HT}}(I_i = 1) &= E(\widehat{t}_{\mathrm{HT}}|Y, I_i = 1) - t_y \\ &= \sum_{j\in U}\Big(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\Big)y_j, \end{aligned} \tag{2.6}$$

where π_{ij} denotes the second-order inclusion probability of units i and j in the sample.

EXAMPLE 2.1. For a stratified simple random sampling design, the conditional bias associated with the sample unit i in stratum h reduces to

$$B_i^{\mathrm{HT}}(I_i = 1) = \frac{N_h}{N_h - 1}\left(\frac{N_h}{n_h} - 1\right)(y_i - \overline{Y}_h),$$

where n_h denotes the sample size in stratum h and $\bar{Y}_h = N_h^{-1}\sum_{i\in U_h} y_i$, with U_h denoting the population of units in stratum h of size N_h, $h = 1, \dots, H$.

EXAMPLE 2.2. For a Poisson sampling design, Tillé (2017), the conditional bias associated with sample unit i is

$$B_i^{\mathrm{HT}}(I_i = 1) = (d_i - 1)y_i. \tag{2.7}$$

The conditional bias of a unit can be obtained for all the sampling designs provided that the second-order inclusion probabilities (or, at least, some approximation of these probabilities) are available. It is not difficult to derive expressions of the conditional bias for two-stage and two-phase designs; see Favre-Martinoz, Haziza and Beaumont (2016).

REMARK 1. The conditional bias (2.6) is generally unknown (except for Poisson sampling) as it depends on the y-values for the nonsampled units. Therefore, it must be estimated. An estimator of $B_i^{\mathrm{HT}}(I_i = 1)$ is given by

$$\widehat{B}_i^{\mathrm{HT}}(I_i = 1) = \sum_{j\in s}\left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_j\pi_{ij}}\right)y_j. \tag{2.8}$$

Provided that $\pi_{ij} > 0$ for all $j \in U$, the estimator (2.8) is conditionally unbiased for $B_i^{\mathrm{HT}}(I_i = 1)$. That is, $E\{\widehat{B}_i^{\mathrm{HT}}(I_i = 1)|Y, I_i = 1\} = B_i^{\mathrm{HT}}(I_i = 1)$.

REMARK 2. The conditional bias (2.6) accounts for the sampling design as it depends on the first-order and second-order inclusion probabilities. If $\pi_i = 1$, we have $B_i^{\mathrm{HT}}(I_i = 1) = 0$. That is, a unit included with probability 1 has no influence on an estimator.

REMARK 3. Although we have focussed on the conditional bias associated with a sample unit, note that a nonsampled unit may have a large influence on the quality of an estimator. For the Horvitz–Thompson estimator, we define the conditional bias of a nonsampled unit as

$$\begin{aligned} B_i^{\mathrm{HT}}(I_i = 0) &= E(\widehat{t}_{\mathrm{HT}}|Y, I_i = 0) - t_y \\ &= -(d_i - 1)^{-1}B_i^{\mathrm{HT}}(I_i = 1). \end{aligned}$$

However, it is not possible to estimate $B_i^{\mathrm{HT}}(I_i = 0)$ from the sample values as the y-values associated with nonsampled units are not observed.

Is the conditional bias of a unit an appropriate measure of influence in the design-based framework? Beaumont, Haziza and Ruiz-Gazen (2013) showed that, for Poisson sampling, the conditional bias of a (sampled or nonsampled) unit can be viewed as the contribution of this unit to the sampling error. This result holds approximately for stratified simple random sampling and high entropy sampling designs, provided that the population size N is large enough. The reader is referred to Tillé (2017) for a discussion of high entropy sampling designs.

Beaumont, Haziza and Ruiz-Gazen (2013) have also established the link between the conditional bias of a unit and the design-variance of $\widehat{t}_{\mathrm{HT}}$. That is, using (2.6),

$$\begin{aligned} V(\widehat{t}_{\mathrm{HT}}|Y) &= \sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)y_iy_j \\ &= \sum_{i\in U} B_i^{\mathrm{HT}}(I_i = 1)y_i. \end{aligned}$$

From the previous expression, it is clear that a sample unit whose conditional bias is equal to 0 does not contribute to the variance of $\widehat{t}_{\mathrm{HT}}$.

Based on the estimated conditional bias, Beaumont, Haziza and Ruiz-Gazen (2013) constructed a robust version of the Horvitz–Thompson estimator:

$$\begin{aligned} \widehat{t}_{\mathrm{RHT}}(c) = \widehat{t}_{\mathrm{HT}} &- \sum_{i\in s}\widehat{B}_i^{\mathrm{HT}}(I_i = 1) \\ &+ \sum_{i\in s}\psi_c(\widehat{B}_i^{\mathrm{HT}}(I_i = 1)), \end{aligned} \tag{2.9}$$

where $\psi_c(\cdot)$ is the Huber function given by

$$\psi_c(x) = \begin{cases} c, & \text{if } x > c, \\ x, & \text{if } |x| \le c, \\ -c, & \text{if } x < -c. \end{cases}$$

Note that $\psi_c(x)/x$ lies between 0 and 1 and reduces the influence of highly influential units.

The value of c may be determined by minimizing the estimated mean square error of (2.9). In general, this is a difficult task that often requires simplifying assumptions. Beaumont, Haziza and Ruiz-Gazen (2013) suggested an alternative criterion, which consists of finding the value of c which minimizes the absolute maximum estimated conditional bias with respect to (2.9). That is, we determine c which minimizes $\max_{i\in s}\{|\hat{B}_i^{\mathrm{RHT}}(I_i = 1)|\}$, where $\hat{B}_i^{\mathrm{RHT}}(I_i = 1) = E(\hat{B}_i^{\mathrm{RHT}}(I_i = 1)|Y, I_i = 1) - t_y$ is the conditional bias attached to unit i with respect to (2.9). It can be shown that the resulting estimator is

$$\widehat{t}_{\mathrm{RHT}}(c_{\mathrm{opt}}) = \widehat{t}_{\mathrm{HT}} - \frac{1}{2}(\widehat{B}_{\min}^{\mathrm{HT}} + \widehat{B}_{\max}^{\mathrm{HT}}), \tag{2.10}$$

where $\widehat{B}_{\min}^{\mathrm{HT}} = \min(\hat{B}_i^{\mathrm{HT}}(I_i = 1) : i \in s)$ and $\widehat{B}_{\max}^{\mathrm{HT}} = \max(\hat{B}_i^{\mathrm{HT}}(I_i = 1) : i \in s)$. The estimator (2.10) is easy to implement and is design-consistent for t_y; see Beaumont, Haziza and Ruiz-Gazen (2013).

2.2.4 *Modifications of the weights in generalized regression (GREG) estimates.* In this section, we briefly discuss the weight trimming methods described in Sections 2.2.1–2.2.3 in the context of GREG type estimation. First, the seminal GREG paper of Deville and Särndal (1992) noted that restrictions on the minimum and maximum weights could be imposed as Ld_i and Ud_i, respectively by replacing g_i in (1.8) with L if $g_i < L$ and similarly with U if $g_i > U$. Wu and Lu (2016) extended this approach to minimize the impact on the resulting calibration by adjusting $L < g_i < U$ by a constant so that, for the adjusted and restricted weights w_{Ci}^*, $\sum_{i\in s} w_{Ci}^* = \sum_{i\in s} w_{Ci}$. While this method and its extensions have the advantage of providing nonnegative weights if $L = 0$, it provides no guidance on the choice of L and U for weight trimming, nor does it impact the design weights. Second, the methods of Potter (1988, 1990) can be readily applied by replacing the original weights $d_i = \pi_i^{-1}$ with the GREG weights $w_{Ci} = d_i g_i$, where g_i is given by (1.8). In the context of winsorized procedures, Kokic (1998) discussed one-sided and two-sided procedures for the GREG estimator. Finally, Beaumont, Haziza and Ruiz-Gazen (2013) constructed a robust version of the GREG estimator based on the concept of conditional bias. The GREG estimator being a complex function of estimated totals, the conditional bias associated with a unit is virtually intractable. One must rely on a first-order Taylor expansion, which leads to

$$B_i^{\mathrm{GREG}}(I_i = 1) \simeq \sum_{j\in U}\left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) E_j, \tag{2.11}$$

where the residual $E_j = y_j - \mathbf{x}_j^\top \mathbf{B}$ with

$$\mathbf{B} = \left(\sum_{i=1}^N q_i \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} \sum_{i=1}^N q_i \mathbf{x}_i y_i,$$

and q_i is defined in (1.8). As for the Horvitz–Thompson estimator, we can estimate $B_i^{\mathrm{GREG}}(I_i = 1)$ by

$$\widehat{B}_i^{\mathrm{GREG}}(I_i = 1) \simeq \sum_{j\in s}\left(\frac{\pi_{ij}}{\pi_{ij}\pi_i \pi_j} - 1\right) e_j, \tag{2.12}$$

where $e_j = y_j - \mathbf{x}_j^\top \hat{\mathbf{B}}$ with

$$\hat{\mathbf{B}} = \left(\sum_{i\in s} \pi_i^{-1} q_i \mathbf{x}_i \mathbf{x}_i^\top\right)^{-1} \sum_{i\in s} \pi_i^{-1} q_i \mathbf{x}_i y_i.$$

Following the approach of Beaumont, Haziza and Ruiz-Gazen (2013), we obtain a robust version of $\widehat{t}_{\mathrm{GREG}}$:

$$\widehat{t}_{\mathrm{RGREG}}(c_{\mathrm{opt}}) = \widehat{t}_{\mathrm{GREG}} - \frac{1}{2}(\widehat{B}_{\min}^{\mathrm{GREG}} + \widehat{B}_{\max}^{\mathrm{GREG}}),$$

where $\widehat{B}_{\min}^{\mathrm{GREG}} = \min(\hat{B}_i^{\mathrm{GREG}}(I_i = 1) : i \in s)$ and $\widehat{B}_{\max}^{\mathrm{GREG}} = \max(\hat{B}_i^{\mathrm{GREG}}(I_i = 1) : i \in s)$.

2.3 Weight Modelling, Marginally or Conditionally on Survey Variables

Weight modelling, as mentioned in Section 1.2, also aims to reduce the mean square error at the cost of introducing some bias. Beaumont (2008) sought estimators that improve on the Horvitz–Thompson (HT) estimator without specification of outcome-specific tuning constants which are problematic in multipurpose surveys. He referred to his method as "generalized design-based inference". Unlike the usual model-based approach in finite population sampling which models the outcome vectors, this method models the design weights, and the inference is conditional on the outcome vectors.

Beaumont began with the smoothed random variable

$$\begin{aligned}\hat{t}_{\mathrm{SHT}} &= E(\hat{t}_{\mathrm{HT}} | I, Y) = E\left(\sum_{i\in U} I_i d_i y_i\right)\\ &= \sum_{i\in s} \tilde{d}_i y_i,\end{aligned} \tag{2.13}$$

where $d_i = \pi_i^{-1}$ is the sampling weight and $\tilde{d}_i = E(d_i | I, Y)$ a smoothed weight for the unit $i \in s$. For a nonsampled unit, that is, for $i \in U - s$, $E(I_i d_i y_i | I, Y) = E(d_i | I, Y) I_i y_i = 0$. The term $\hat{t}_{\mathrm{SHT}}$ is introduced to reduce the variability of the d_i, but cannot be used as such, because it involves the unknown $\tilde{d}_i$. To address this issue, Beaumont modeled the d_i to obtain an estimator $\hat{d}_i$ of $\tilde{d}_i$. The final smoothed estimator of t_y is then $\hat{t}_{\mathrm{SHT}} = \sum_{i\in s} \hat{d}_i y_i$.

Beaumont proposed two simple models, although he recognized the potential for others. Beaumont's first model is a linear regression model given by $d_i = \mathbf{h}_i^\top \boldsymbol{\beta} + v_i^{1/2} \varepsilon_i$, where h_i and v_i are known functions of y_i, and where conditional on (I, Y), the ε_i are i.i.d. with zero mean and variance σ^2. For this model, $\tilde{d}_i = \mathbf{h}_i^\top \boldsymbol{\beta}$ and is estimated by $\hat{d}_i = \mathbf{h}_i^\top \hat{\boldsymbol{\beta}}$, where

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i\in s} \mathbf{h}_i \mathbf{h}_i^\top / v_i\right)^{-1} \left(\sum_{i\in s} \mathbf{h}_i d_i / v_i\right). \tag{2.14}$$

Choosing an appropriate model in the context of weight modelling is important as a misspecified model may lead to biases. Standard model selection as well as model diagnostics tools can be applied for selecting a reasonable model. Also, if one removes conditioning on I, $E(I_i d_i y_i | Y) = y_i$ for both $i \in s$ and $i \in U - s$, and then there is no need for weighting.

Two extreme special cases of the above linear model are as follows. First, suppose $h_i = v_i = 1$ for all i so that the model reduces to $d_i = \beta + \varepsilon$ for all i. In this case, $\hat{d}_i = \hat{N}/n$ for all i, where $\hat{N} = \sum_{i \in s} d_i$, so that the variability of the design weights is entirely removed. The resulting smoothed HT estimator is then given by $\hat{t}_{\text{SHT}} = (\hat{N}/n) \sum_{i \in s} y_i$.

The other extreme case is when the h_i are perfect predictors of the design weights d_i without any error so that the error variables ε_i are zero with probability 1. In this case, $d_i = \tilde{d}_i = \hat{d}_i = \mathbf{h}_i^\top \boldsymbol{\beta}$ so that the smoothed HT estimator is identical with the original HT estimator, and smoothing does not lead to any added efficiency. In practice, models that are not so extreme in either direction will be more appropriate.

The above linear model suffers from the drawback that it can produce estimators $\hat{d}_i$ smaller than 1 when the true d_i are known to be bigger than 1. In order to overcome this problem, Beaumont introduced a second model with $d_i = 1 + \exp(\mathbf{h}_i^\top \boldsymbol{\beta} + v_i^{1/2} \varepsilon_i)$, again only for $i \in s$. Then the smoothed weight $\tilde{d}_{i*} = E(d_i | I, Y) = 1 + \exp(\mathbf{h}_i^\top \boldsymbol{\beta}) E\{\exp(v_i^{1/2} \varepsilon_i)\}$, $i \in s$. An analytical expression for an estimator $\hat{d}_i$ of d_i is difficult without further assumptions, but Beaumont approximated $\tilde{d}_{i*}$ by $\tilde{d}_{i*}^a(\boldsymbol{\beta})$, where $E\{\exp(v_i^{1/2} \varepsilon_i)\}$ is replaced by the corresponding sample average $n^{-1} \sum_{l \in s} \exp\{v_l^{1/2} \varepsilon_l(\boldsymbol{\beta})\}$. The random vector $\varepsilon_l(\boldsymbol{\beta}) = \{\log(d_l - 1) - \mathbf{h}_l^\top \boldsymbol{\beta}\}/v_l^{1/2}$ is a function of $\boldsymbol{\beta}$ as justified from the second model. It is easy to check that $E\{\tilde{d}_{i*}^a(\boldsymbol{\beta}) | I, Y\} = \tilde{d}_i$, and the final smoothed estimator is given by $\tilde{d}_{i*}^a(\hat{\boldsymbol{\beta}})$ with $\hat{\boldsymbol{\beta}} = (\sum_{i \in s} \mathbf{h}_i \mathbf{h}_i^\top / v_i)^{-1} \times \{\sum_{i \in s} (\mathbf{h}_i / v_i) \log(d_i - 1)\}$. Further, when $h_i = v_i = 1$ for all i, $\hat{\boldsymbol{\beta}}$ simplifies to $\hat{N}_*/n$, where $\hat{N}_* = \sum_{i \in s} \log(d_i - 1)$. Note that, while the discussion above uses design weights as the model outcome, one could just as easily use GREG weights or other calibration weights as a model outcome.

Kim and Skinner (2013), like Beaumont (2008), also considered weight modification in the original HT estimator. They proposed two methods, one involving multiplication of the inverse probability weights by functions of covariates. The second, quite in the spirit of Beaumont, was smoothing weights involving outcome variables and the covariates.

Any sampling mechanism where the selection probabilities of units depend on outcome variables after conditioning on covariates is referred to as *informative sampling*. This area, developed recently in a series of articles by Pfeffermann and his colleagues, has found application in many complex surveys, for example, in case-control sampling. A source for this is the review article of Pfeffermann and Sverchkov (2009), which contains many useful references. The weight modification approaches of both Beaumont (2008) and Kim and Skinner (2013) are particularly relevant in this context.

2.4 Incorporating Weights Via Models for Survey Variables

A model-based approach in finite population sampling relates survey outcomes to auxiliary variables through a model. This can lead to potentially unreliable estimates of finite population quantities in the event of model failure unless the survey weights are also incorporated as part of the model. One way to address this is to build models that treat the survey weights as covariates, predicting the outcome of interest as a function of the weight, and estimating population-level quantities of interest using Bayesian finite population inference (Little, 1983, 1991; Rubin, 1983). This approach can be extended to the hierarchical model setting, allowing "data-driven" weight trimming that maintains associations between outcomes and weights when the data suggest such associations exist, and smooths them toward 0 otherwise.

Elliott and Little (2000) developed two approaches to induce weight smoothing under stratified designs, as described in Example 1.1. The first of these, termed "weight pooling", used a variable selection method that mimicked weight trimming, except that the trimming cutpoint was treated as an unknown parameter. Elliott (2008) extended this to allow for any number of trimming cutpoints associated with the data to be used:

$$
\begin{aligned}
y_{hi} | \boldsymbol{\mu}_l, \sigma^2, L = l &\overset{\text{ind}}{\sim} N(\mathbf{Z}_{lhi}^\top \boldsymbol{\mu}_l, \sigma^2), \\
\boldsymbol{\mu}_l | \sigma^2, L = l &\sim N(\boldsymbol{\mu}_{0l}, \sigma^2 \boldsymbol{\Sigma}_{0l}), \\
\sigma^2 | L = l &\sim \text{Inv}-\chi^2(a, s^2), \\
p(L = l) &= 2^{-(H-1)},
\end{aligned}
\tag{2.15}
$$

where L indexes the $2^{(H-1)}$ possible patterns of pooling coterminous strata, $\mathbf{Z}_{lhi}$ is a vector of dummy variables of length H^* set equal to 1 for the pooled stratum to which the hth stratum belongs, and 0 otherwise, and $\boldsymbol{\mu}_l = (\mu_1, \dots, \mu_{H^*})^\top$ corresponds to the vector of means associated with these pooled strata. The model includes as special cases both the unweighted estimator (when $l = 1$, $H^* = 1$ and $Z_{lhi} = 1$ for all h) and fully-weighted estimator, $N^{-1}\sum_{h=1}^{H} N_h\bar{y}_h$, of $\bar{Y}$ (when $l = 2^{(H-1)}$, $H^* = H$ and $\mathbf{Z}_{lhi}$ is a vector consisting of 0s except for a 1 in the hth element). The posterior mean of $\bar{Y}$ is obtained by averaging all possible poolings of coterminous inclusion strata, where each estimator contributes to the final average based on the posterior probability that the pooling is "correct":

$$\begin{aligned} p(\bar{Y}|\mathbf{y}) = \sum_l \int\int & p(\bar{Y}|\boldsymbol{\mu}_l, \sigma^2, L = l, \mathbf{y}) \\ & \cdot p(\boldsymbol{\mu}_l|\sigma^2, L = l, \mathbf{y}) p(\sigma^2|L = l, \mathbf{y}) \\ & \cdot p(L = l|\mathbf{y})\, d\boldsymbol{\mu}_l\, d\sigma^2. \end{aligned} \tag{2.16}$$

This integral cannot be computed analytically, but can be obtained via simulation. Elliott (2008, 2009) showed that weight pooling models were robust and had substantial efficiency gains over standard weighted estimators of population means when associations between probability of selection and means or regression parameters were weak, particular when fractional Bayes factors (O'Hagan, 1995) were employed.

A second approach, first suggested by Holt and Smith (1979), considered random effects or "weight smoothing" models that used a hierarchical structure to induce shrinkage in the weight stratum terms:

$$y_{hi}|\mu_h \overset{\text{ind}}{\sim} N(\mu_h, \sigma^2), \quad \boldsymbol{\mu} \sim N_H(\boldsymbol{\phi}, \mathbf{D}), \tag{2.17}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_H)'$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_H)'$, and a weak or noninformative hyperprior $p(\boldsymbol{\phi}, \mathbf{D})$ is posited. Based on (2.17), the posterior predictive mean of $\bar{Y}$ is

$$E(\bar{Y}|\mathbf{y}) = N^{-1}\sum_{h=1}^{H}(n_h\bar{y}_h + (N_h - n_h)\hat{\mu}_h), \tag{2.18}$$

where $\hat{\mu}_h = E(\bar{Y}_h|\mathbf{y}) = E(\mu_h|\mathbf{y})$.

A simple weight smoothing model is the exchangeable random effects model, where $\phi_h = \phi_*$ for all h and $\mathbf{D} = \tau^2\mathbf{I}_H$ (Holt and Smith, 1979; Ghosh and Meeden, 1986; Little, 1991; Lazzeroni and Little, 1998). Under this model, with a uniform prior for ϕ_*, one gets $E\{\bar{Y}|\mathbf{y}\} = N^{-1}\sum_{h=1}^{H} N_h[w_h\bar{y}_h + (1 - w_h)\bar{y}]$, where $w_h = \tau^2 n_h/(\sigma^2 + \tau^2 n_h) = n_h/(n_h + \sigma^2\tau^{-2})$ and $\bar{y} = \sum_{h=1}^{H} n_h\bar{y}_h/\sum_{h=1}^{H} n_h$. In the case when $\tau^2 \to 0$, $w_h \to 0$ for all h, leading thereby to the estimator $N^{-1}\sum_{h=1}^{H}[n_h\bar{y}_h + (N_h - n_h)\bar{y}] = \bar{y}$ of $\bar{Y}$. Thus, all the unsampled units are estimated by the pooled mean, which is sensible since the model assumes now that the observations in all strata have a common mean. On the other hand, when $\tau^2 \to \infty$, $w_h \to 1$ for all h, one gets the fully weighted estimator.

Elliott and Little (2000) considered the following generalizations of the exchangeable model.

(I) Linear: $\phi_h = \alpha + \beta_h$ for all h, $\mathbf{D} = \tau^2\mathbf{I}_H$ (Lazzeroni and Little, 1998).

(II) Autoregressive: $\phi_h = m$ for all h, $\mathbf{D} = \tau^2(\rho^{|i-j|})$ (Lazzeroni and Little, 1998).

(III) Nonparametric: $\phi_h = g(h)$, $\mathbf{D} = 0$, where g is a twice differentiable smooth function of h satisfying (i) g and g' absolutely continuous and (ii) $\int(g''(u))^2\,du < \infty$ (Wahba, 1978; Hastie and Tibshirani, 1990).

The function g minimizes the residual sum of squares plus a roughness penalty given by

$$\sum_{h=1}^{H}\sum_{i\in s_h}(y_{hi} - g(h))^2 + \lambda\int(g''(u))^2\,du,$$

λ denoting the penalty parameter.

Elliott and Little found that models with little structure in $\boldsymbol{\phi}$ and $\mathbf{D}$ had larger gains in efficiency when the association between the probability of inclusion and the mean was weak, but were vulnerable to "oversmoothing" when this association was strong. The nonparametric mean prior (III) yielded a highly robust estimator of the population mean when weights were needed to adjust for bias, with a moderate increase in efficiency when bias correction was unnecessary. Elliott (2007) developed extensions of weight smoothing models for linear and generalized linear model regression.

A related approach uses penalized spline (p-spline) models to predict the outcome using the probabilities of inclusion for the nonsampled units. This approach does not require a stratified sample design; instead, sampling weights for nonsampled units are typically available in probability-proportional-to-size sample designs, where the probability of inclusion π_i for the ith element in the population is proportional to a measure of size variable x_i known for all elements in the population: $\pi_i = \frac{nx_i}{\sum_{i=1}^{N} x_i}$. Zheng and Little (2003, 2005) considered p-spline models of the form:

$$\begin{aligned} y_i = \beta_0 + & \sum_{j=1}^{p}\beta_j\pi_i^j \\ & + \sum_{l=1}^{m}\beta_{l+p}(\pi_i - \kappa_l)_+^p + \varepsilon_i, \\ \varepsilon_i & \overset{\text{ind}}{\sim} N(0, \pi_i^{2k}\sigma^2), \\ \beta_{l+p} & \sim N(0, \tau^2), \quad l = 1, \ldots, m, \end{aligned} \tag{2.19}$$

where $(x)_+ = x$ if $x > 0$ and 0 otherwise, and k is known (to allow for heteroscedasticity). This can be viewed as an extension of the model implied by the Horvitz–Thompson estimator (1.5). In settings where the implied HT model is correct, Zheng and Little showed that the p-spline estimator was almost as efficient as the HT estimator, whereas in other settings it had much lower mean square error. Zheng and Little (2005) also showed that the p-spline estimator using jackknife standard errors yielded inferences that are superior to the HT and the GREG estimators. Results were not sensitive to the choice and number of knots as long as there were a sufficient number (Zheng and Little suggested 15). Chen, Elliott and Little (2012) extended model (2.19) to estimate finite population quantiles by estimating the variance of y_i using a second p-spline function of π_i.

Chen, Elliott and Little (2010) also extended this into a setting with binary outcomes to estimate population proportions using a probit p-spline model:

$$\begin{aligned}&\Psi\big(P(y_i = 1)\big)^{-1}\\ &\quad = \beta_0 + \sum_{j=1}^{p} \beta_j \pi_i^j + \sum_{l=1}^{m} \beta_{l+p}(\pi_i - \kappa_l)_+^p,\end{aligned} \tag{2.20}$$

where $\Psi(\cdot)^{-1}$ denotes the inverse CDF of a standard normal distribution. The population proportions were estimated using $N^{-1}(\sum_{i \in s} y_i + \sum_{j \notin s} \hat{y}_j)$. They showed by simulation studies and real data applications that the p-spline estimators yielded substantial gains over the classical weighted estimators for population proportions with respect to mean square error, confidence coverage and interval length, especially when the sample size is small.

3. ESTIMATION FOR DOMAINS

Estimates are often required not only at the population level but also for subpopulations called domains. If the population is divided into mutually disjoint domains $U_1, \dots, U_G$ such that $\bigcup_{g=1}^{G} U_g = U$, it seems desirable that the estimates of domain totals of y sum to the estimate of the total of y in the population. This property is often referred to as external consistency. This condition might not be satisfied if weight modification is performed at the population level and within each domain independently. One possible solution for achieving external consistency is to obtain the estimate at the population level by aggregating the domain estimates. However, the compounding of biases in the domain estimates may lead to appreciable bias in the aggregate (Rivest and Hidiroglou, 2004). To overcome this problem, a simple solution was proposed by Favre-Martinoz, Haziza and Beaumont (2015). It consists of performing weight modification within each domain separately and obtaining G estimates, $\widehat{t}_1^*, \dots, \widehat{t}_G^*$. Independently, an estimator at the overall level, $\widehat{t}_0^*$, is also obtained using the same weight modification method. The idea is then to force consistency by determining final estimates, $\widetilde{t}_0, \widetilde{t}_1, \dots, \widetilde{t}_G$ as close as possible to the initial estimates $\widehat{t}_0^*, \widehat{t}_1^*, \dots, \widehat{t}_G^*$, subject to external consistency being satisfied.

4. SIMULATION STUDY

4.1 Introduction to Simulations

The overarching goal of the simulation study is to relate the performance of the various estimators to:

(i) the nature and distribution of the weights;

(ii) the shape and nature of the relationship between the response variable and the weights;

(iii) the degree of agreement with models [such as the Horvitz–Thompson (1.5) and Hájek (1.6) models] that lead to particular estimators; and

(iv) the proportions, sizes and degree of symmetry of outliers.

This study extends the one carried out by Henry and Valliant (2012). A selected set of estimators of the population mean of a continuous response variable y was considered with populations generated from a variety of models (see Table 1). A similar simulation for the population proportion of a binary variable y is shown in the online supplementary materials (Chen et al., 2017).

The population size was 20,000 and the sample size was 200 for each scenario considered. Where there was stratification, the stratum sizes were 5000, 6000 and 9000, and the sample sizes were allocated proportionally to the stratum sizes.

4.1.1 *Calculation and distribution of weights.* The sampling design for most of the scenarios was single stage systematic PPS sampling, using ppss() from the PPS package in R, and using as size variable the values of a positive random variable x. The basic design weights were the reciprocals of the inclusion probabilities. When making inclusion probabilities proportional to x resulted in some of them exceeding 1, the inclusion probabilities for the corresponding units were set equal to 1 and the rest recalculated—iteratively if necessary—so that inclusion probabilities

TABLE 1
Scenarios: PPS sampling with x as the size variable, ε is distributed as $N(0,1)$, independently of x

x^a	Scenario	$y \mid x =$	Note
A	1	$10 + 0.5 \log x + \varepsilon$	stratified, same slopes
	2	$10 + 0.5 \log x + 0.5Z_1^b - Z_2^b - 0.2Z_1 \log x + 0.3Z_2 \log x + \varepsilon$	stratified, varying slopes
	3	$(10 + 0.5 \log x + 0.5Z_1 - Z_2 - 0.2Z_1 \log x + 0.3Z_2 \log x + \varepsilon)U_0^d + (10 + 0.5 \log x + e)(1 - U_0); \log e \sim N(0, 1)$	stratified, mixed slopes
B[c]	4	ε	no association
	5	$100\pi + 10\pi\varepsilon$	HT model
	6	$\log(\pi + 0.0001n) + \pi^{0.25}\varepsilon$	log function of π
	7	$\{\pi > q_{20}^{\pi}\} + \{\pi > q_{40}^{\pi}\} - \{\pi > q_{60}^{\pi}\} - \{\pi > q_{80}^{\pi}\} + \varepsilon$	stepwise function of π
C	32	$3.5x + 0.8x\varepsilon$	HT model
	33	$50 + 3\sqrt{x}\varepsilon$	Hájek model
	35	$(3x + 0.8x\varepsilon)U_1^d + (7000 + 5000\varepsilon)(1 - U_1)$	HT model w/outliers
	36	$500 + 3x + 250\varepsilon$	linear w/intercept
	37	$(500 + 3x + 250\varepsilon)U_1 + (7500 + 3000\varepsilon)(1 - U_1)$	linear w/int, high outliers
	38	$(10{,}000 + 3x + 500\varepsilon)U_2^d + (10{,}750 + 3x + 10{,}000 * \varepsilon)(1 - U_2)$	linear w/int, high/low outliers
	40	$2500 + \mathrm{e}^{0.01x} + x\varepsilon$	nonlinear heteroscedastic
	41	$(2500 + e^{0.0105x} + 500\varepsilon)U_1 + (10{,}000 + 5000\varepsilon)(1 - U_1)$	nonlinear w/outliers
	47b	$10{,}000 + 3x + 10{,}000\varepsilon$	$y \mid x$ of 47 and x of C
D	42	$2500 + x^2 + 1000\varepsilon$	nonlinear homoscedastic
	43	$2500 + x^2 + 20x\varepsilon$	nonlinear heteroscedastic
E	44	$(50 + 3\sqrt{x}\varepsilon)U_3^d + (100 + 30\varepsilon)(1 - U_3)$	Hájek model w/outliers
F	36b	$500 + 3x + 250\varepsilon$	$y \mid x$ of 36 and x of F
	47	$10{,}000 + 3x + 10{,}000\varepsilon$	linear w/int, $\log Nx$
	48	$3x + 100\sqrt{x}\varepsilon$	linear, $\log Nx$
	49	$500 + \sqrt{x}\varepsilon$	Hájek, $\log Nx$

[a] $A = 100 \log N(0.5, 0.25)$, $B = \Gamma(1.5, 0.001)$, $C = 50\Gamma(5, 1)$, $D = 10\Gamma(5, 1)$, $E = 5\Gamma(1, 1)$, and $F = 100 \log N(0, 4)$. [b] $Z_1 = 1$ if stratum 1 and 0 otherwise, and $Z_2 = 1$ if stratum 2, and 0 otherwise. [c] $\pi = nx / \sum_{i \in U} x$. [d] $U_0 \sim$ Bernoulli(0.8), $U_1 \sim$ Bernoulli(0.995), $U_2 \sim$ Bernoulli(0.99), $U_3 \sim$ Bernoulli(0.95).

overall would sum to the sample size. The basic design weights were then the reciprocals of the recalculated inclusion probabilities.

The x distributions considered ranged from somewhat skew to highly skew. Specifically, we used: 100 × lognormal(0.5, 0.25); 100 × lognormal(0, 4) (leading to some calculated inclusion probabilities exceeding 1); $\Gamma(1.5, 0.001)$ (corresponding to highly variable inclusion probabilities); 50 × $\Gamma(5, 1)$; 10 × $\Gamma(5, 1)$; and 5 × $\Gamma(1, 1)$. The density curves of the x distributions are displayed in Figure 1A (online supplementary material).

In each scenario, the population values, including the x variable values, were generated newly for all population units before each realization of the sample from the sampling design. Where there was stratification, the strata were formed independently of x.

4.1.2 *Relationship between the response variable y and the size variable x.* The relationship between y and the size variable x was determined through a specification of a model, in some cases depending on stratum. There were no other explanatory variables. See Table 1 for the list of models.

4.1.3 *Proportions, sizes and symmetry/asymmetry of outliers.* The proportions of outliers introduced in the response variable ranged from 0 through 0.005 to 0.05. Outliers were both symmetric and asymmetric. With a sample size of 200, the smaller proportion provided on average one outlier per sample, and the larger on average 10 outliers per sample.

4.1.4 *Estimators.* The following estimators of the finite population mean of y were computed, for comparisons of relative bias and relative root mean square error (RMSE):

(a) The Horvitz–Thompson (HT) estimator, given by equation (1.2).

(b) The Hájek (HA) estimator, given by equation (1.3).

(c) The robust HT (HT-ROB) estimator of Section 2.2.3, given by equation (2.10).

(d) The robust Hájek (HA-ROB) estimator, analogous to (c).

(e) A trimmed weight Hájek (TRIM-3W) estimator with weights exceeding 3 times the mean weight set equal to that value and the weights rescaled to preserve the total of sample weights, iterated to convergence.

(f) The penalized spline of propensity prediction (PSPP-HOM) estimator, from prediction of y in terms of a penalized linear spline function of x with $m = 10$ equally spaced knots, assuming homoscedastic errors (Zheng and Little, 2003). The model was fitted using a fully Bayesian approach. The model is given by equation (2.19) with $k = 0$, $p = 1$, locally uniform priors on β_0 and β_1, and σ^2 and τ^2 having $\text{InvGamma}(10^{-5}, 10^{-5})$ priors. The posterior MCMC computation uses 10,000 iterations, with 1000 burn-in and subsequently drawing every 10th iteration, for 900 draws altogether.

(g) The PSPP estimator with the same settings, but allowing heteroscedastic errors in the sense that the error variances are proportional to the $2k$th power of the inclusion probability (PSPP-HET); the prior for k is uniform on $(-2, 2)$.

(h) The estimator of Beaumont, given by equation (2.13), where the original weight is replaced by the predicted value from a regression of the weight on a spline function of y (BEAU-PS).

(i) The generalized regression (GREG) estimator, with $\mathbf{x}_i = (1, x_i)^\top$.

(j) The robust GREG estimator (GREG-ROB) of Section 2.2.4, analogous to (c).

Of these, a subset for which variance estimation is straightforward were compared in terms of associated confidence or credible interval (for PSPP) widths and empirical coverage probabilities:

(a) The HT estimator.

(b) The Hájek estimator.

(f) The PSPP estimator assuming homoscedastic errors.

(g) The PSPP estimator allowing heteroscedastic errors.

(i) The GREG estimator.

The empirical bias of the estimator of variance was also obtained in each case.

Variance estimation for (a) was based on the following expression:

$$\hat{V} = \frac{1}{N^2} \sum_{i \in s} \phi_i \left(\frac{y_i}{\pi_i} - \hat{b} \right)^2, \tag{4.1}$$

where

$$\hat{b} = \frac{\sum_{i \in s} \alpha_i \frac{y_i}{\pi_i}}{\sum_{i \in s} \alpha_i}$$

with $\phi_i = \alpha_i = \frac{n}{n-1}(1 - \pi_i)$; see Matei and Tillé (2005) and Haziza, Mecatti and Rao (2008).

For the Hájek estimator, the variance was estimated using (4.1) with $z_i = (y_i - \hat{\mu})$ replacing y_i, where $\hat{\mu}$ is the Hájek estimator of the mean, and $\hat{N} = \sum_{i \in s} \pi_i^{-1}$ replacing N.

Similarly, the variance of the GREG estimator (with $q_i = 1$) was estimated using (4.1) with $z_i = (y_i - \mathbf{x}_i^\top \hat{\mathbf{B}})$ replacing y_i, where

$$\hat{\mathbf{B}} = \left(\sum_{i \in s} \pi_i^{-1} \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \sum_{i \in s} \pi_i^{-1} \mathbf{x}_i y_i,$$

and $\hat{N}$ replacing N.

For the Hájek estimator, replacement of N by $\hat{N}$ is suggested by the study of Wu and Deng (1983). For the GREG estimator, replacement of N by $\hat{N}$ is by analogy.

It should be noted that the formula for $\hat{V}$ assumes that it is the finite population mean which is being estimated, and thus a slight underestimation of the variance of estimation of the model mean, the estimand in these simulations, is expected. However, in additional simulations where the finite population mean was taken to be the estimand, the overall conclusions did not change.

4.1.5 *Intervals*. For the HT, Hájek and GREG estimators, the intervals were the standard normal-based confidence intervals: $\hat{\mu} \pm 1.96 SE(\hat{\mu})$. Credible intervals for the PSPP estimators were the 2.5th and 97.5th percentiles of the posterior draws.

4.1.6 *Simulation design*. As indicated above, for each sample in each scenario, following generation of the population x values and the stratification, the values of the y variable were generated for all population units given the x values and stratification. The true value of the estimand was taken to be the expectation under the model of the finite population mean. There were 200 iterations for each scenario initially, and subsequently 20,000 iterations for the HT, Hájek and GREG methods.

For each estimator and scenario, the empirical Relative Bias was computed as the absolute value of the empirical bias of the estimator (as an estimator of the superpopulation mean) divided by the absolute value of the empirical bias of the Hájek estimator. The empirical Relative RMSE was computed as the square root of the empirical MSE of the estimator divided by either the square root of the empirical MSE of the Hájek estimator or the minimum (over the set of estimators considered) of the set of square root empirical MSEs. Where the confidence or credible intervals were calculable, the average width (CIW) and the empirical coverage probabilities were obtained for each estimator and scenario. The relative RMSEs are displayed in Figure 2A (online supplementary material) in dot plots for all estimators while the relative RMSE, CIW and noncoverage are displayed in dot plots for the HT, Hájek, PSPP and GREG methods in Figure 3A (online supplementary material). We also looked at an additional set of plots, that is, for each scenario, (a) scatterplots of y vs. x and y versus sampling weight for a finite population and a selected PPS sample of values, (b) bar charts of relative (to Hájek) bias and RMSE, (c) bar charts of relative (to Hájek) CIW and percent coverage for the HT, Hájek, PSPP and GREG methods, (d) histograms of $t = (\hat{\mu} - \mu)/SE(\hat{\mu})$ and the Monte Carlo relative bias of the variance estimator of $\hat{\mu}$ for the HT, Hájek, PSPP and GREG estimators, and (e) scatterplots of CIW against $(\hat{\mu} - \mu)$. This set of plots is given in Figures 4A and 5A (online supplementary material) for scenarios 35 and 38, respectively.

4.2 Results

4.2.1 *Bias and RMSE of estimators.* We have evaluated the ten estimators (listed in Section 4.1.4), in terms of their absolute biases and RMSEs; see Figure 2A in the online supplementary material for the RMSEs. A summary comparison of the RMSE of the ten estimators is presented as a heat map in Figure 1.

The heat map is constructed as follows. Letting $\mathrm{RMSE}(i, r)$ denote the root mean square error for estimator i under scenario r, let

$$R(i, r) = \mathrm{RMSE}(i, r) \\ /(\min\{\mathrm{RMSE}(j, r), j = 1, \dots, 10\}).$$

The columns (estimators) are ordered in terms of the mean of $R(i, r)$ over all scenarios r, from smallest (PSPP-HET) to largest (HT). The rows (scenarios) are ordered so that those with similar color patterns are close together. Specifically, they are ordered by the 70th percentile of $R(i, r)$ for fixed r and varying i. There are 13 colors, with lighter colors for smaller

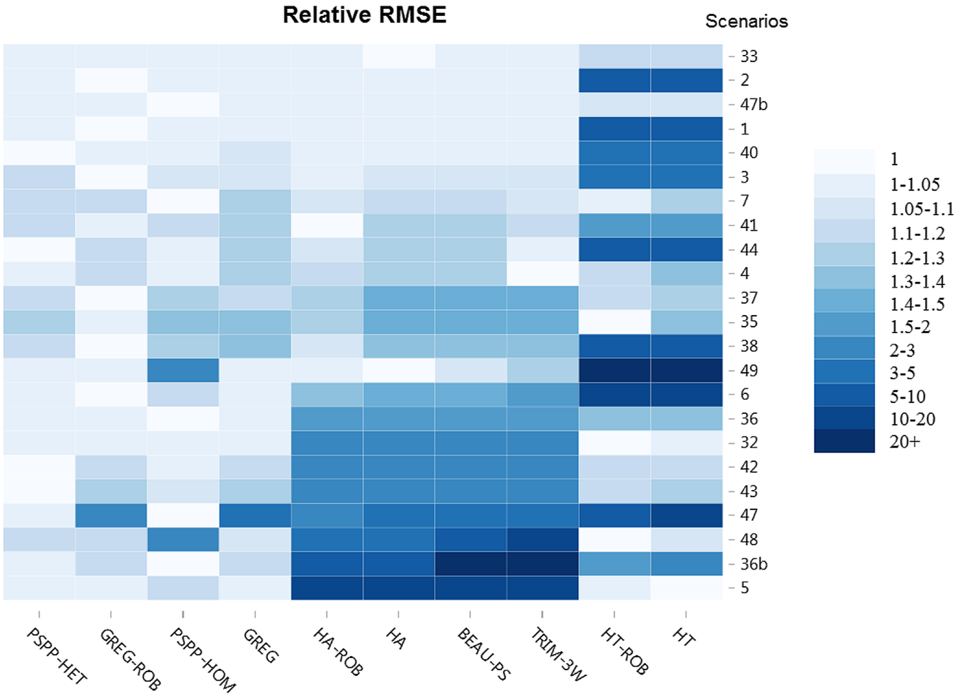

FIG. 1. *Relative RMSE of each estimator compared to the estimator with the smallest RMSE under each model scenario. The columns are ordered by the mean relative RMSE for each method and the rows are ordered by the 70th percentile of relative RMSE for each scenario.*

$R(i,r)$ and darker colors for larger $R(i,r)$. The 13 corresponding intervals for $R(i,r)$ are 1, (1, 1.05], (1.05, 1.1], (1.1, 1.2], (1.2, 1.3], (1.3, 1.4], (1.4, 1.5], (1.5, 2], (2, 3], (3, 5], (5, 10], (10, 20], 20+. If all of the rectangles in a column have a light color, then that estimator has small RMSE over all scenarios. Correspondingly, if a column has many dark rectangles then that estimator has a substantial number of large values of RMSE. For example, PSPP-HET (first column) has small RMSE overall while HA (sixth column) does not.

In terms of RMSE and for the scenarios presented, there are three distinct groups of estimators. In the first group, the PSPP-HET, GREG-ROB, GREG and PSPP-HOM methods perform similarly, with the PSPP-HET and GREG-ROB methods being superior. The allowance for nonconstant variance in PSPP-HET results in gains for PSPP-HET over PSPP-HOM in a few scenarios. The robustness feature of GREG-ROB results in gains over GREG in some problems, including those with outliers. The second group includes HA and related estimators: HA-ROB, BEAU-PS and TRIM-3W. The third group comprises HT and HT-ROB. The second and third groups perform relatively well for complementary sets of scenarios, the second group working well when the data are generated by a model that is not too far from the Hájek model (1.6), and the third group working well when the data are generated by a model that is not too far from the HT model (1.5).

For interpretation of simulation results for the trimmed estimator it is helpful to note that the observations with large weights are those with small x values and the observations with small weights are those with large x values. The more highly variable the weights, the greater the proportion of observations that will have trimmed weights for a given value of the cut-off constant c. Trimming weights with high values will decrease the contribution of points with small x values and increase the contribution of points with large x values. Thus, if y is associated with x, the bias of the trimmed estimator will increase, the larger the proportion of weights that are trimmed. In the heat map, it is apparent that the trimmed estimator does well in terms of MSE in cases of no association between y and x (scenario 4) and the Hájek model with outliers (scenario 44).

In general, the robust HT, Hájek and GREG estimators perform about the same as their nonrobust counterparts. In the scenarios with a difference, the bias tends to be higher but the RMSE lower for the robust versions, as might be expected; most of these scenarios involve outliers.

4.2.2 *Interval coverage and width.* We have also investigated the coverage and interval widths of those procedures where methods for constructing intervals are straightforward. Doing so means that we have not assessed those procedures where bootstrap methods would be needed. Thus, this evaluation is limited to the HT, HA, PSPP-HOM, PSPP-HET and GREG methods.

We start with an overall evaluation of these methods. A heat map, Figure 2, presents for each method and scenario the amount of noncoverage and relative interval width. In Figure 2, an appearance of light (blue scale) colors in a column indicates good coverage while an appearance of light (red scale) colors indicates small relative width. Looking at the scenarios and estimators where coverages are close to the nominal 95%, and widths are relatively small, it is apparent that most of them correspond to PSPP-HET and GREG. Thus, for this set of scenarios the PSPP-HET and GREG are the dominant methods. It is notable that the large widths for the HT intervals negate the generally good coverage for this method, this being true to a somewhat lesser extent for HA. Overall, the balance between coverage and width is best for PSPP-HET.

Another general conclusion is that most of the scenarios where the performance of several of the methods is poor are those where there are outliers.

We now discuss some of our findings. We look first at HT, noting that, for specificity, we define satisfactory coverage as between 0.92 and 0.98. While, as expected, the HT intervals are fully satisfactory when the model generating the data is close to the HT model, this is also true for other scenarios, for example, some where the model has an intercept. Overall, the performance of the HT method is satisfactory in twelve of the twenty-three scenarios. Most of the exceptions are where there are outliers, or there is a substantial departure from the HT model.

Two cases illustrate common themes in these scenarios. The first is scenario 35 where the histogram of the t statistic, $t = (\hat{\mu} - \mu)/SE(\hat{\mu})$, is highly (left) skewed, and the coverage is quite low. Here, the HT estimator is too small relative to the expected value of y when there is no outlier in the sample. Considering only those samples where there are no observations from the non-HT part of the mixture distribution, the histogram of the t statistic is symmetric, while those histograms for samples with at least one outlier are highly left skewed. The second is scenario 38 where the distribution of the t statistic is symmetric, the coverage is satisfactory but the average interval width is large. In this case, the distribution of the variance estimator is shifted to the right

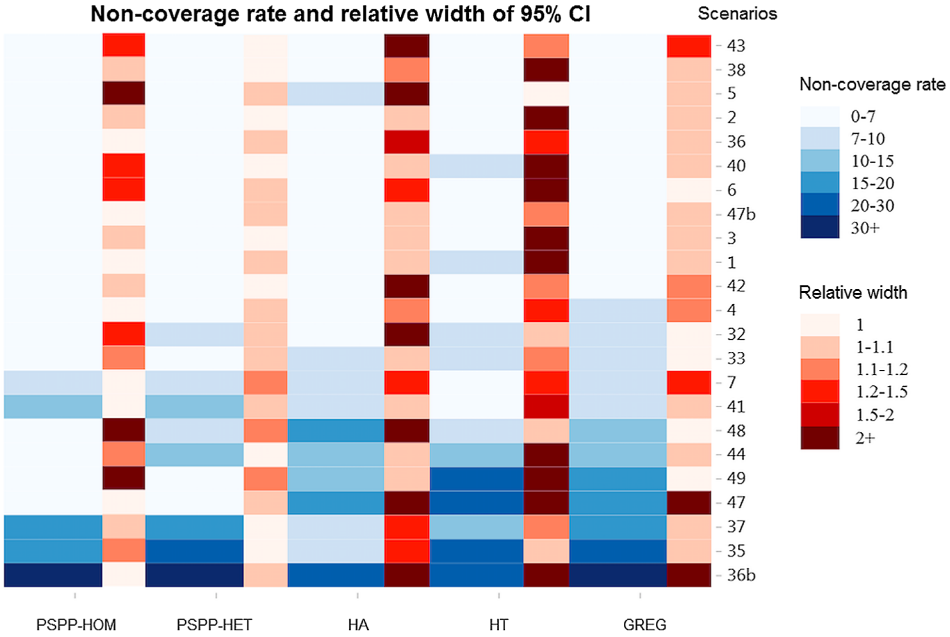

FIG. 2. *Relative average interval width and noncoverage rate of* 95% *CI under each model scenario. The relative width is the ratio of the average interval width of each estimator compared to the estimator with the smallest avearge interval width under each scenario. The columns are ordered by the mean noncoverage rate for each method and the rows are ordered by the* 75th *percentile of the noncoverage rate for each scenario.*

of the other variance estimators, that is, it is often too large. One can see from (4.1) that if some of the outliers in the sample are associated with small values of x, the estimated variance gives greater weight to those contributions to the square error, yielding a large estimated variance and a wide interval. See Figures 4A and 5A in the online supplementary material for the plots relevant to this analysis.

Of the twenty-three scenarios, there are ten where the HA method has either low coverage or large interval width. In most of these scenarios, either there are outliers, or the model generating the data is far from the HA model.

For the GREG method, there are seven scenarios where the performance is unsatisfactory. Three of these have outliers, and the explanation is the same as that given for HT for scenario 35. It is surprising that the GREG method has poor coverage in scenarios 47–49 where, in each case, there is a linear relationship between y and x. Noting that for those scenarios, the number of simulations may have been insufficient because of the very high variability of the weights, we investigated further, using 20,000 simulations. Then the coverage is almost 0.92 for both scenarios 48 and 49. However, for scenario 47 the relative bias of each of the variance estimators is very large, and the coverage is very low. By using $x \sim 50\Gamma(5, 1)$, scenario 47b, rather than $x \sim 100 \log N(0, 4)$ in scenario 47, the bias of the basic variance estimator is now small, and the coverage is satisfactory. We have seen this result in other scenarios, that is, that the distribution of x may make a large difference in the properties of the intervals.

In general, the PSPP method tends to outperform the other methods: there are only five scenarios where the coverage or width for PSPP-HET is poor. Four of these are ones where there are outliers and the explanation given for the low coverage for scenario 35 for HT apparently holds here as well.

5. APPLICATION: ESTIMATION OF STUDENT SOCIOECONOMIC STATUS MEASURES

To compare the methods using a real dataset, a single-stage PPS sample of 200 Texas schools was drawn from the population of schools in the National Center for Education Statistics Core of Common Data (NCES/CCD) database, using the number of students attending the school as a measure of size. The NCES/CCD is a national database of US primary and secondary public schools (http://nces.ed.gov/ccd/aboutCCD.asp). The population was restricted to the 7171 schools with complete data on location, Title I funding status and number of children on free or reduced lunch status. We considered two sets of base weights: the PPS sampling weights themselves, and these weights calibrated (post-stratified) so that the distribution of urbanicity matched the "known" distribution of urbanicity taken from the population data. [Urbanicity was defined by $4 \times 3 = 12$ levels; 4 levels of city, suburb, town and rural crossed with 3 size levels (for city and suburb) or three location levels relative to urbanized area (for town and rural).] We focus on estimating the mean number of free lunch students in Texas public schools.

We obtained estimates using all of the methods described in the simulation study: the Horvitz–Thompson and Hájek estimators, the Hájek estimator trimmed to have a maximum weight value 3 times the mean of the sampling weights, the PSPP estimators using the equal and unequal variance assumptions, the Beaumont estimator using the predicted values of the weights from a spline model to the free lunch count and the GREG estimator using total number of students in Texas as a calibrating variable; robust versions of the Horvitz–Thompson, Hájek and GREG estimators were also obtained (see Table 2). All confidence intervals in Table 2 were computed using with-replacement approximations, which were found to be approximately correct for the Horvitz–Thompson and Hájek sample weighted estimators; confidence intervals were not computed for the robust estimators. Figure 3 shows the distribution of the original Horvitz–Thompson and calibrated weights, as well as for the crude trimming and Beaumont methods that involve direct adjustment of the weights (either original or calibrated). Calibration to location generates a long tail of large weights; the crude trimming heaps the maximum weights at 3 times the mean, while the Beaumont method does a more general shrinkage away from the tail. Figure 4 shows the population distribution of the sampling weights and the outcome; it suggests that the underlying Horvitz–Thompson "model", (1.6), is a reasonable approximation in this setting, and consequently it has little bias and stable variance, although the GREG estimator also performs reasonably well. The Hájek estimator, which ignores the heteroscedacity in Figure 4, has larger degrees of bias, with the Beaumont and trimming estimators also performing relatively poorly. (Note that the point estimate for the calibrated Horvitz–Thompson estimator corresponds to the Hájek estimator, since the calibrated weights sum to the population size.) PSPP-HET continues to outperform the other methods in this example, although it closely approximates the Horvitz–Thompson estimator in this setting.

TABLE 2
Estimated mean number of free lunch students in Texas public schools by various estimators (95% *CIs in parenthesis*). *Population mean* = 325.05

Estimator	Sample weighted	Calibration weighted
HT	321.09 (298.81, 343.36)	364.86 (325.86, 403.86)
HT-ROB	321.18	NA
HA	371.53 (333.62, 409.44)	364.86 (325.86, 403.86)
HA-ROB	372.88	NA
TRIM	374.90 (338.10, 411.70)	372.04 (334.86, 409.22)
PSPP-HOM	340.57 (286.91, 395.93)	NA
PSPP-HET	322.79 (299.62, 346.60)	NA
BEAUMONT	371.56 (336.40, 406.72)	364.86 (330.21, 399,51)
GREG	329.43 (285.27, 373.59)	328.37
GREG-ROB	329.50	NA

6. DISCUSSION AND CONCLUSIONS

This article has mainly focused on the estimation of the finite population mean of y, leaving consideration of descriptive inference for other finite population parameters and analytic inference for future study.

Most of the estimators considered in detail and in the simulation study could be regarded as modifications of basic Horvitz–Thompson or Hájek estimators. One advantage of these two basic forms is that they are computable as a weighted sum, with the same form for any variable y, using weights which may be provided with the survey file. The same is true of the GREG estimator of (1.7) and (1.8). The methods of trimming large weights described in Section 2.2.1 have this same property, because the weights $\tilde{w}_i$ in $\hat{t}_{\text{TR}}$ do not depend on y_i.

Some of the other estimators are of a similar form but require computation by the user to obtain y-specific weights or adjustments. In the case of the winsorization estimator $\hat{t}_{\text{WIN}}$ of Section 2.2.2, the estimator is a weighted sum, but the weight $\tilde{w}_i$ in (2.2) depends on y_i. Beaumont's estimator of Section 2.3 also is a weighted sum, where the weights $\hat{d}_i$ depend not only on y_i but on functions of the sampled values of y. The robust Horvitz–Thompson estimator of (2.10) in Section 2.2.3 is a weighted sum minus a sample-based constant correction for conditional biases which is dependent on sampled values of y; the robust GREG estimator of Section 2.2.4 has this form as well.

The PSPP estimators of Section 2.4 use the survey weights through a model of the y variable as a function of the design inclusion probabilities, and thus do not employ them as weights in the traditional sense.

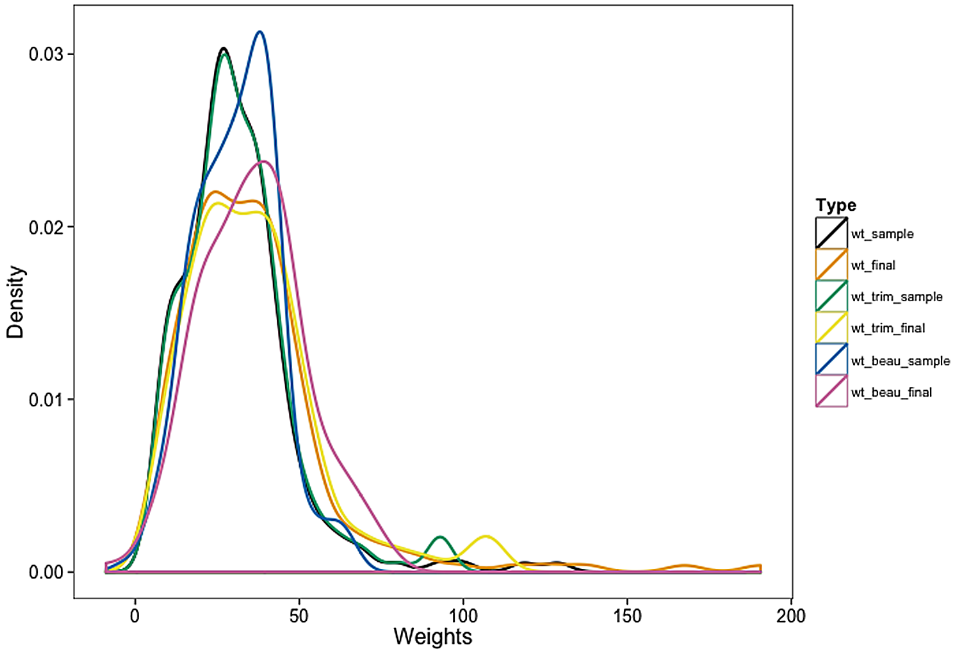

FIG. 3. *Distribution of weights: uncalibrated (wt_sample = HT weights; wt_trim_sample = trimmed to a normalized maximum of* 3*; wt_beau_sample = Beaumont modeling adjustment) and calibrated to location (wt_final = HT weights; wt_trim_final = trimmed to a normalized maximum of* 3*; wt_beau_final = Beaumont modeling adjustment).*

The PSPP estimators used here require knowledge of the inclusion probabilities for both sampled and nonsampled units, although extensions in which the sampling probabilities are observed only for the sample are available (Zangeneh and Little, 2015). The GREG and robust GREG estimators require knowing the x values for the sampled units and (only) the population total of x.

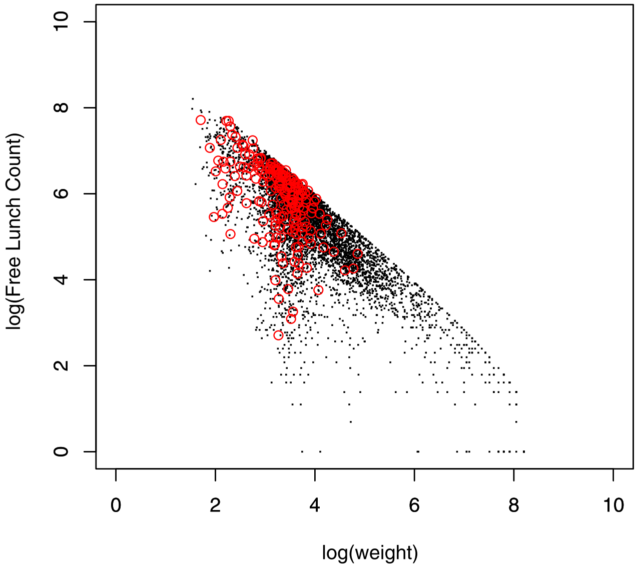

FIG. 4. *Number of free lunch students by sampling weight; nonsampled (black dots) and sampled (red circles).*

The simulation scenarios are special in that the size variable in PPS sampling, proportional to the inclusion probabilities, is also the principal auxiliary variable. The purpose was to separate the performance of the estimators, and to be favorable or unfavorable to specific estimators. That being said, there are practical scenarios where any procedure does poorly. Another limitation is that the true models for continuous y in terms of x or π assume linear, quadratic, or exponential forms, and the error term is normal in all cases. Thus, cases where the error term has a skewed distribution, so that it might be profitable to employ a transformation on the y variable, have been excluded. Finally, the population size (20,000) in each case is large enough that there is essentially no difference between the finite population mean and its expectation under the model; the sample size (200) is large enough that simple estimators of variance suffice; at the same time the ratio of the sample size to the population size (0.01) is small.

The following are general recommendations:

- In choosing or constructing an estimator of the finite population mean of y, it is important to pay attention to the relationship between y (or its mean function) and π, along with other auxiliary information.

- It is also important to examine the distribution of π to detect very high weight variability, and possible outlier weights.
- Trimming methods of the kind considered in the simulations to deal with outliers in y given x are seldom effective in estimation of the mean of y.
- Although the performance of the simplest estimators is good when the data are concordant with their associated models, this is not the case when there are outliers or significant deviations from such models.
- Some more sophisticated estimators/procedures are less dependent on closeness to associated models; in particular PSPP-HET and GREG-ROB are flexible all-purpose models.

ACKNOWLEDGMENTS

The authors thank Neung Soo Ha, Meena Khare and Yajuan Si for their participation in our working group, part of the 2013-4 SAMSI program, Computational Methods in the Social Sciences.

SUPPLEMENTARY MATERIAL

Supplement to "Approaches to Improving Survey-Weighted Estimates"
(DOI: 10.1214/17-STS609SUPP; .pdf). The Supplementary Material includes the density plots of the size variables, and dot plots summarizing the relative mean square errors, interval widths, and percent noncoverage for the methods in Section 4. Results of the simulations for binary outcomes, and plots corresponding to a thorough study of two scenarios with continuous outcomes are also presented.

REFERENCES

BASU, D. (1971). An essay on the logical foundations of survey sampling. Part I. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) 203–242. Holt, Rinehart and Winston, Toronto. MR0423625

BEAUMONT, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika* **95** 539–553. MR2443174

BEAUMONT, J.-F., HAZIZA, D. and RUIZ-GAZEN, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika* **100** 555–569. MR3094437

BENRUD, C. H. et al. (1978). Final report on national assessment of educational progress: sampling and weighting activities for assessment year 08. Research Triangle Park, North Carolina: National Assessment of Education Progress.

BINDER, D. A. (1982). Nonparametric Bayesian models for samples from finite populations. *J. R. Stat. Soc.*, *Ser. B.* **44** 388–393. MR0693238

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. R. Stat. Soc.*, *Ser. A* **143** 383–430. MR0603745

BREWER, K. R. W. (1979). A class of robust sampling designs for large-scale surveys. *J. Amer. Statist. Assoc.* **74** 911–915. MR0556487

CHEN, Q., ELLIOTT, M. R. and LITTLE, R. J. A. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Surv. Methodol.* **36** 23–34.

CHEN, Q., ELLIOTT, M. R. and LITTLE, R. J. A. (2012). Bayesian inference for finite population quantiles from unequal probability samples. *Surv. Methodol.* **38** 203–214.

CHEN, Q., ELLIOTT, M. R., HAZIZA, D., YANG, Y., GHOSH, M., LITTLE, R. J., SEDRANSK, J. and THOMPSON, M. (2017). Supplement to "Approaches to Improving Survey-Weighted Estimates." DOI:10.1214/17-STS609SUPP.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York. MR0474575

DALÉN, J. (1986). Sampling from finite populations: Actual coverage probabilities for confidence intervals on the population mean. *J. Off. Stat.* **2** 13–24.

DEVILLE, J.-C. and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87** 376–382. MR1173804

ELLIOTT, M. R. (2007). Bayesian weight trimming for generalized linear regression models. *Surv. Methodol.* **33** 23–34.

ELLIOTT, M. R. (2008). Model averaging methods for weight trimming. *J. Off. Stat.* **24** 517–540.

ELLIOTT, M. R. (2009). Model averaging methods for weight trimming in generalized linear regression models. *J. Off. Stat.* **25** 1–20.

ELLIOTT, M. R. and LITTLE, R. J. A. (2000). Model-based alternatives to trimming survey weights. *J. Off. Stat.* **16** 191–209.

ERICSON, W. A. (1969). Subjective Bayesian models in sampling finite populations. *J. R. Stat. Soc.*, *Ser. B.* **31** 195–233. MR0270494

ERICSON, W. A. (1988). Bayesian inference in finite populations. In *Handbook of Statistics* (P. R. C. Krishnaiah and C. R. Rao, eds.) 213–246. North-Holland, Amsterdam. MR1020084

FAVRE-MARTINOZ, C., HAZIZA, D. and BEAUMONT, J.-F. (2015). A method of determining the winsorization threshold, with an application to domain estimation. *Surv. Methodol.* **41** 57–77.

FAVRE-MARTINOZ, C., HAZIZA, D. and BEAUMONT, J.-F. (2016). Robust inference in two-phase sampling designs with application to unit nonresponse. *Scand. J. Stat.* **43** 1019–1034. MR3573673

FULLER, W. A. (2009). *Sampling Statistics*. Wiley, New York.

GELMAN, A. (2007). Struggles with survey weighting and regression modeling. *Statist. Sci.* **22** 153–164. MR2408951

GHOSH, M. and MEEDEN, G. (1986). Empirical Bayes estimation in finite population sampling. *J. Amer. Statist. Assoc.* **81** 1058–1062. MR0867632

GHOSH, M. and MEEDEN, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall, London.

GODAMBE, V. and THOMPSON, M. (1986). Parameters of superpopulation and survey populations: Their relationships and estimation. *Int. Stat. Rev.* **54** 127–138. MR1469494

HÁJEK, J. (1971). Comment on a paper by D. Basu. *Foundations of Statistical Inference* 236.

HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953). *Sample Survey Methods and Theory*. *Methods and Applications* **1**. Wiley, New York. MR0058171

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. *Monographs on Statistics and Applied Probability* **43**. Chapman & Hall, London. MR1082147

HAZIZA, D. and BEAUMONT, J. (2017). Construction of weights in surveys: A review. *Statist*. *Sci*. **32** 206–226.

HAZIZA, D., MECATTI, F. and RAO, J. N. K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron* **66** 91–108.

HENRY, K. and VALLIANT, R. (2012). Comparing alternative weight adjustment methods. In *Proceedings of the Section on Survey Research Methods* 4696–4710. Amer. Statist. Assoc., Alexandria, VA.

HOLT, D. and SMITH, T. M. F. (1979). Post stratification. *J*. *R*. *Stat*. *Soc*., *Ser*. *A* **142** 33–46.

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J*. *Amer*. *Statist*. *Assoc*. **47** 663–685. MR0053460

HULLIGER, B. (1995). Outlier robust Horvitz–Thompson estimator. *Surv*. *Methodol*. **21** 79–87.

ISAKI, C. T. and FULLER, W. A. (1982). Survey design under the regression superpopulation model. *J*. *Amer*. *Statist*. *Assoc*. **77** 89–96. MR0648029

KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist*. *Sci*. **22** 523–539. MR2420458

KIM, J. K. and SKINNER, C. J. (2013). Weighting in survey analysis under informative sampling. *Biometrika* **100** 385–398. MR3068441

KISH, L. (1995). The hundred years' wars of survey sampling. *Stat*. *Transit*. **2** 813–830.

KOKIC, P. (1998). On winsorization in business surveys. In *Proceedings of the Survey Methods Section* 237–239. Statistical Society of Canada, Ottawa.

KOKIC, P. N. and BELL, P. A. (1994). Optimal winsorising cutoffs for a stratified finite population estimator. *J*. *Off*. *Stat*. **10** 419–435.

LAZZERONI, L. C. and LITTLE, R. J. A. (1998). Random effects models for smoothing post-stratification weights. *J*. *Off*. *Stat*. **14** 61–78.

LITTLE, R. J. A. (1983). Comment on "An evaluation of model-dependent and probability sampling inferences in sample surveys" by M. H. Hansen, W. G. Madow and B. J. Tepping. *J*. *Amer*. *Statist*. *Assoc*. **78** 797–799.

LITTLE, R. J. A. (1991). Inference with survey weights. *J*. *Off*. *Stat*. **7** 405–424.

LITTLE, R. J. A. (2003). Bayesian methods for unit and item nonresponse. In *Analysis of Survey Data* (R. L. Chambers and C. J. Skinner, eds.) 289–306. Wiley, Chichester. MR1978857

LITTLE, R. J. A. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *Amer*. *Statist*. **60** 213–223.

LITTLE, R. J. A. (2012). Calibrated Bayes: An alternative inferential paradigm for official statistics. *J*. *Off*. *Stat*. **28** 309–372.

MATEI, A. and TILLÉ, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *J*. *Off*. *Stat*. **21** 543–570.

MOLINA, C. and SKINNER, C. (1992). Pseudo-likelihood and quasi-likelihood estimation for complex sampling schemes. *Comput*. *Statist*. *Data Anal*. **13** 395–405.

MORENO-REBOLLO, J. L., MUÑOZ-REYES, A. and MUÑOZ-PICHARDO, J. (1999). Influence diagnostic in survey sampling: Conditional bias. *Biometrika* **86** 923–928. MR1741987

MORENO-REBOLLO, J. L., MUÑOZ-REYES, A., JIMÉNEZ-GAMERO, M. D. and MUÑOZ-PICHARDO, J. (2002). Influence diagnostic in survey sampling: Estimating the conditional bias. *Metrika* **55** 209–214. MR1912392

NEYMAN, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J*. *Roy*. *Statist*. *Soc*. **97** 558–625. MR0121942

O'HAGAN, A. (1995). Fractional Bayes factors for model comparison. *J*. *R*. *Stat*. *Soc*., *Ser*. *B*. **57** 99–138. MR1325379

PENG, Y., LITTLE, R. J. A. and RAGHUNATHAN, T. E. (2004). An extended general location model for causal inferences from data subject to noncompliance and missing values. *Biometrics* **60** 598–607. MR2089434

PFEFFERMANN, D. and SVERCHKOV, M. (2009). Inference under informative sampling. In *Sample Surveys*: *Inference and Analysis* (D. Pfeffermann and C. R. Rao, eds.) **29B** 455–487. Elsevier, Amsterdam.

POTTER, F. (1988). Survey of procedures to control extreme sampling weights. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* 453–458. Amer. Statist. Assoc., Alexandria, VA.

POTTER, F. (1990). A study of procedures to identify and trim extreme sampling weights. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* 225–230.

RAO, J. N. K. (1966). Alternative estimators in sampling for multiple characteristics. *Sankhyā Ser*. *A* **28** 47–60. MR0258176

RAO, J. N. K. and WU, C.-F. J. (1988). Resampling inference with complex survey data. *J*. *Amer*. *Statist*. *Assoc*. **83** 231–241. MR0941020

RAO, J. N. K., WU, C. F. J. and YUE, K. (1992). Some recent work on resampling methods for complex surveys. *Surv*. *Methodol*. **18** 209–217.

RIVEST, L.-P. and HIDIROGLOU, M. (2004). Outlier treatment for disaggregated estimates. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* 4248–4256. Amer. Statist. Assoc., Alexandria, VA.

RIVEST, L.-P. and HURTUBISE, D. (1995). On Searls' winsorized means for skewed populations. *Surv*. *Methodol*. **21** 107–116.

ROYALL, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika* **57** 377–387.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann*. *Statist*. **6** 34–58. MR0472152

RUBIN, D. B. (1983). Comment on "An evaluation of model-dependent and probability-sampling inferences in sample surveys" by M. H. Hansen, W. G. Madow, and B. J. Tepping. *J*. *Amer*. *Statist*. *Assoc*. **78** 803–805.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519

SÄRNDAL, C.-E. (2007). The calibration approach in survey theory and practice. *Surv. Methodol.* **33** 99–119.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer, New York. MR1140409

SCOTT, A. J. (1977). On the problem of randomization in survey sampling. *Sankhyā* **39** 1–9.

SCOTT, A. J. and SMITH, T. M. F. (1969). Estimation in multistage surveys. *J. Amer. Statist. Assoc.* **64** 830–840.

SUGDEN, R. A. and SMITH, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika* **71** 495–506. MR0775395

TAMBAY, J. L. (1988). Integrated approach for the treatment of outliers in sub-annual economic surveys. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* 229–234. Amer. Statist. Assoc., Alexandra, VA.

THOMPSON, M. E. (1997). *Theory of Sample Surveys*. *Monographs on Statistics and Applied Probability* **74**. Chapman & Hall, London. MR1462619

TILLÉ, Y. (2017). Sampling designs: New methods and guidelines. *Statist. Sci.* To appear.

VALLIANT, R., DORFMAN, A. H. and ROYALL, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, New York. MR1784794

WAHBA, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. R. Stat. Soc., Ser. B.* **40** 364–372. MR0522220

WU, C.-F. and DENG, L.-Y. (1983). Estimation of variance of the ratio estimator: An empirical study. In *Scientific Inference, Data Analysis and Robustness* (G. E. P. Box, T. Leonard and C. F. J. Wu, eds.) 245–277. Academic Press, Orlando, FL. MR0772772

WU, C. and LU, W. W. (2016). Calibration weighting methods for complex surveys. *Int. Stat. Rev.* **84** 79–98. MR3491280

ZANGENEH, S. Z. and LITTLE, R. J. A. (2015). Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample. *Journal of Survey Statistics and Methodology* **3** 162–192.

ZHENG, H. and LITTLE, R. J. A. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *J. Off. Stat.* **19** 99–117.

ZHENG, H. and LITTLE, R. J. A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *J. Off. Stat.* **21** 1–20.