Published in Biometrika 95, issue 3, 521–537, 2008
which should be used for any reference to this work

Optimal sampling and estimation strategies under the linear model

BY DESISLAVA NEDYALKOVA AND YVES TILL

Institute of Statistics, University of Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel, Switzerland
desislava.nedyalkova@unine.ch yves.tille@unine.ch

SUMMARY

In some cases model-based and model-assisted inferences can lead to very different estimators. These two paradigms are not so different if we search for an optimal strategy rather than just an optimal estimator, a strategy being a pair composed of a sampling design and an estimator. We show that, under a linear model, the optimal model-assisted strategy consists of a balanced sampling design with inclusion probabilities that are proportional to the standard deviations of the errors of the model and the Horvitz–Thompson estimator. If the heteroscedasticity of the model is 'fully explainable' by the auxiliary variables, then this strategy is also optimal in a model-based sense. Moreover, under balanced sampling and with inclusion probabilities that are proportional to the standard deviation of the model, the best linear unbiased estimator and the Horvitz–Thompson estimator are equal. Finally, it is possible to construct a single estimator for both the design and model variance. The inference can thus be valid under the sampling design and under the model.

Some key words: Balanced sampling; Design-based inference; Finite population sampling; Fully explainable heteroscedasticity; Model-assisted inference; Model-based inference; Optimal strategy.

1. INTRODUCTION

In survey sampling theory there have long been contrasting views on which approach to use in order to obtain a valid inference in estimating population totals: a prediction theory based on a superpopulation model or a probability sampling theory based on a sampling design. Neither of these paradigms is false. Numerous articles compare the two approaches (Brewer, 1994, 1999b, 2002; B r e w e r e t a l . , 1988; Hansen et al., 1983; Iachan, 1984; Royall, 1988; Smith, 1976, 1984, 1994). Valliant et al. (2000, p. 14), who favour the model-based theory, say that 'there is no doubt of the mathematical validity of either of the two theories'. Nevertheless, we believe that the choice between them depends on the point of view of the analyst.

In the model-based, or prediction, approach studied by Royall (1976, 1992), Royall & Cumberland (1981) and Chambers (1996), the optimality is conceived only with respect to the regression model without taking into account the sampling design. Royall (1976) proposed the use of the best linear unbiased predictor when the data are assumed to follow a linear model. Royall (1992) showed that under certain conditions there exists a lower bound for the error variance of the best linear unbiased predictor, and that this bound is only achieved when the sample is balanced. Royall & Herson (1973a,b) and Scott et al. (1978) discussed the importance of balanced sampling in order to protect the inference against a misspecifie model. These authors conclude that the sample must be balanced, but not necessarily random.

In the model-assisted approach advocated by S¨arndal et al. (1992), the estimator must be approximately design-unbiased under the sampling design. The generalized regression estimator

uses auxiliary information from the linear model, but is approximately design-unbiased. Deville & Särndal (1992) proposed a purely design-based methodology that takes into account auxiliary information without considering a model. The main difference between the design-based and the model-based approaches arises because the statistical properties of an estimator are evaluated with respect to the sampling design and not with respect to the model.

Recently, Deville & Tillé (2004) developed the cube method, an algorithm that can select randomly balanced samples and that satisfie exactly the given inclusion probabilities. In the model-based framework, balanced samples are essential for achieving the lower bound for the error variance proposed by Royall (1992). Moreover, it can be shown that balanced sampling is also optimal under model-assisted inference. H´ajek (1981) define a strategy as a pair comprising a sampling design and an estimator. The purpose of this paper is to show that, if we search for an optimal strategy rather than just an optimal estimator, most of the differences between model-based and model-assisted inferences can be reconciled.

2. Notation and definitions

We consider a finit population U of size N. Each unit of the population can be identifie by a label $k = 1, \ldots, N$. Let $x_k = (x_{k1}, \ldots, x_{kq})'$ be the vector of the values of q auxiliary variables for unit k, for $k = 1, \ldots, N$, and let $X = \sum_{k\in U} x_k$ be the vector of totals, which is also known. The values $y_1, \ldots, y_N$ of the variables of interest are unknown. The aim is to estimate the population total $Y = \sum_{k\in U} y_k$. A sample s is a subset of the population U. Let $p(s)$ denote the probability of selecting the sample s, S being the random sample such that $p(s) = \mathrm{pr}(S = s)$ and let $n(S)$ be the size of the sample S. The expected sample size is $n = E_p\{n(S)\}$, where E_p denotes the expected value under the sampling design $p(\cdot)$. Let $\bar{S}$ denote the set of units of the population which are not in S. Let $\pi_k = \mathrm{pr}(k \in S)$ denote the inclusion probability of unit k, and let $\pi_{k\ell} = \mathrm{pr}(k \in S \text{ and } \ell \in S)$ denote, for $k \neq \ell$, the joint inclusion probability of units k and ℓ. The variable y is observed on the sample only.

Under model-based inference, the values $y_1, \ldots, y_N$ are assumed to be the realization of a superpopulation model ξ. The model which we will study is the general linear model with uncorrelated errors, given by

$$y_k = x_k'\beta + \varepsilon_k, \tag{1}$$

where the x_ks are not random, $\beta = (\beta_1, \ldots, \beta_q)'$, $E_\xi(\varepsilon_k) = 0$, $\mathrm{var}_\xi(\varepsilon_k) = v_k^2\sigma^2$, for all $k \in U$, and $\mathrm{cov}_\xi(\varepsilon_k, \varepsilon_\ell) = 0$, when $k \neq \ell \in U$. The quantities $v_k, k \in U$, are assumed known. Moreover, we scale them so that $\sum_{k\in U} v_k = N$. The superpopulation model (1) includes the possibility of heteroscedasticity. Under homoscedasticity, $v_k = 1$ for all $k \in U$. An important and common hypothesis is that the random sample S and the errors ε_k of (1) are independent. The symbols E_ξ, var_ξ and cov_ξ denote, respectively, expected value, variance and covariance under the model.

In order to estimate the total Y, we will only use linear estimators which can be written as

$$\hat{Y}_w = \sum\nolimits_{k\in S} w_{kS} y_k = \sum\nolimits_{k\in U} w_{kS} y_k I_k,$$

where the $w_{kS}, k \in S$ are weights that can depend on the sample, and where I_k is equal to 1 if $k \in S$ and equal to 0 if $k \notin S$.

Definition 1 (Hájek, 1981, p. 153). *A strategy is a pair* $\{p(\cdot), \hat{Y}\}$ *comprising a sampling design and an estimator.*

Definition 2. *An estimator* $\hat{Y}$ *is said to be model-unbiased if* $E_\xi(\hat{Y} - Y) = 0$.

DEFINITION 3. *An estimator $\hat{Y}$ is said to be design-unbiased if $E_p(\hat{Y}) - Y = 0$.*

DEFINITION 4. *A linear estimator $\hat{Y}_w$ is said to be calibrated on a set of auxiliary variables x_k if and only if its weights satisfy*

$$\sum_{k\in S} w_{kS}x_k = \sum_{k\in U} x_k.$$

DEFINITION 5. *The design variance of an estimator $\hat{Y}$ is define by*

$$\mathrm{var}_p(\hat{Y}) = E_p\{\hat{Y} - E_p(\hat{Y})\}^2.$$

DEFINITION 6. *The design mean-squared error of an estimator $\hat{Y}$ is define by*

$$\mathrm{MSE}_p(\hat{Y}) = E_p(\hat{Y} - Y)^2.$$

DEFINITION 7. *The model variance of an estimator $\hat{Y}$ is define by*

$$\mathrm{var}_\xi(\hat{Y}) = E_\xi\{\hat{Y} - E_\xi(\hat{Y})\}^2.$$

DEFINITION 8. *The model mean-squared error of an estimator $\hat{Y}$ is define by*

$$E_\xi(\hat{Y} - Y)^2.$$

The model mean-squared error is sometimes called the error variance. The model mean-squared error of an estimator $\hat{Y}$ is generally smaller than its model variance because $\hat{Y}$ is closer to Y than to $E_\xi(\hat{Y})$.

DEFINITION 9. *The anticipated mean-squared error of an estimator $\hat{Y}$ is define by*

$$\mathrm{MSE}_{p\xi}(\hat{Y}) = E_pE_\xi(\hat{Y} - Y)^2 = E_\xi E_p(\hat{Y} - Y)^2.$$

The anticipated mean-squared error is also called the anticipated variance, for example, by Isaki & Fuller (1982).

3. LINEAR ESTIMATORS

Consider the class of linear estimators, $\hat{Y}_w = \sum_{k\in S} w_{kS}y_k$. For all $k \in U$, define $C_k = E_p(w_{kS}\, I_k) = \pi_k E_p(w_{kS} \mid I_k = 1)$. Godambe (1955) showed that $\hat{Y}_w$ is design-unbiased if and only if $C_k = 1$ or, equivalently, if $E_p(w_{kS} \mid I_k = 1) = 1/\pi_k$. Moreover, its model bias is

$$E_\xi(\hat{Y}_w - Y) = \sum_{k\in S} w_{kS}x_k'\beta - \sum_{k\in U} x_k'\beta,$$

for any value of $\beta \in \mathbb{R}^q$. Therefore, for the class of linear estimators under the linear model ξ, the definition of a model-unbiased and a calibrated estimator are equivalent. For any linear estimator, a general expression of the anticipated mean-squared error can be given.

RESULT 1. *If $\hat{Y}_w$ is a linear estimator, then*

$$\begin{aligned}
&E_pE_\xi(\hat{Y}_w - Y)^2 \\
&= \sigma^2 E_p\left\{\sum_{k\in S}(w_{kS} - 1)^2\nu_k^2 + \sum_{k\in\bar{S}}\nu_k^2\right\} + E_p\left(\sum_{k\in S} w_{kS}x_k'\beta - \sum_{k\in U} x_k'\beta\right)^2 \\
&= \sigma^2\sum_{k\in U}\nu_k^2\left\{C_k^2\frac{1-\pi_k}{\pi_k} + \pi_k\mathrm{var}_p(w_{kS} \mid I_k = 1) + (C_k - 1)^2\right\} \\
&\quad + \mathrm{var}_p\left(\sum_{k\in S} w_{kS}x_k'\beta\right) + \left(\sum_{k\in U} C_kx_k'\beta - \sum_{k\in U} x_k'\beta\right)^2.
\end{aligned}$$

The proof is given in the Appendix.

The anticipated mean-squared error $E_p E_\xi(\hat{Y}_w - Y)^2$ is the sum of fi e nonnegative terms,

$$E_p E_\xi(\hat{Y}_w - Y)^2 = A + B + C + D + E, \tag{2}$$

where

$$A = \sigma^2 \sum_{k\in U} \nu_k^2 C_k^2 \frac{1-\pi_k}{\pi_k}, \quad B = \sigma^2 \sum_{k\in U} \nu_k^2 \pi_k \mathrm{var}_p(w_{kS} \mid I_k = 1),$$
$$C = \sigma^2 \sum_{k\in U} \nu_k^2 (C_k - 1)^2, \quad D = \mathrm{var}_p\left(\sum_{k\in S} w_{kS} x_k' \beta\right),$$
$$E = \left(\sum_{k\in U} C_k x_k' \beta - \sum_{k\in U} x_k' \beta\right)^2.$$

Term A is a part of the anticipated mean-squared error; it depends on the inclusion probabilities and the variance of the errors. Term B is only relevant if the weights w_{kS} differ from sample to sample. Term C depends on the design bias and the variance of the errors of the model; it is null if the estimator is design-unbiased. Term D is the design variance of the model expectation of the estimator; it is null when the estimator is calibrated, or model-unbiased. Term E is the square of the design bias of the model expectation of the estimator; it is also null when the estimator is calibrated, or model-unbiased or when the estimator is design-unbiased.

Some particular cases of Result 1 are interesting.

COROLLARY 1. *If $\hat{Y}_w$ is a model-unbiased linear estimator, or a calibrated estimator, then* $E_p E_\xi(\hat{Y}_w - Y)^2 = A + B + C$.

COROLLARY 2. *If $\hat{Y}_w$ is a design-unbiased linear estimator, then $C_k = 1$ for all k in U and* $E_p E_\xi(\hat{Y}_w - Y)^2 = A + B + D$.

COROLLARY 3. *If $\hat{Y}_w$ is a design-unbiased linear estimator with weights w_{ks} that are constant from sample to sample, then $C_k = 1$, for all k in U, and* $E_p E_\xi(\hat{Y}_w - Y)^2 = A + D$.

COROLLARY 4. *If $\hat{Y}_w$ is a design-unbiased and model-unbiased linear estimator, then* $E_p E_\xi(\hat{Y}_w - Y)^2 = A + B$.

Example 1. The Horvitz–Thompson estimator, given by

$$\hat{Y}_\pi = \sum_{k\in S} \frac{y_k}{\pi_k},$$

is linear and design-unbiased when $\pi_k > 0$, for all $k \in U$, because

$$E_p(\hat{Y}_\pi) = \sum_{k\in U} \frac{y_k}{\pi_k} E(I_k) = Y.$$

Under any sampling design, the design variance of this estimator is

$$\mathrm{var}_p(\hat{Y}_\pi) = \sum_{k\in U} \sum_{\ell\in U} \frac{y_k}{\pi_k} \Delta_{k\ell} \frac{y_\ell}{\pi_\ell}, \tag{3}$$

where $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell,\ k, \ell \in U$. The Horvitz–Thompson estimator is, however, model-biased and its bias is

$$E_\xi(\hat{Y}_\pi - Y) = \left(\sum_{k\in S} \frac{x_k'}{\pi_k} - \sum_{k\in U} x_k'\right) \beta. \tag{4}$$

Since the Horvitz–Thompson estimator is design-unbiased with weights $w_{ks} = 1/\pi_k$ that are constant from sample to sample, its anticipated mean-squared error can be deduced from Corollary 3,

$$E_p E_\xi (\hat{Y}_\pi - Y)^2 = A + D = \sigma^2 \sum\nolimits_{k\in U} v_k^2 \frac{1-\pi_k}{\pi_k} + \sum\nolimits_{k\in U} \sum\nolimits_{\ell\in U} \frac{x_k'\beta}{\pi_k} \Delta_{k\ell} \frac{x_\ell'\beta}{\pi_\ell}.$$

4. Balanced sampling

There exist several different definitions of the concept of balancing. A first definition of a balanced sample is that the sample mean is equal to the population mean. According to this definition balancing is a property of a sample and a balanced sample can be constructed deliberately and deterministically without reference to a random procedure. A balanced sample is then associated with the purposive selection and is thus in contradiction to the random selection of the sample (Brewer, 1999b).

A balanced sample can also be selected randomly by a procedure called a balanced sampling design. According to the definition of Deville & Tillé (2004), a sampling design $p(\cdot)$ is said to be balanced on the auxiliary variables $x_1, \ldots, x_q$ if the Horvitz–Thompson estimator satisfies the relationship

$$\hat{X}_\pi = \sum\nolimits_{k\in S} \frac{x_k}{\pi_k} = \sum\nolimits_{k\in U} x_k = X. \tag{5}$$

Authors such as Cumberland & Royall (1981) and Kott (1986) would call this a 'π-balanced sampling', as opposed to a mean-balanced sampling defined by the equation

$$\frac{1}{n} \sum\nolimits_{k\in S} x_k = \frac{1}{N} \sum\nolimits_{k\in U} x_k.$$

Below, we use the expression 'balanced sampling' to denote a sampling design that satisfies equation (5) for one or more auxiliary variables, a mean-balanced sampling being a particular case of this balanced sampling when the sample is selected with inclusion probabilities n/N.

The definition of balanced sampling includes the definition of sampling with fixed sample size. Suppose that one of the balancing variables is proportional to the inclusion probabilities or, more generally, that there exists a vector λ such that $\lambda' x_k = \pi_k$, for all $k \in U$. In this case, the balancing equation

$$\sum\nolimits_{k\in S} \frac{x_k}{\pi_k} = \sum\nolimits_{k\in U} x_k$$

becomes for this variable, by multiplication by λ',

$$\sum\nolimits_{k\in S} \frac{\pi_k}{\pi_k} = \sum\nolimits_{k\in U} \pi_k,$$

or equivalently,

$$\sum\nolimits_{k\in S} 1 = \sum\nolimits_{k\in U} \pi_k,$$

which means that the sample size must be fixed. In practice, it is always recommended to add the vector of inclusion probabilities in the balancing variables, because this allows one to fix the sample size and thus the cost of the survey.

If a sampling design is balanced on the auxiliary variables, then $\hat{X}_\pi$ is not a random variable. For a long time, balanced samples were considered difficult to construct, except for particular special cases such as sampling with fixed sample size or stratification. Partial procedures of balanced

sampling have been proposed by Yates (1946), Thionet (1953), Deville et al. (1988), Ardilly (1991), Deville (1992) and Hedayat & Majumdar (1995), and a list of methods for constructing balanced samples is given in Valliant et al. (2000, pp. 65–78). Several of these methods are rejective: they consist of generating randomly a sequence of samples with an original sampling design until a sample is obtained that is sufficientl well balanced. Rejective methods are actually a way of constructing a conditional sampling design and have the important drawback that the inclusion probabilities of the balanced design are not necessarily the same as the inclusion probabilities of the original design. Moreover, if the number of balancing variables is large, rejective methods can be very slow.

The cube method, proposed by Deville & Tillé (2004), is a non-rejective procedure that directly allows the random selection of balanced or nearly balanced samples and that satisfie exactly the given first-orde inclusion probabilities. The cube method works with equal or unequal inclusion probabilities (Tillé, 2006, pp. 147–76). If one of the balancing variables is proportional to the inclusion probabilities, then the cube method will produce samples of fixe size. However, it is not always possible for such a sample to be exactly balanced because of the rounding problem. For instance, in proportional stratification which is a particular case of balanced sampling, it is generally impossible to select an exactly balanced sample because the sample sizes of the strata, $n_h = nN_h/N$, are seldom integers. Deville & Till´e (2004) also showed that the rounding problem, under reasonable hypotheses, is bounded by $O(q/n)$, where q is the number of balancing variables and n is the sample size. Thus, the rounding problem becomes negligible if the sample size is reasonably large relative to the number of balancing variables.

Under model (1) and balanced sampling, the Horvitz–Thompson estimator is model-unbiased. Indeed, by equations (4) and (5), it follows that

$$E_\xi(\hat{Y}_\pi - Y) = \left(\sum_{k\in S} \frac{x_k}{\pi_k} - \sum_{k\in U} x_k\right)' \beta = 0.$$

Under model (1) and balanced sampling, we can compute the error variance and the anticipated mean-squared error of the Horvitz–Thompson estimator.

RESULT 2. *Under model* (1), *if the sample is balanced on* x_k *and selected with inclusion probabilities* π_k, *then*

$$E_p E_\xi(\hat{Y}_\pi - Y)^2 = \sigma^2 \sum_{k\in U} v_k^2 \frac{1-\pi_k}{\pi_k}.$$

The proof is given in the Appendix.

If we fi the inclusion probabilities, then the expectation of the sample size is also fixed The design mean-squared error of a balanced sampling design is, unfortunately, more difficul to determine. In their Method 4, Deville & Tillé (2005) have proposed the following approximation of the design variance given in (3):

$$\mathrm{var}_p(\hat{Y}_\pi) \doteq \mathrm{var}_{\mathrm{app}}(\hat{Y}_\pi) = \sum_{k\in U} d_k \frac{(y_k - x_k' b)^2}{\pi_k^2}, \tag{6}$$

where

$$b = \left(\sum_{k\in U} d_k \frac{x_k x_k'}{\pi_k^2}\right)^{-1} \sum_{k\in U} d_k \frac{x_k y_k}{\pi_k^2},$$

and the d_k are the solutions of the nonlinear system

$$\pi_k(1-\pi_k) = d_k - \frac{d_k x_k'}{\pi_k}\left(\sum_{\ell\in U} d_\ell \frac{x_\ell x_\ell'}{\pi_\ell^2}\right)^{-1}\frac{d_k x_k}{\pi_k}, \quad k \in U. \tag{7}$$

This approximation, which uses only the first-orde inclusion probabilities, was validated by Deville & Tillé (2005) under a variety of balanced samples regardless of how the y-values were generated. An additional argument in favour of using this approximation is that its model expectation is equal to its anticipated mean-squared error, as we see below.

RESULT 3. *Under model* (1), *if the sample is balanced on* x_k, *then*

$$E_\xi\{\mathrm{var}_{\mathrm{app}}(\hat{Y}_\pi)\} = E_p E_\xi(\hat{Y}_\pi - Y)^2.$$

The proof is given in the Appendix.

5. THE MODEL-ASSISTED APPROACH

One approach to estimating Y consists of findin the 'best' strategy that provides a valid inference under the sampling design. Godambe (1955) showed that there is no optimal estimator in the class of linear estimators for all $y_1, \ldots, y_N$ that minimizes the design mean-squared error. It is, however, not possible to determine an optimal design-based strategy without formalizing the link between the auxiliary variables x_k and the variables of interest y_k . A model must therefore be used to guide the choice of the estimator. S¨arndal et al. (1992) proposed the concept of 'model-assisted inference'. To be model-assisted, the estimator must be chosen so that it leads to a valid inference with respect to the sampling design, even if the model is misspecified In order to make the inference, we need to estimate $E_p(\hat{Y}_w - Y)^2$, but in order to fin the optimal strategy, we need to minimize $E_\xi E_p(\hat{Y}_w - Y)^2$ under the constraint that the estimator is design-unbiased or that its design bias is small with respect to its design mean-squared error.

A bound for the model-assisted strategy given by Godambe & Joshi (1965) for a set of fixe inclusion probabilities can be derived directly from Corollary 2. I f $\hat{Y}_w$ is a design-unbiased linear estimator, then

$$E_p E_\xi(\hat{Y}_w - Y)^2 \geqslant L_p = \sigma^2 \sum_{k\in U} v_k^2 \frac{1-\pi_k}{\pi_k}. \tag{8}$$

If we suppose at least tentatively that the v_k are known, a judicious choice of the inclusion probabilities allows a smaller anticipated mean-squared error to be determined. If we minimize L_p in π_k subject to

$$\sum_{k\in U} \pi_k = n, \quad 0 \leqslant \pi_k \leqslant 1, \tag{9}$$

for all k in U, then we obtain the optimal inclusion probabilities $\pi_k^* = \min(1, \alpha v_k/N)$, where α is such that

$$\sum_{k\in U} \min\left(1, \frac{\alpha v_k}{N}\right) = n.$$

The following general result gives a bound for any design-unbiased strategy with a sample size n.

RESULT 4. *For any design-unbiased strategy,*

$$\begin{aligned} E_p E_\xi(\hat{Y}_w - Y)^2 &\geqslant L_p = \sigma^2 \sum\nolimits_{k\in U} v_k^2 \frac{1-\pi_k}{\pi_k} \\ &\geqslant \sigma^2 \sum\nolimits_{k\in U} v_k^2 \frac{1-\pi_k^*}{\pi_k^*} = \sigma^2 \left(\frac{N}{\alpha} \sum\nolimits_{\substack{k\in U \\ \pi_k^*<1}} v_k - \sum\nolimits_{\substack{k\in U \\ \pi_k^*<1}} v_k^2 \right) \\ &\geqslant \sigma^2 \left(\frac{N^2}{n} - \sum\nolimits_{k\in U} v_k^2 \right) = \sigma^2 N^2 \frac{N-n}{Nn} - \sigma^2 \sum\nolimits_{k\in U} (v_k - 1)^2 . \end{aligned}$$

The proof is given in the Appendix.

DEFINITION 10. *An optimal model-assisted strategy is one with a design-unbiased estimator that, subject to* (9)*, minimizes the anticipated mean-squared error of that estimator.*

From § 4 and Result 4, we obtain directly an optimal model-assisted strategy.

STRATEGY 1. *Under the superpopulation model* (1)*, an optimal model-assisted strategy consists of using inclusion probabilities that are proportional to* v_k *subject to* (9)*, selecting the sample by means of a balanced sampling design on* x_k *, and using the Horvitz–Thompson estimator.*

6. THE MODEL-BASED APPROACH

Under the model-based approach, the aim is to fin a strategy that leads to a valid inference with respect to the model, i.e. a model-unbiased or approximately model-unbiased estimator and a sample that minimizes the error variance $E_\xi(\hat{Y} - Y)^2$.

DEFINITION 11. *An optimal model-based strategy is one with a linear model-unbiased estimator that, subject to a fixe sample size n, minimizes the error variance of that estimator.*

In the model-based approach, this strategy is strictly applied under ideal circumstances, which occur when the model is known to hold. In practice, the modeller must bear model failure in mind, and the model-based approach strongly emphasizes robustness to deviations from the working model. The strictly optimal strategies that are not robust in case of misspecificatio of the model are thus clearly rejected.

A well-known result (Royall, 1976) is that the model-unbiased linear estimator of Y that minimizes the error variance is the best linear unbiased estimator

$$\hat{Y}_{\text{BLU}} = \sum\nolimits_{k\in S} y_k + \sum\nolimits_{k\in \bar{S}} x_k' \hat{\beta}_{\text{BLU}},$$

where $\hat{\beta}_{\text{BLU}}$ is the weighted least-squares estimator of the regression coefficient vector β

$$\hat{\beta}_{\text{BLU}} = A^{-1} \sum\nolimits_{k\in S} \frac{x_k y_k}{v_k^2},$$

where

$$A = \sum\nolimits_{k\in S} \frac{x_k x_k'}{v_k^2}.$$

The error variance of the best linear unbiased estimator is

$$E_\xi(\hat{Y}_{\text{BLU}} - Y)^2 = \sigma^2 \left(\sum\nolimits_{k\in \bar{S}} x_k' A^{-1} \sum\nolimits_{\ell\in \bar{S}} x_\ell + \sum\nolimits_{k\in \bar{S}} v_k^2 \right). \tag{10}$$

Consequently, to determine a model-based strategy, we look for a sample s that minimizes (10), this sample being not necessarily unique.

STRATEGY 2. *Under the superpopulation model* (1)*, an optimal model-unbiased strategy consists of using the best linear unbiased estimator, and choosing a sample of size n that minimizes expression* (10)*.*

Again, this strategy must be put into perspective with respect to possible misspecificatio of the model. If the sample that minimizes (10) is very particular, then a more robust strategy should be considered.

With certain superpopulation models, expression (10) can be considerably simplified More-over, minimizing the anticipated mean-squared error given in (11) below in the class of linear model-unbiased estimators also leads to Strategy 2,

$$E_p E_\xi(\hat{Y}_{\text{BLU}} - Y)^2 = \sigma^2 \left\{ E_p \left(\sum_{k \in \bar{S}} x_k' A^{-1} \sum_{\ell \in \bar{S}} x_\ell \right) + \sum_{k \in U} (1 - \pi_k) v_k^2 \right\}. \tag{11}$$

Unfortunately, expression (11) cannot be much simplified

DEFINITION 12. *Model* (1) *is said to have fully explainable heteroscedasticity if*

(i) *there exists a vector* $\lambda \in \mathbb{R}^q$ *such that* $\lambda' x_k = v_k^2$;
(ii) *there exists a vector* $\theta \in \mathbb{R}^q$ *such that* $\theta' x_k = v_k$.

RESULT 5 (*Royall,* 1992). *If the superpopulation model* (1) *is such that condition* (i) *of Definitio* 12 *is met, then* $\hat{Y}_{\text{BLU}} = \sum_{k \in U} x_k' \beta_{\text{BLU}}$, *and* $E_\xi(Y_{\text{BLU}} - Y) = \sigma^2(X' A^{-1} X - \sum_{k \in U} v_k)$.

RESULT 6 (*Royall,* 1992). *If the superpopulation model* (1) *has fully explainable hetero-scedasticity, then*

$$E_\xi(\hat{Y}_{\text{BLU}} - Y)^2 \geqslant \sigma^2 \left(\frac{N^2}{n} - \sum_{k \in U} v_k^2 \right),$$

and, if the sample is such that

$$\frac{1}{n} \sum_{k \in S} \frac{x_k}{v_k} = \frac{\sum_{k \in U} x_k}{N},$$

then the bound for the error variance is achieved.

Royall (1992) and later Valliant et al. (2000, pp. 98–100) in their Theorem 4·2·1 and consequent Remark 4 present results which from a design-based point of view can be used to prove the following result.

RESULT 7. *If the superpopulation model* (1) *has fully explainable heteroscedasticity and if the sample is balanced with inclusion probabilities proportional to* v_k*, then the best linear unbiased estimator* $\hat{Y}_{\text{BLU}}$ *equals the Horvitz–Thompson estimator* $\hat{Y}_\pi$ *and the bound for the error variance is achieved.*

Under the conditions of Result 7, $E_\xi(\hat{Y}_\pi - Y)^2 = E_p E_\xi(\hat{Y}_\pi - Y)^2$.

7. A COMBINED MODEL-BASED AND MODEL-ASSISTED APPROACH

A third option for estimating Y consists of findin a strategy that is simultaneously design-unbiased and model-unbiased. From Corollary 4, we know that such a strategy has an anticipated

mean-squared error equal to

$$E_p E_\xi(\hat{Y}_w - Y)^2 = \sigma^2 \sum_{k\in U} \nu_k^2 \left\{ \pi_k \mathrm{var}_p(w_{kS} \mid I_k = 1) + \frac{1-\pi_k}{\pi_k} \right\}.$$

If the weights w_{ks} are not random, then we obtain the Godambe–Joshi bound

$$E_p E_\xi(\hat{Y}_w - Y)^2 \geqslant L_p = \sigma^2 \sum_{k\in U} \nu_k^2 \frac{1-\pi_k}{\pi_k}. \tag{12}$$

Thus, an optimal strategy that is at the same time model-unbiased and design-unbiased consists simply of adopting Strategy 1, in which case the bound in expression (12) is achieved.

8. Estimation of variance

From the previous sections, it clearly appears that the Horvitz–Thompson estimator with a balanced sampling design is a strategy that leads to valid inference under the model and under the sampling design. The estimation of the total should be complemented by a confidenc interval. We will show that it is possible to construct a variance estimator that leads to a valid inference under the model and under the sampling design.

In order to estimate the variance, it is prudent to treat the ν_k as if they were unknown, even if the sample has been selected assuming known ν_k . This will make the estimation of model variance in some sense robust to the failure of that assumption; see, for example, Cumberland & Royall (1981). In the model-assisted framework, Deville & Tillé (2005) have proposed a family of variance estimators for balanced sampling, of the form

$$\hat{\mathrm{var}}(\hat{Y}_\pi) = \sum_{k\in S} c_k \frac{(y_k - x_k'\hat{b})^2}{\pi_k^2},$$

where

$$\hat{b} = \left(\sum_{\ell\in S} c_\ell \frac{x_\ell x_\ell'}{\pi_\ell^2} \right)^{-1} \sum_{\ell\in S} c_\ell \frac{x_\ell y_\ell}{\pi_\ell^2}$$

and the c_k are the solutions of the nonlinear system

$$1 - \pi_k = c_k - \frac{c_k x_k'}{\pi_k} \left(\sum_{\ell\in S} c_\ell \frac{x_\ell x_\ell'}{\pi_\ell^2} \right)^{-1} \frac{c_k x_k}{\pi_k},$$

which can be solved by a fi ed-point algorithm.

In Deville & Tillé (2005), simpler variants of c_k are also proposed, based on the fact that $c_k \doteq n(1-\pi_k)/(n-q)$. The estimator $\hat{\mathrm{var}}(\hat{Y}_\pi)$ is approximately design-unbiased because it is an estimator by substitution (Deville, 1999) of the approximation given in expression (6), which is a reasonable approximation of the variance under the sampling design.

For the model-based framework, the question of estimating $E_\xi(\hat{Y}_\pi - Y)^2$ is complicated because it depends on all the ν_k of the population and not just on the ν_k of the sample. The following result shows that $\hat{\mathrm{var}}(\hat{Y}_\pi)$ is also a pertinent estimator of $E_\xi(\hat{Y}_\pi - Y)^2$ and can be model-unbiased.

Result 8. *Under model* (1)*, if the sample is balanced on* x_k*, then*

$$E_\xi\{\hat{\mathrm{var}}(\hat{Y}_\pi)\} = E_\xi(\hat{Y}_\pi - Y)^2 + \sigma^2 \left(\sum_{k\in S} \frac{\nu_k^2}{\pi_k} - \sum_{k\in U} \nu_k^2 \right),$$

$$E_p E_\xi\{\hat{\mathrm{var}}(\hat{Y}_\pi)\} = E_p E_\xi(\hat{Y}_\pi - Y)^2.$$

If condition (i) *of Definitio* 12 *is met, then* $\hat{\mathrm{var}}(\hat{Y}_\pi)$ *is a model-unbiased estimator of* $E_\xi(\hat{Y}_\pi - Y)^2$.

The proof is given in the Appendix.

If $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal variable, the confidenc interval

$$\left[\hat{Y}_\pi - z_{1-\alpha/2}\sqrt{\{\hat{\text{var}}(\hat{Y}_\pi)\}},\ \hat{Y}_\pi + z_{1-\alpha/2}\sqrt{\{\hat{\text{var}}(\hat{Y}_\pi)\}}\right]$$

leads to a reasonable design-based inference and a valid model-based inference, provided that the v_k^2 can be expressed as linear combinations of the auxiliary variables. This inference does not depend on assumed values of the standard deviations of the errors of the model.

9. Examples

In the examples, we will use the notation

$$\bar{X} = \frac{1}{N}\sum_{k\in U} x_k, \quad \bar{x} = \frac{1}{n}\sum_{k\in S} x_k, \quad \bar{y} = \frac{1}{n}\sum_{k\in S} y_k, \quad \bar{y}_h = \frac{1}{n_h}\sum_{k\in U_h\cap S} y_k,$$

where $U_1, \ldots, U_H$ are strata, i.e. the U_h $(h = 1, \ldots, H)$, are a partition of U. Moreover,

$$s_x^2 = \frac{1}{n-1}\sum_{k\in S}(x_k - \bar{x})^2, \quad s_y^2 = \frac{1}{n-1}\sum_{k\in S}(y_k - \bar{y})^2,$$

$$s_{xy}^2 = \frac{1}{n-1}\sum_{k\in S}(x_k - \bar{x})(y_k - \bar{y}), \quad s_{yh}^2 = \frac{1}{n_h-1}\sum_{k\in U_h\cap S}(y_k - \bar{y}_h)^2.$$

Example 2. Suppose that the superpopulation model is the constant model $y_k = \beta + \varepsilon_k$, for all $k \in U$, with $\text{var}_\xi(\varepsilon_k) = \sigma^2$. This simple model is homoscedastic and has fully explainable heteroscedasticity, which implies that the optimal model-assisted strategy is also an optimal model-based strategy. The optimal model-based strategy consists of selecting any sample of fi ed sample size n, deliberately or randomly. The optimal model-assisted strategy consists of selecting a sample that is balanced on the constant, which implies that it has a fi ed sample size. This sample must be selected with equal inclusion probabilities n/N. In practice, a simple random sampling can be applied and the anticipated mean-squared error is

$$E_p E_\xi(\hat{Y}_\pi - Y)^2 = \sigma^2 N^2 \frac{N-n}{Nn}.$$

In this case, $\hat{Y}_\pi = N\bar{y}$,

$$c_k = \frac{(N-n)n}{N(n-1)}, \quad \hat{\text{var}}(\hat{Y}_\pi) = N^2\frac{N-n}{Nn}s_y^2.$$

Example 3. Suppose that the superpopulation model consists of a constant and only one independent variable, i.e. $y_k = \beta_0 + x_k\beta_1 + \varepsilon_k$, for all $k \in U$, with $\text{var}_\xi(\varepsilon_k) = \sigma^2$. This model is homoscedastic and has fully explainable heteroscedasticity, which implies that the optimal model-assisted strategy is also an optimal model-based strategy. For a particular sample S, balanced or not, and with fi ed sample size, the error variance of the best linear unbiased estimator is

$$E_\xi(\hat{Y}_{\text{BLU}} - Y)^2 = \sigma^2 N^2\frac{N-n}{Nn} + \sigma^2 N^2\frac{(\bar{x}-\bar{X})^2}{(n-1)Ns_x^2}.$$

The optimal model-based strategy consists of selecting a fi ed-sample-size balanced sample in the sense that $\bar{x} = \bar{X}$. The optimal model-assisted strategy consists of selecting a sample that is balanced on x_k, of fi ed sample size and with equal inclusion probabilities. This can be done by using the cube method. Next, one uses the Horvitz–Thompson estimator. The anticipated mean-squared error is then

$$E_p E_\xi(\hat{Y}_\pi - Y)^2 = \sigma^2 N^2 \frac{N-n}{Nn}.$$

By using the approximation $c_k \doteq (N-n)n/\{N(n-2)\}$, we obtain

$$\hat{\text{var}}(\hat{Y}_\pi) = N^2 \frac{N-n}{Nn} \frac{1}{n-2} \sum\nolimits_{k\in S} (y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k)^2,$$

where $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and $\hat{\beta}_1 = s_{xy}/s_x^2$.

Example 4. Suppose that the superpopulation model has only one independent variable, i.e. $y_k = x_k\beta + \varepsilon_k$, for all $k \in U$, with $\text{var}_\xi(\varepsilon_k) = v_k^2\sigma^2$, where $v_k = Nx_k/X$, $x_k \geqslant 0$ and $X = \sum_{k\in U} x_k$. This model does not have fully explainable heteroscedasticity, which implies that the model-assisted and model-based optimal strategies are not the same. The optimal model-based strategy consists of using the best linear unbiased estimator. From expression (10), knowing that $A = X^2 n/N^2$, we obtain the anticipated mean-squared error,

$$E_p E_\xi(\hat{Y}_{\text{BLU}} - Y)^2 = \sigma^2 E_p \left\{ \frac{1}{n} \left(\sum\nolimits_{k\in \bar{S}} v_k \right)^2 + \sum\nolimits_{k\in \bar{S}} v_k^2 \right\}. \tag{13}$$

In this case, the best strictly model-based strategy consists of selecting a nonrandom sample containing the largest n units. However, Valliant et al. (2000, p. 55) point out that, in this case, 'selecting this sample may be risky if the working model is wrong' because it fails to protect against model failure. By using an alternative more general model, they conclude that a balanced sample will protect against model bias resulting from misspecification From a design-based point of view, the strictly best model-based strategy leads to an incorrect design-based inference. The optimal model-assisted strategy consists of using a sampling design that is balanced on x_k and has unequal inclusion probabilities proportional to x_k with the Horvitz–Thompson estimator. The anticipated mean-squared error is then

$$E_p E_\xi(\hat{Y}_\pi - Y)^2 = \sigma^2 \left(\frac{N^2}{n} - \sum\nolimits_{k\in U} v_k^2 \right).$$

This strategy has a larger anticipated mean-squared error than (13), but leads to correct model-assisted and model-based inferences. In this case, the estimator of the variance is

$$\hat{\text{var}}(\hat{Y}_\pi) = \sum\nolimits_{k\in S} \frac{c_k}{\pi_k^2} \left(y_k - \pi_k \frac{\sum_{\ell\in S} c_\ell y_\ell/\pi_\ell}{\sum_{\ell\in S} c_\ell} \right)^2,$$

where c_k are the solutions of the nonlinear system $1 - \pi_k = c_k - c_k^2(\sum_{\ell\in S} c_\ell)^{-1}$ or more simply can be approximated by $c_k \doteq (1-\pi_k)n/(n-1)$.

Example 5. We consider the superpopulation model presented in Kott (1986), given by $y_k = x_k\beta_1 + x_k^2\beta_2 + \varepsilon_k$, for all $k \in U$, with $\text{var}_\xi(\varepsilon_k) = v_k^2\sigma^2$, where $v_k = Nx_k/X$ and $X = \sum_{k\in U} x_k$. This model has fully explainable heteroscedasticity, which implies that the model-assisted and the model-based optimal strategies are the same. Therefore, a strategy that is optimal for both the model-assisted and model-based frameworks consists of selecting a sample balanced on x_k

and x_k^2 with inclusion probabilities that are proportional to x_k, and using the Horvitz–Thompson estimator. The anticipated mean-squared error is then

$$E_p E_\xi(\hat{Y}_\pi - Y)^2 = \sigma^2 \left(\frac{N^2}{n} - \sum\nolimits_{k \in U} v_k^2 \right).$$

This strategy leads to correct model-assisted and model-based inferences.

Example 6. Consider the stratifie superpopulation model $y_{kh} = \alpha_h + \varepsilon_k$, for all $k \in U_h$, $h = 1, \ldots, H$, and suppose that $\mathrm{var}_\xi(\varepsilon_{kh}) = v_h^2 \sigma^2$, with $\sum_{h=1}^H N_h v_h = N$. The stratifie model has fully explainable heteroscedasticity, which implies that the optimal model-assisted strategy is also an optimal model-based strategy. The optimal model-based strategy consists of definin the inclusion probabilities proportional to v_h, which gives $\pi_{kh} = n v_h / N$, which is an optimal stratification Next, a sample is selected with a fi ed sample size $n_h = n N_h v_h / N$ in each stratum U_h. The Horvitz–Thompson estimator, $\hat{Y}_\pi = \sum_{h=1}^H N_h \bar{y}_h$ has anticipated mean-squared error

$$E_p E_\xi(\hat{Y}_\pi - Y)^2 = \sigma^2 \left(\frac{N^2}{n} - \sum_{h=1}^H N_h v_h^2 \right) = \sigma^2 \frac{N^2}{n} \left(1 - \frac{1}{n} \sum_{h=1}^H \frac{n_h^2}{N_h} \right).$$

In this case,

$$c_k = \frac{(N_h - n_h) n_h}{N_h (n_h - 1)}, \quad k \in U_h,$$

and thus

$$\hat{\mathrm{var}}(\hat{Y}_\pi) = \sum_{h=1}^H N_h^2 \frac{N_h - n_h}{N_h n_h} s_{yh}^2.$$

10. Discussion

The search for an optimal strategy rather than an optimal estimator allows the proponents of the model-based and the model-assisted approaches to resolve their differences because, when the superpopulation model has fully explainable heteroscedasticity, one chooses the same sampling design, which is a balanced sampling design with inclusion probabilities that are proportional to the standard deviations of the errors of the model. In this case, the best linear unbiased estimator is the Horvitz–Thompson estimator. As a complement to this estimator, an estimator of the variance can be given, which in turn leads to valid model-based and design-based inferences. The controversy makes sense only if the sample is chosen inappropriately. If the superpopulation model has fully explainable heteroscedasticity, then Strategy 1 is the best strategy in the model-based, model-assisted and combined model-based and model-assisted frameworks, as presented in Table 1.

If the heteroscedasticity is not fully explainable, the optimal strategy is not the same in the model-assisted and model-based frameworks. In fact, Strategy 1 always leads to the selection of a balanced sample, while the strict application of Strategy 2 can lead either to the selection of a balanced sample or to the purposive selection of the sample as in Example 4 in § 9. In this second case, a robustness argument is usually used by the modeller in order to protect against misspecificatio of the model. The robustness is obtained by balancing the sample for the variables that are in the alternative model, which gives the same strategy as in the model-assisted framework. Thus the two approaches are not far apart. In any case, it can also be wise to balance

Table 1. *Optimal strategies in the model-assisted, model-based and combined model-based and model-assisted approaches*

Approach	Fully explainable heteroscedasticity	Non-fully explainable heteroscedasticity
MB	Strategy 1	Strategy 2
MA	Strategy 1	Strategy 1
CMBMA	Strategy 1	Strategy 1

MB, model-based; MA, model-assisted; CMBMA, combined model-based and model-assisted.

the sampling design with respect to additional variables in order to protect against failure of the model, such as the presence of curvature or an intercept. However, we suggest the use of models that have fully explainable heteroscedasticity, which can be easily achieved by systematically using v_k and ${v_k}^2$ as independent variables in the model. This was the advantage of the model developed by Kott (1986) and summarized in Example 5 over the model given in Example 4, which does not have a fully explainable heteroscedasticity.

The theory developed in this paper shows that the best approach is to select a sample that is balanced on the auxiliary variables. If exact balancing is not possible, a nearly balanced sample must firs be selected. In this case, the rounding problem can be solved by a small calibration, by using either the calibration estimator (Deville & Särndal, 1992) or the best linear unbiased estimator, depending on the basis of the inference. An interesting particular case is the so-called cosmetic calibration proposed by Brewer (1999a). In a set of simulations, Deville & Tillé (2004) showed that the balanced sampling design with a calibration estimator strategy achieves the best results among the following four strategies: (i) non-balanced sampling with the Horvitz–Thompson estimator, (ii) balanced sampling with the Horvitz–Thompson estimator,(iii) non-balanced sampling with a calibration estimator and (iv) balanced sampling with a calibration estimator. With strategy (iv), the weights w_{ks} are less random than in the case of strategy (iii), and this leads to a more accurate estimator.

Acknowledgement

The authors would like to thank Alina Matei, Phil Kott and the two reviewers for their helpful comments and suggestions. This work is in part supported by a grant from the Swiss National Science Foundation.

Appendix

Proofs

Proof of Result 1. Since

$$\begin{aligned}\hat{Y}_w - Y &= \sum\nolimits_{k\in S} w_{kS} y_k - \sum\nolimits_{k\in U} y_k \\ &= \sum\nolimits_{k\in S} w_{kS} x_k'\beta + \sum\nolimits_{k\in S} w_{kS}\varepsilon_k - \sum\nolimits_{k\in U} x_k'\beta - \sum\nolimits_{k\in U}\varepsilon_k \\ &= \sum\nolimits_{k\in S}(w_{kS}-1)\varepsilon_k - \sum\nolimits_{k\in \bar{S}}\varepsilon_k + \sum\nolimits_{k\in S} w_{kS} x_k'\beta - \sum\nolimits_{k\in U} x_k'\beta,\end{aligned}$$

we have that

$$E_\xi(\hat{Y}_w - Y)^2 = \sigma^2\left\{\sum\nolimits_{k\in S}(w_{kS}-1)^2 v_k^2 + \sum\nolimits_{k\in \bar{S}} v_k^2\right\} + \left(\sum\nolimits_{k\in S} w_{kS} x_k'\beta - \sum\nolimits_{k\in U} x_k'\beta\right)^2, \qquad \text{(A1)}$$

which leads to the firs equality of Result 1. The second term of (A1) can be simplified Indeed,

$$
\begin{aligned}
&E_p\left(\sum_{k\in S} w_{kS}x_k'\beta - \sum_{k\in U} x_k'\beta\right)^2 \\
&= E_p\left\{\sum_{k\in S} w_{kS}x_k'\beta - E_p\left(\sum_{k\in S} w_{kS}x_k'\beta\right) + E_p\left(\sum_{k\in S} w_{kS}x_k'\beta\right) - \sum_{k\in U} x_k'\beta\right\}^2 \\
&= E_p\left\{\sum_{k\in S} w_{kS}x_k'\beta - E_p\left(\sum_{k\in S} w_{kS}x_k'\beta\right)\right\}^2 + E_p\left\{\sum_{k\in U} E_p(w_{kS}I_k)x_k'\beta - \sum_{k\in U} x_k'\beta\right\}^2 \\
&\quad + 2E_p\left[\left\{\sum_{k\in S} w_{kS}x_k'\beta - E_p\left(\sum_{k\in S} w_{kS}x_k'\beta\right)\right\}\left\{\sum_{k\in U} E_p(w_{kS}I_k)x_k'\beta - \sum_{k\in U} x_k'\beta\right\}\right] \\
&= \mathrm{var}_p\left(\sum_{k\in S} w_{kS}x_k'\beta\right) + \left(\sum_{k\in U} C_k x_k'\beta - \sum_{k\in U} x_k'\beta\right)^2. \qquad \text{(A2)}
\end{aligned}
$$

The firs term of (A1) g i v e s

$$
\begin{aligned}
&\sigma^2 E_p\left\{\sum_{k\in S}(w_{kS}-1)^2\nu_k^2 + \sum_{k\in \bar{S}}\nu_k^2\right\} \\
&= \sigma^2\left[\sum_{k\in U} E_p\left\{(w_{kS}-1)^2 I_k\right\}\nu_k^2 + \sum_{k\in U}(1-\pi_k)\nu_k^2\right] \\
&= \sigma^2\sum_{k\in U}\nu_k^2\left[E_p\left\{(w_{kS}-1)^2 I_k\right\} + 1 - \pi_k\right] \\
&= \sigma^2\sum_{k\in U}\nu_k^2\left\{E_p\left(w_{kS}^2 I_k\right) - 2E_p(w_{kS}I_k) + \pi_k + 1 - \pi_k\right\} \\
&= \sigma^2\sum_{k\in U}\nu_k^2\left\{E_p\left(w_{kS}^2 I_k\right) - E_p^2(w_{kS}I_k) + E_p^2(w_{kS}I_k) - 2E_p(w_{kS}I_k) + 1\right\} \\
&= \sigma^2\sum_{k\in U}\nu_k^2\left\{\mathrm{var}_p(w_{kS}I_k) + (C_k-1)^2\right\}. \qquad \text{(A3)}
\end{aligned}
$$

By the law of total variance,

$$
\begin{aligned}
\mathrm{var}_p(w_{kS}I_k) &= \mathrm{var}_p E_p(w_{kS}I_k \mid I_k) + E_p\mathrm{var}_p(w_{kS}I_k \mid I_k) \\
&= \pi_k\{E_p(w_{kS} \mid I_k = 1)\}^2 - \{E_p(w_{kS}I_k)\}^2 + \pi_k\mathrm{var}_p(w_{kS} \mid I_k = 1) \\
&= \frac{1-\pi_k}{\pi_k}C_k^2 + \pi_k\mathrm{var}_p(w_{kS} \mid I_k = 1). \qquad \text{(A4)}
\end{aligned}
$$

By inserting (A4) i n t o (A3), and by adding (A2) and (A3), we finall obtain the second equality of Result 1. □

Proof of Result 2. Result 2 comes directly from equation (2). Term B vanishes because the weights $1/\pi_k$ do not differ from sample to sample. Term C vanishes because the estimator is design-unbiased. Terms D and E vanish because the estimator is model-unbiased under balanced sampling. All that remains is term A with $C_k = 1$ because the estimator is design-unbiased. □

Proof of Result 3. Since $y_k = x_k'\beta + \varepsilon_k$,

$$
\begin{aligned}
\mathrm{var}_{\mathrm{app}}(\hat{Y}_\pi) &= \sum_{k\in U} d_k\frac{(y_k - x_k'b)^2}{\pi_k^2} \\
&= \sum_{k\in U} d_k\left\{\frac{\varepsilon_k}{\pi_k} - \frac{x_k}{\pi_k}'\left(\sum_{\ell\in U} d_\ell\frac{x_\ell x_\ell'}{\pi_\ell^2}\right)^{-1}\sum_{\ell\in U} d_\ell\frac{x_\ell\varepsilon_\ell}{\pi_\ell^2}\right\}^2 \\
&= \sum_{k\in U} d_k\frac{\varepsilon_k^2}{\pi_k^2} - \sum_{k\in U}\frac{d_k x_k'\varepsilon_k}{\pi_k^2}\left(\sum_{\ell\in U}\frac{d_\ell x_\ell x_\ell'}{\pi_\ell^2}\right)^{-1}\sum_{\ell\in U}\frac{d_\ell x_\ell\varepsilon_\ell}{\pi_\ell^2}.
\end{aligned}
$$

Thus,

$$E_\xi\{\mathrm{var}_{\mathrm{app}}(\hat{Y}_\pi)\} = \sigma^2 \sum_{k\in U} d_k \frac{v_k^2}{\pi_k^2} - \sigma^2 \sum_{k\in U} \frac{v_k^2}{\pi_k^2} \frac{d_k x_k'}{\pi_k} \left(\sum_{\ell\in U} d_\ell \frac{x_\ell x_\ell'}{\pi_k^2} \right)^{-1} \frac{d_k x_k}{\pi_k}.$$

By using the definitio of d_k, given in expression (7), we obtain

$$E_\xi\{\mathrm{var}_{\mathrm{app}}(\hat{Y}_\pi)\} = \sigma^2 \sum_{k\in U} \pi_k(1-\pi_k)\frac{v_k^2}{\pi_k^2} = E_p E_\xi(\hat{Y}_\pi - Y)^2,$$

which holds even when the v_k are unknown. □

Proof of Result 4. The optimal inclusion probabilities π_k^* are obtained by minimizing (8) subject to

$$\sum_{k\in U} \pi_k = n, \quad 0 \leqslant \pi_k \leqslant 1,$$

which gives the second inequality. Now, if we minimize (8) subject to $\sum_{k\in U} \pi_k = n$, but without the constraint $\pi_k \leqslant 1$, then we obtain $\tilde{\pi}_k = n v_k / N$, and we obtain a still lower bound in the third inequality. □

Proof of Result 8. By Result 3, following the same steps, we obtain

$$\begin{aligned} E_\xi\{\hat{\mathrm{var}}(\hat{Y}_\pi)\} &= \sigma^2 \sum_{k\in S}(1-\pi_k)\frac{v_k^2}{\pi_k^2} \\ &= \sigma^2 \left\{ \sum_{k\in S}(1-\pi_k)^2 \frac{v_k^2}{\pi_k^2} + \sum_{k\in \bar{S}} v_k^2 \right\} + \sigma^2 \left(\sum_{k\in S} \frac{v_k^2}{\pi_k} - \sum_{k\in U} v_k^2 \right) \\ &= E_\xi(\hat{Y}_\pi - Y)^2 + \sigma^2 \left(\sum_{k\in S} \frac{v_k^2}{\pi_k} - \sum_{k\in U} v_k^2 \right). \end{aligned}$$

Obviously, if there exists a vector λ such that $\lambda' x_k = v_k^2$, then

$$\sum_{k\in S} \frac{v_k^2}{\pi_k} - \sum_{k\in U} v_k^2 = 0.$$

□

References

Ardilly, P. (1991). Échantillonnage représentatif optimum à probabilités inégales. *Ann. D´Econ. Statist.* **23**, 91–113.

Brewer, K. R. W. (1994). Survey sampling inference: Some past perspectives and present prospects. *Pak. J. Statist.* **10**, 15–30.

Brewer, K. R. W. (1999a). Cosmetic calibration for unequal probability sample. *Survey Methodol.* **25**, 205–12.

Brewer, K. R. W. (1999b). Design-based or prediction-based inference? Stratifie random vs stratifie balanced sampling. *Int. Statist. Rev.* **67**, 35–47.

Brewer, K. R. W. (2002). *Combined Survey Sampling Inference, Weighing Basu's Elephants*. London: Arnold.

Brewer, K. R. W., Hanif, M. & Tam, S. M. (1988). How nearly can model-based prediction and design-based estimation be reconciled. *J. Am. Statist. Assoc.* **83**, 128–32.

Chambers, R. L. (1996). Robust case-weighting for multipurpose establishment surveys. *J. Offic Statist.* **12**, 3–32.

Cumberland, W. G. & Royall, R. M. (1981). Prediction models in unequal probability sampling. *J. R. Statist. Soc.* B **43**, 353–67.

Deville, J.-C. (1992). Constrained samples, conditional inference, weighting: Three aspects of the utilisation of auxiliary information. In *Proc. Workshop on the Uses of Auxiliary Information in Surveys*, pp. 21–40. Örebro, Sweden: Statistics Sweden.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodol.* **25**, 193–204.

Deville, J.-C., Grosbras, J.-M. & Roth, N. (1988). Efficien sampling algorithms and balanced samples. In *COMPSTAT, Proceedings in Computational Statistics*, Ed. R. Payne and P. Green, pp. 255–66, Heidelberg: Physica.

Deville, J.-C. & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *J. Am. Statist. Assoc.* **87**, 376–82.

Deville, J.-C. & Tillé, Y. (2004). Efficien balanced sampling: The cube method. *Biometrika* **91**, 893–912.

DEVILLE, J.-C. & TILLÉ, Y. (2005). Variance approximation under balanced sampling. *J. Statist. Plan. Infer.* **128**, 411–25.

GODAMBE, V. P. (1955). A unifie theory of sampling from finit population. *J. R. Statist. Soc.* B **17**, 269–78.

GODAMBE, V. P. & JOSHI, V. M. (1965). Admissibility and Bayes estimation in sampling finit populations I. *Ann. Math. Statist.* **36**, 1707–22.

HÁJEK, J. (1981). *Sampling from a Finite Population*. New York: Marcel Dekker.

HANSEN, M. H., MADOW, W. G. & TEPPING, B. J. (1983). An evaluation of model dependent and probability-sampling inferences in sample surveys (with Discussion). *J. Am. Statist. Assoc.* **78**, 776–807.

HEDAYAT, A. S. & MAJUMDAR, D. (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *J. Statist. Plan. Infer.* **44**, 237–47.

IACHAN, R. (1984). Sampling strategies, robustness and efficien y: The state of the art. *Int. Statist. Rev.* **52**, 209–18.

ISAKI, C. T. & FULLER, W. A. (1982). Survey design under a regression population model. *J. Am. Statist. Assoc.* **77**, 89–96.

KOTT, P. S. (1986). When a mean-of-ratios is the best linear unbiased estimator under a model. *Am. Statist.* **40**, 202–4.

ROYALL, R. M. (1976). The linear least squares prediction approach to two-stage sampling. *J. Am. Statist. Assoc.* **71**, 657–64.

ROYALL, R. M. (1988). The prediction approach to sampling theory. In *Handbook of Statistics*, Vol. 6. pp. 399–413. Amsterdam, Holland: Elsevier Science Publishers.

ROYALL, R. M. (1992). Robustness and optimal design under prediction models for finit populations. *Survey Methodol.* **18**, 179–85.

ROYALL, R. M. & CUMBERLAND, W. G. (1981). The finit population linear regression estimator and estimators of its variance. An empirical study. *J. Am. Statist. Assoc.* **76**, 924–30.

ROYALL, R. M. & HERSON, J. (1973a). Robust estimation in finit populations I. *J. Am. Statist. Assoc.* **68**, 880–9.

ROYALL, R. M. & HERSON, J. (1973b). Robust estimation in finit populations II: Stratificatio on a size variable. *J. Am. Statist. Assoc.* **68**, 891–3.

SÄRNDAL, C.-E., SWENSSON, B. & WRETMAN, J. H. (1992). *Model Assisted Survey Sampling*. New York: Spinger.

SCOTT, A. J., BREWER, K. R. W. & HO, E. W. H. (1978). Finite population sampling and robust estimation. *J. Am. Statist. Assoc.* **73**, 359–61.

SMITH, T. M. F. (1976). The foundations of survey sampling: A review. *J. R. Statist. Soc.* A **139**, 183–204.

SMITH, T. M. F. (1984). Sample surveys, present position and potential developments: Some personal views (with Discussion). *J. R. Statist. Soc.* A **147**, 208–21.

SMITH, T. M. F. (1994). Sample surveys 1975–1990; An age of reconciliation (with Discussion)? *Int. Statist. Rev.* **62**, 5–34.

THIONET, P. (1953). *La théorie des sondages*. Paris: INSEE, Imprimerie Nationale.

TILLÉ, Y. (2006). *Sampling Algorithms*. New York: Springer.

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

YATES, F. (1946). A review of recent statistical developments in sampling and sampling surveys (with Discussion). *J. R. Statist. Soc.* A **109**, 12–43.