Design-based predictive inference

Li-Chun Zhang[1,2], Luis Sanguiao-Sande[3], and Danhyang Lee[4]

[1]*Statistisk sentralbyrå (lcz@ssb.no)*
[2]*University of Southampton (L.Zhang@soton.ac.uk)*
[3]*Instituto Nacional de Estadística (luis.sanguiao.sande@ine.es)*
[4]*Baylor University (Danhyang_Lee@baylor.edu)*

Abstract

Design-based inference from probability samples is valid by construction for target parameters that are descriptive summaries of finite populations. We develop a novel approach of design-based predictive inference for finite populations, where the individual-level predictor is learned from a probability sample using any models or algorithms for incorporating the relevant auxiliary information, and the uncertainty of estimation is evaluated with respect to the known probability design while the outcome and auxiliary values for modelling are treated as constants. Unlike the existing theory of design-based model-assisted estimation for finite populations, design-based predictive inference is as well suited for individual-level prediction *in addition* to producing population-level estimates.

Keywords: Probability sampling, model-assisted estimation, sample split, Rao-Blackwellisation, administrative register, big data

1 Introduction

Throughout the 20th century, design-based inference from finite-population probability sampling has been established as the standard approach to official statistics; see Hansen (1987), Smith (1994), Kalton (2002), Rao (2005, 2011), Beaumont and Haziza (2022) for reviews and appraisals. In this context the target parameters for estimation are descriptive, observable summaries of a given finite population, such as the population total, mean or quantiles of some specific values associated with the given population units, and the inference is characterised as descriptive or predictive (Smith, 1983; Geisser, 1993), in contrast to analytic inference of theoretical, unobservable targets such as the life expectancy (of a hypothetical cohort of individuals) or a parametric model that can be used to understand the given population.

By design-based inference from probability samples, the uncertainty of estimation is evaluated with respect to hypothetically repeated sampling from the same finite population, while all the other values involved are treated as constants associated with the given population. Design-based inference is valid by construction because it is based on the *known* sampling design, "whatever the unknown properties of the population" (Neyman, 1934). In contrast, by model-based inference, the uncertainty of estimation is evaluated with respect to an *assumed* statistical model of the observations, while the available sample is typically treated as fixed; see e.g. Valliant et al. (2000). Although models are necessary for analytic targets or if the available observations are not obtained by probability sampling, model-based inference may be invalid to the extent the assumed model is misspecified in respects that matter to the task at hand.

Design-based inference can be made more efficient by using auxiliary information in addition to the sampling design. This can largely recover the 'loss of efficiency' compared to model-based inference that uses the same auxiliary information optimally under the assumed model. For instance, calibration estimation (Deville and Särndal, 1992) is a general approach that makes adjustments to the design weights with respect to the known auxiliary population totals. Or, empirical likelihood methods can yield confidence intervals that have better properties than normal approximation based on the central limit theorem (Hartley and Rao, 1968; Rao and Wu, 2010; Berger and De La Riva Torres, 2016). More relevant to our development is the model-assisted approach where an assisting model is explicitly formulated but inference remains design-based, whether or not the adopted estimator has optimal properties with respect to the assisting model. One can use linear models (Särndal et al, 1992), generalised linear models (Wu and Sitter, 2001), or many other models under a unified "construction recipe" as reviewed by Breidt and Opsomer (2017).

To justify any model-assisted estimator that is not design-unbiased, it is common to seek a proof that it can be design consistent asymptotically for a hypothetical sequence of populations of increasing sizes. As Smith (1994) points out, this "asymptotic notion of consistency" is not immediately applicable to the given population as "a real entity". In contrast, for a given population and sampling method, if $t(1), ..., t(k)$ are unbiased estimators of the population totals $T(1), ..., T(k)$, then $g(t(1), ..., t(k))$ is called "consistent" for $g(T(1), ..., T(k))$ by Fisher (1956), in that replacing $t(j)$ by $T(j)$ gives the true target population parameter. Similarly, an interval estimator of "a collective character... of a population" is called "consistent" by Neyman (1934), if it achieves the designated level of coverage given the finite population and sampling method.

We emphasise that asymptotic consistency of a point estimator or an interval estimator is unnecessary, if the estimator is *Neyman-Fisher consistent* in the sense Neyman and Fisher have used the term "consistent" for finite-population inference. This will be our perspective to *design-based predictive inference* in this paper, which may also be called *fully design-based* inference in contrast

to *asymptotically design-consistent* inference (that has been traditionally more common for model-assisted finite population estimation).

Now, we notice that design-unbiased ratio or linear regression estimators *for population totals* have been proposed by Hartley and Ross (1954) and Mickey (1959), which are finite-population Neyman-Fisher consistent. More recently, Sanguiao-Sande and Zhang (2021) developed a design-unbiased approach, called *subsampling Rao-Blackwellisation (SRB)*, which allows for *any* assisting Machine Learning (ML) models or algorithms that have become increasingly common. The SRB approach combines three classic ideas in statistical inference, (i) model-assisted estimation for survey sampling, (ii) cross-validation for error estimation by ML methods, and (iii) the Rao-Blackwell Theorem (Rao, 1945; Blackwell, 1947) for efficiency improvement.

In this paper, we extend the SRB approach to a larger class of population estimators, which are commonly referred to as the prediction estimators, *as well as* the associated individual-level predictors for the out-of-sample units. Notice that, traditionally, due to the lack of a design-based prediction theory, individual outcomes must be treated as random variables for model-based prediction and the term predictor is common in this context. Although from a design-based inference perspective the term prediction estimator would seem more appropriate also at the unit level, we shall keep the term predictor at the individual level for convenience and familiarity reasons. Notice also that the two terms "unit" and "individual" are used interchangeably in this paper, such as in 'statistical unit' or 'individual prediction'.

It may be helpful to make some remarks immediately regarding the nature of our inference approach and its advantages compared to the more familiar model-assisted inference.

First, we consider predictive inference by definition, where the sample-based prediction estimator (using any given ML model or algorithm) aims at some out-of-sample quantity that *varies with the sample*, the property of which is evaluated *only* with respect to repeated sampling from the given population. This design-based predictive inference outlook differs from model-assisted estimation that is aimed at *fixed* population parameters (such as totals or means).

Next, we develop Neyman-Fisher consistent uncertainty estimators, which accommodate any given assisting model (or algorithm) and apply generally to sampling from finite populations. In contrast, asymptotically design-consistent inference may not hold in a given setting of finite population sampling, but still requires tailored asymptotic arguments to be developed for different nonparametric assisting models, such as random forest or support vector machine, which will remain a challenge as new models or algorithms emerge.

Finally, while our approach is model-assisted in the sense that the sampling design remains the inference basis despite the model introduced, it provides as well a design-theoretical basis for individual level prediction (estimation). This is another important difference to standard model-assisted estimation that is

only applicable to population parameter estimation.

1.1 Prediction estimator

Denote by $U = \{1, ..., N\}$ a given finite population that is of size N. Let $y_U = \{y_i : i \in U\}$ be the associated values of interest. Denote by $x_U = \{x_i : i \in U\}$ the collection of feature vectors, where x_i is the vector associated with each unit $i \in U$. Given any sample of units from U, denoted by $s \subset U$, let $\mu(x, s)$ be a predictor for any out-of-sample unit, say, $j \in R = U \setminus s$, whose feature vector takes value x, i.e. $x_j = x$. Notice that any function of $\{y_i : i \in U \setminus s\}$ is a random quantity that varies with the sample s, just like $\mu(x, s)$ itself as the notation emphasises. When the target parameter is the population total $Y = \sum_{i \in U} y_i$, the prediction estimator of Y is given as

$$\hat{Y} = \sum_{i \in s} y_i + \sum_{j \in U \setminus s} \mu(x_j, s) \tag{1}$$

We shall consider $\mu(x, s)$ as the associated individual-level predictor of y for any unit with given features x, where y is treated as a constant just like x.

Notice that individual level features are required to compute the prediction estimator (1), except e.g. in the case of using linear models. However, this is a requirement common to all individual-level prediction models, regardless the inference framework. There may be situations where such information is unavailable, which would limit one's choice of models. But the ability to utilise individual level covariates is certainly not a limitation of our approach not least because, as we shall explain, design-based individual-level predictive inference can provide a valid theoretical basis for producing census-like statistical data, which fills a gap in the current literature.

Note that many design-based estimators in survey sampling can as well be given as prediction estimators (1). For example, let $x_i = \pi_i N/n$ for any $i \in U$, given $\pi_i = \Pr(i \in s)$ and $n = |s|$, such that the Horovitz-Thompson (HT) estimator (Horvitz and Thompson, 1952) can be given in the form (1) with

$$\mu_{HT}(x_j, s) = x_j \beta_s + \frac{1}{N - n} \sum_{i \in s} (x_i \beta_s - y_i)$$

where $\beta_s = n^{-1} \sum_{i \in s} y_i / x_i$. Other examples include the generalised regression estimator, model-calibrated linear estimator, as well as the SRB estimator of Sanguiao-Sande and Zhang (2021).

However, we are interested in design-unbiased inference of the prediction estimator (1) generally, including when μ is given by an arbitrary ML method *without regard* to the sampling design and may not have a wieldy expression, such as a random forest trained on $\{(y_i, x_i) : i \in s\}$ by ready-made software.

The theory of design-based predictive inference for population totals and individuals will be developed and illustrated in Sections 2 and 3, respectively,

an illustrative application to Structural Business Survey is given in Section 4, and some final remarks are given in Section 5. However, before we get into the details of the development, let us first motivate below what design-based predictive inference can do for official statistics.

1.2 Introduction to total estimation

Design-based predictive inference is clearly relevant to the perennial design vs. model controversy in survey sampling, as traditionally the two main strands of approaches to finite population estimation.

In the design-based approach, estimators depend on the sampling design through the sample inclusion probabilities or other known sampling probabilities. Although auxiliary information in addition to the sampling design can be incorporated by various techniques, the validity and the associated uncertainty of the resulting estimator are still based on the given sampling design. In contrast, the model-based prediction approach, frequentist or Bayesian, depends on an assumed working model, which typically ignores the sampling design and treats the available sample as fixed when it comes to the assessment of the associated uncertainty.

Meanwhile, it is possible to evaluate any design-based estimator, such as the HT estimator or a generalised regression estimator, with respect to an assumed model, in which case it is common to conclude that design-based estimation is inefficient or lacks desirable conditional properties (e.g. Valliant et al., 2000). Conversely, a prediction estimator derived optimally under a working model can be evaluated with respect to the sampling design, in which case the danger of model misspecification is frequently noted (e.g. Hansen et al., 1983).

Design-based predictive inference takes the last analysis further, whereby one explicitly evaluates the design-based bias and variance of any given model-based prediction estimator (1). One can then compare the given model-based estimator to any other estimator, whether the latter is dependent on the design or a working model, and choose according to their design-based properties regardless how they are constructed. The merit of such an approach rests now on the fact that design-based uncertainty assessment is valid, which does not require the model underlying any given estimator to be correct.

To illustrate, under the linear model $E_M(y_i \mid x_i) = x^\top \beta$ and $V_M(y_i \mid x_i) = \sigma^2$ where E_M and V_M denote expectation and variance under the model, the best linear unbiased predictor (BLUP) of a population total $Y = \sum_{i \in U} y_i$ is

$$\hat{Y} = X^\top b \quad \text{and} \quad b = (\sum_{i \in s} x_i x_i^\top)^{-1} \sum_{i \in s} x_i y_i$$

where $X = \sum_{i \in U} x_i$. Given $\pi_i = \Pr(i \in s)$, we have

$$E_p(\hat{Y}) = X^\top E_p(b) = Y - \sum_{i \in U} \{y_i - x_i^\top E_p(b)\}$$

where $E_p(b) \approx (\sum_{i\in U} \pi_i x_i x_i^\top)^{-1} \sum_{i\in U} \pi_i x_i y_i$ with respect to sampling, and

$$V_p(\hat{Y}) = X^\top V_p(b \mid s) X \approx X^\top E_p\{V_M(b \mid s)\} X$$

with respect to sampling, where

$$E_p\{V_M(b \mid s)\} = E_p\{\sigma^2 (\sum_{i\in s} x_i x_i')^{-1}\} \approx \Big(\frac{1}{N} \sum_{i\in U} \{y_i - x_i^\top E_p(b)\}^2\Big) \big(\sum_{i\in U} \pi_i x_i x_i^\top\big)^{-1} .$$

Both $E_p(\hat{Y})$ and $V_p(\hat{Y})$ can then be estimated from s and compared to, say, the approximate variance of a generalised regression estimator of Y.

Thus, a chief advantage of adopting design-based predictive inference to population total estimation based on probability sampling is to circumvent the design vs. model controversy, by providing a valid common ground for uncertainty assessment. A theory applicable to the class of prediction estimators (1), which will be developed in this paper, would allow one to use any assisting ML models or algorithms that can often make more efficient use of auxiliary information than the standard design-based calibration estimation or model-assisted estimation methods.

1.3 Introduction to individual estimation

Individual estimation requires the most extremely disaggregated results. It can be useful for constructing statistical registers or census-like statistical data as the basis for descriptive official statistics. However, there has never been a design-based theory for estimation *at the individual level*.

For instance, having taken a simple random sample of *all but two* units in a given population, the traditional design-based estimation theory would only allow one to make inference about the total (or mean) of the two out-of-sample units, but not each on its own, no matter how large the sample is or how much auxiliary information one has in addition. This is clearly unsatisfactory, which requires extension of the design-based inference theory.

To illustrate the conceptual issue at hand, suppose on observing $\{y_i : i \in s\}$ in a subset s of the population U, one would like to predict the y-value for each unit out of s by $\mu(s) = \sum_{i\in s} y_i/n$, where n is the number of units in s. How can one infer about the loss $D_s = \sum_{j\in U\setminus s} \{\mu(s) - y_j\}^2$ that is unobserved?

One possibility is to assume a model. For instance, under the model that y_i is independent and identically distributed (IID) for any $i \in U$, we have

$$E_M(D_s \mid s) = (N-n)\big(1 + n^{-1}\big)\sigma^2$$

with respect to the IID model conditional on the given subset s, where σ^2 is the variance of y_i under the model and N is the number of units in U.

However, we notice that a fundamentally different, design-based approach would in fact be possible if s is selected from U by a known sampling design,

denoted by $s \sim p(s)$, where $\sum_{s\in\Omega} p(s) = 1$ and Ω contains all the possible samples from U. For instance, suppose s is selected from U by simple random sampling without replacement (SRSWOR), where $p(s) = 1/\binom{N}{n}$, such that

$$E_p(D_s) = (N-n)\big(1+n^{-1}\big)S_y^2$$

with respect to $p(s)$, where $S_y^2 = \sum_{i\in U}(y_i - \bar{Y})^2/(N-1)$ and $\bar{Y} = \sum_{i\in U} y_i/N$.

Since $s_y^2 = \sum_{i\in s}\{y_i - \mu(s)\}^2/(n-1)$ is both unbiased for σ^2 under the IID model and unbiased for S_y^2 under SRSWOR, numerically one would obtain the same estimate of the expected loss, even though they have completely different interpretations. While an assumed model would be necessary for evaluating $E_M(D_s|s)$ if the selection mechanism of s is unknown, it could be invalid if the observed data distribution actually differs to that of the unobserved ones. While the design-based loss $E_p(D_s)$ here requires one to plan and implement the SRSWOR design, it is necessarily valid because $p(s)$ is known.

We shall develop a general design-based theory for the out-of-sample loss, provided the observations are obtained by probability sampling. This would yield valid inference of the associated risk with respect to the given sampling design, where all the outcomes y_U and features x_U are treated as constants.

2 Total prediction estimator

Consider the prediction estimator (1), where $\mu(x, s)$ can be obtained by any model or algorithm fitted to the full sample s. Now, hypothetically speaking, it is clear that design-unbiased estimation of $\hat{Y} - Y = \sum_{j\in R}\mu(x_j, s) - y_j$, or some function of it, would be possible given an *additional* probability sample r selected from $U \setminus s$, because one can then observe the error $e_j = \mu(x_j, s) - y_j$ for any $j \in r$. In the absence of extra observations $\{y_j; j \in r\}$, valid design-based inference requires creating observed errors within the sample s.

Denote by $s_1 \cup s_2 = s$ and $s_1 \cap s_2 = \emptyset$ a *training-test sample split*, where s_1 is selected by a *subsampling design*, denoted by $q(s_1 \mid s)$. For example, s_1 of size n_1 can be randomly sampled from s with or without replacement. Or, in T-fold cross-validation, s is first randomly partitioned into T clusters and then each cluster is selected as s_2 one by one systematically, yielding $s_1 = s\setminus s_2$ accordingly. Denote by $\mu(x, s_1)$ the predictor obtained from the subsample s_1, in the same way as $\mu(x, s)$ from s. Its error $\mu(x_j, s_1) - y_j$ can be observed for any $j \in s_2$.

As in Sanguiao-Sande and Zhang (2021), we shall refer to the sampling design that yields (s_1, s) as the pq-*design*, denoted by

$$f(s_1, s) = q(s_1 \mid s)p(s) = f(s \mid s_1)f(s_1) \tag{2}$$

where the last product indicates that, conditional on the training set s_1, one can view the test set s_2 as a probability sample from $U \setminus s_1$, according to which

s can vary under the pq-design. In particular, for any $i \in U$, let

$$\pi_{2i} = \Pr(i \in s_2 \mid s_1) = \sum_{s \ni i, i \notin s_1} f(s \mid s_1) \tag{3}$$

be its conditional s_2-inclusion probability given s_1 under the pq-design.

The theory of design-based predictive inference we develop below, including both Theorems 1 and 2, apply generally provided any well-defined pq-design (2). However, the typical q-designs mentioned above may need to be modified in practice, in order to accommodate the additional complex sampling features that may exist. For instance, given a stratified p-design of s, it may be natural to subsample s_1 within each stratum as well. Or, given a multistage p-design, the q-design must involve subsampling of the selected primary sampling units (PSUs), instead of only subsampling elements within all the selected PSUs, such that conditional sampling by $f(s \mid s_1)$ covers the whole population.

Now, conditional on any given s_1, the design-based bias and mean squared error (MSE) of the prediction estimator using $\mu(x, s_1)$ can be easily derived. The theory below explains how one can infer the design-based bias and MSE of the prediction estimator (1) that is trained using all the observations in s, with respect to $s \sim p(s)$, by appropriate averaging over the q-design.

2.1 SRB prediction estimator

Given s_1 under the pq-design, the *subsample-trained* prediction estimator is

$$\hat{Y}_1^* = \sum_{i \in s} y_i + \sum_{j \in U \setminus s} \mu(x_j, s_1)$$

whose *total error* for Y is given by

$$B = \sum_{j \in U \setminus s} e_{1j} = B_1 - B(s_2)$$

where $e_{1j} = \mu(x_j, s_1) - y_j$ for any $j \notin s_1$, and

$$B_1 = \sum_{j \in U \setminus s_1} e_{1j} \qquad \text{and} \qquad B(s_2) = \sum_{j \in s_2} e_{1j}$$

Note that B and $B(s_2)$ vary with (s, s_2) conditional on s_1 while B_1 is fixed.

Applying Rao-Blackwellisation to $\hat{Y}_1^*$ yields a corresponding *SRB prediction estimator* in the form of (1), which is given by

$$\hat{Y}^{RB} = \sum_{i \in s} y_i + \sum_{j \in U \setminus s} \bar{\mu}(x_j, s) \tag{4}$$

where

$$\bar{\mu}(x_j, s) = E_q\{\mu(x_j, s_1) \mid s\} \tag{5}$$

Notice that $\bar{\mu}(x, s)$ is a particular full-sample trained predictor, and its special notation $\bar{\mu}$ is introduced to distinguish it from any $\mu(x, s)$ that uses the same μ but is directly trained *once* on the full sample s. Sanguiao-Sande and Zhang (2021) refer to the operation $E_q(\cdot \mid s)$ as SRB, since it yields the conditional expectation over subsampling $s_1 \sim q(s_1 \mid s)$, where the unordered s is the minimal sufficient statistic with respect to the sampling distribution $p(s)$.

It follows from the definition that the bias of $\hat{Y}^{RB}$ is given by

$$E_p(\hat{Y}^{RB}) - Y = E_{pq}(\hat{Y}_1^*) - Y = E_{pq}(B) \ .$$

Given s_1, a conditionally unbiased predictor of B is given by

$$\hat{B} = \sum_{j \in s_2} (\pi_{2j}^{-1} \ - 1) e_{1j}$$

in the sense that, as s varies conditional on s_1, we have

$$E_s(\hat{B} \mid s_1) = B_1 - E_s\{B(s_2) \mid s_1\} = E_s(B \mid s_1)$$

since e_{1j} resulting from $\mu(x_j, s_1)$ trained on the subsample s_1 is fixed with respect to sampling of s_2 by $f(s \mid s_1)$ conditional on s_1. Applying Rao-Blackwellisation to $\hat{B}$ yields then a more efficient estimator

$$\hat{B}^{RB} = E_q(\hat{B} \mid s) \tag{6}$$

i.e. as an unbiased estimator of the bias of $\hat{Y}^{RB}$ with respect to $p(s)$, since

$$E_p(\hat{B}^{RB}) = E_p\{E_q(\hat{B} \mid s)\} = E_{s_1}\{E_s(\hat{B} \mid s_1)\} = E_{s_1}\{E_s(B \mid s_1)\} = E_{pq}(B) \ .$$

Notice that one needs at least $|s_2| = 1$ to calculate $\hat{B}$, in which case $\mu(x, s_1)$ is trained on $n - 1$ units, i.e. the so-called leave-one-out (LOO) predictor, whose difference to $\mu(x, s)$ is only due to one randomly selected unit.

Moreover, one can estimate unbiasedly the MSE of $\hat{Y}^{RB}$ by the following result, the proof of which is given in Appendix A.

Theorem 1. *For any given $\mu(\cdot)$, an unbiased estimator of the MSE of the SRB prediction estimator* (4)*, over $s \sim p(s)$, is given by*

$$\mathit{mse}^{RB} = E_q\{\hat{B}^2 - \hat{V}_s(\hat{B} \mid s_1) + \hat{V}_s\{B(s_2) \mid s_1\} \mid s\} - V_q(\hat{Y}_1^* \mid s) \tag{7}$$

where $\hat{B} = \sum_{j \in s_2} (\pi_{2j}^{-1} \ - 1)\{\mu(x_j, s_1) - y_j\}$, and $\hat{V}_s(\hat{B} \mid s_1)$ is unbiased for

$$V_s(\hat{B} \mid s_1) = \sum_{i \notin s_1} \sum_{j \notin s_1} (\pi_{2ij} - \pi_{2i}\pi_{2j})(\frac{1}{\pi_{2i}} - 1)(\frac{1}{\pi_{2j}} - 1) e_{1i} e_{1j}$$

where $\pi_{2ij} = \Pr(i, j \in s_2 \mid s_1)$*, and* $\hat{V}_s\{B(s_2) \mid s_1\}$ *is unbiased for*

$$V_s\{B(s_2) \mid s_1\} = \sum_{i \notin s_1} \sum_{j \notin s_1} (\pi_{2ij} - \pi_{2i}\pi_{2j}) e_{1i} e_{1j} \ .$$

Notice that one needs at least $|s_2| = 2$ to calculate the variance estimators in (7), in which case $\mu(x, s_1)$ is trained on $n - 2$ units, i.e. the leave-two-out (LTO) predictor, which differs to $\mu(x, s)$ only due to two randomly selected units.

2.2 Discussion

In a recent discussion of cross-validation for prediction error estimation under the IID model of (y_i, x_i), Bates et al. (2023, Theorem 3) have shown that, given a sample s of size n, it is possible to estimate unbiasedly the MSE of $(K-1)$-fold cross-validation on a reduced sample of size $n(K-1)/K$, but not for the K-fold cross-validation on the sample of size n which one actually does. Unbiased MSE estimation for the latter is generally difficult if $\mu(x, s)$ does not have a wieldy expression, because by definition one cannot observe the error of $\mu(x, s)$ without extra out-of-sample observations.

We have obtained design-unbiased MSE estimator (7) for the SRB prediction estimator $\hat{Y}^{RB}$ that uses $\bar{\mu}(x, s)$, but not for $\hat{Y}$ using the more familiar $\mu(x, s)$ that is trained once on the full sample. Design-unbiased estimation of MSE$(\hat{Y})$ faces the same general difficulty as model-based MSE estimaton. Indeed, similarly to reduced-sample cross-validation, we could easily estimate unbiasedly the design-based MSE of the hypothetical prediction estimator

$$\hat{Y}_1 = \sum_{i \in s_1} y_i + \sum_{j \in U \setminus s_1} \mu(x_j, s_1)$$

i.e. as if s_1 was the sample (instead of the actual s) such that $\mu(x, s_1)$ was the full-sample once-trained predictor. Analogously to Theorem 1, we would have $E_q\{\hat{B}_1 \mid s\}$ as an unbiased estimator of the design-based bias of $\hat{Y}_1$, where

$$\hat{B}_1 = \sum_{i \in s_2} \pi_{2i}^{-1} \{\mu(x_i, s_1) - y_i\}$$

and $E_q\{\hat{B}_1^2 - \hat{V}_s(\hat{B}_1) \mid s\}$ unbiasedly for MSE$(\hat{Y}_1)$, where $\hat{V}_s(\hat{B}_1)$ is unbiased for

$$V_s(\hat{B}_1) = \sum_{i \notin s_1} \sum_{j \notin s_1} \big(\frac{\pi_{2ij}}{\pi_{2i}\pi_{2j}} - 1\big)\{\mu(x_i, s_1) - y_i\}\{\mu(x_j, s_1) - y_i\}$$

Next, as will be illustrated later, we note that MSE$(\hat{Y}^{RB})$ using $\bar{\mu}(x, s)$ can provide a good approximation to MSE$(\hat{Y})$ using $\mu(x, s)$, where $\hat{Y}^{RB}$ and $\hat{Y}$ should be close to each other now that both of them are trained on the full sample, as

long as μ is 'stable' in a suitable sense. Specifically, we have

$$\hat{Y} - Y \equiv E_q(\hat{Y} - Y \mid s) = E_q\{\sum_{j \notin s} \mu(x_j, s) - \mu(x_j, s_1) \mid s\} + (\hat{Y}^{RB} - Y)$$

where the term $E_q\{\cdot\}$ is of a lower order to $\hat{Y}^{RB} - Y$, if μ is n_2-times q-stable. Sanguiao-Sande and Zhang (2021) define the LOO-predictor μ to be q-stable, if $\mu(x, s) - \mu(x, s_1) \xrightarrow{P} 0$ asymptotically, as $n, N \to \infty$, while $n_2 = 1$ is held fixed. Here, μ is said to be n_2-times q-stable if the same holds given any $n_2 \geq 1$.

Finally, notice that, apart from familiarity and custom, there is no reason why one cannot adopt the SRB prediction estimator (4), for which MSE estimation is unbiased, instead of using the once-trained predictor $\mu(x, s)$.

2.3 Notes on implementation

2.3.1 Sampling probability

By definition (3), π_{2i} is the conditional s_2-inclusion probability given s_1, the calculation of which generally requires $f(s \mid s_1)$ that is derived from $q(s_1 \mid s)p(s)$. However, $p(s)$ is unknown for many unequal-probability sampling methods in practice, such as the cube method (Deville and Tille, 2004), although the inclusion probability $\pi_i = \Pr(i \in s)$ is always known.

One can use instead another sampling probability of the pq-design. For any $i \in U$, let its conditional test-set inclusion probability *given* $i \notin s_1$ be

$$\phi_{2i} = \Pr(i \in s_2 \mid i \notin s_1) = \frac{\Pr(i \in s_2, i \notin s_1)}{\Pr(i \notin s_1)} = \frac{\pi_i\{1 - \Pr(i \in s_1 \mid i \in s)\}}{1 - \pi_i \Pr(i \in s_1 \mid i \in s)} \ . \qquad (8)$$

Given π_i, the probability ϕ_{2i} can be calculated as long as $\Pr(i \in s_1 \mid i \in s)$ does not depend on i under the subsampling design and can be specified regardless the realised s, such as SRS of s_1 from s with or without replacement, or T-fold cross-validation. Since

$$\phi_{2i} = \frac{\sum_{s_1: i \notin s_1} \sum_{s: i \in s \cap i \notin s_1} f(s \mid s_1) f(s_1)}{\sum_{s_1: i \notin s_1} f(s_1)} = \frac{\sum_{s_1: i \notin s_1} \pi_{2i} f(s_1)}{\sum_{s_1: i \notin s_1} f(s_1)} = E_{s_1}\{\pi_{2i} \mid i \notin s_1\} \ ,$$

it is the conditional expectation of non-zero π_{2i}, where $\pi_{2i} = 0$ iff $i \in s_1$.

To illustrate, take the special case of SRSWOR of s from U and SRSWOR of s_1 from s, with sample sizes $n = |s|$, $n_1 = |s_1|$ and $n_2 = |s_2| = n - n_1$. For any given s_1 and $i \notin s_1$, we have exactly

$$\pi_{2i} = \frac{\Pr(i \in s_2) f(s_1 \mid i \in s_2)}{f(s_1)} = \frac{\frac{n_2}{N}\binom{N-1}{n_1}^{-1}}{\binom{N}{n_1}^{-1}} = \frac{n_2}{N - n_1} = \phi_{2i} \ .$$

Similarly, instead of π_{2ij} in (7), one can use another conditional joint s_2-

inclusion probability

$$\phi_{2ij} = \Pr(i,j \in s_2 \mid i,j \notin s_1) = \frac{\Pr(i,j \in s_2 \mid i,j \in s)\pi_{ij}}{\Pr(i,j \notin s_1)}$$

where $\Pr(i,j \in s_2 \mid i,j \in s)$ is known for random subsampling of s_1 or T-fold cross-validation, and $\Pr(i,j \notin s_1) = 1 - \Pr(i \in s_1) - \Pr(j \in s_1) + \Pr(i,j \in s_1)$.

2.3.2 Monte Carlo SRB

By (7), the exact mse^{RB} for the LTO-SRB prediction estimator $\hat{Y}^{RB}$ requires $\mathcal{C}(n,2) = \binom{n}{2}$ sample splits. If feasible, it would also provide a good estimator of $\text{MSE}(\hat{Y})$ given any 2-times q-stable μ as explained earlier. The variance of mse^{RB} should be comparable to that of the HT variance estimator; indeed, the former tends to be smaller than the latter if the model-assisted $\hat{Y}^{RB}$ is more efficient than the HT-estimator 'assisted' only by $x_i = \pi_i N/n$ (Section 1).

The exact SRB operation may be infeasible if $\mathcal{C}(n,2)$ is too large, in which case it needs to be replaced by Monte Carlo Rao-Blackwellisation (MC-RB) using T sample splits, $T \ll \mathcal{C}(n,2)$. The corresponding MC-SRB prediction estimator, denoted by $\tilde{Y}^{RB}$, is given by replacing $\bar{\mu}(x,s)$ in (4) by

$$\tilde{\mu}(x,s) = \frac{1}{T}\sum_{t=1}^{T} \mu(x, s_1^{(t)})$$

based on sample splits $(s_1^{(t)}, s_2^{(t)})$ for $t = 1,...,T$. The MSE of $\tilde{Y}^{RB}$ is then

$$\text{MSE}(\tilde{Y}^{RB}) = \text{MSE}(\hat{Y}^{RB}) + E_p\{V_q(\hat{Y}_1^* \mid s)\}/T \ .$$

The unbiased exact-RB estimator of $\text{MSE}(\tilde{Y}^{RB})$ follows from (7), i.e.

$$\text{mse}^{RB}(\tilde{Y}^{RB}) = E_q\{\hat{B}^2 - \hat{V}_s(\hat{B} \mid s_1) + \hat{V}_s\{B(s_2) \mid s_1\} \mid s\} - \frac{T-1}{T} V_q(\hat{Y}_1^* \mid s),$$

such that the unbiased MC-RB estimator of $\text{MSE}(\tilde{Y}^{RB})$ is given by

$$\widetilde{\text{mse}}^{RB} = \frac{1}{T}\sum_{t=1}^{T}\Big(\hat{B}^{2(t)} - \hat{V}_s(\hat{B} \mid s_1^{(t)}) + \hat{V}_s\{B(s_2) \mid s_1^{(t)}\} - \{\hat{Y}_1^{*(t)} - \tilde{Y}^{RB}\}^2\Big), \qquad (9)$$

where the index t explicates the computation results for $t = 1,...,T$.

For the LTO $\tilde{Y}^{RB}$, the ratio $V_{pq}(\widetilde{\text{mse}}^{RB})/V_p(\text{mse}^{RB})$ converges to 1 much slower than $\text{MSE}(\tilde{Y}^{RB})/\text{MSE}(\hat{Y}^{RB})$, as T increases. The inflation of $V_{pq}(\widetilde{\text{mse}}^{RB})$ is almost entirely due to approximating the $E_q\{\cdot\}$-term in (7) by MC in (9), wherein the terms are all estimated from $s_2^{(t)}$ of the size $n_2 = 2$. While the MC variance of $V_{pq}(\widetilde{\text{mse}}^{RB})$ can be reduced by increasing either T or n_2, the reduction is considerably faster as n_2 increases. Note that the condition of n_2-times q-stability of μ is put under a greater pressure as n_2 increases, which might affect whether

MSE($\tilde{Y}^{RB}$) would remain close to MSE($\hat{Y}$), where $\hat{Y}$ is based on $\mu(x, s)$.

Therefore, in practice, one can first explore the MC variance of $\widetilde{\text{mse}}^{RB}$ in relation to n_2 using a moderately large T, in order to choose a value of n_2 that reduces the MC variance as much as possible without jeopardising the n_2-times stability condition, and then compute $\widetilde{\text{mse}}^{RB}$ with the chosen n_2 using T that is as large as computationally practical.

2.4 Ilustration

Let us illustrate with a simple example. Generate and fix a population of size $N = 1000$ by $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$ with IID $x_{1i} \sim \text{LogN}(1, 1)$, $x_{2i} \sim \text{Poisson}(5)$ and $\epsilon_i \sim N(0, \sigma^2/4)$, where σ^2 is the population variance of x_{1i}. Let s be given by SRSWOR with $n = 100$. Let SRSWOR be the subsampling q-design of s_1 with size n_1 that is to be specified. Let the mis-specified predictor be $\mu(x, s) = x^\top \beta$, with $x = (1, x_1)$ but omitting x_2. The full-sample once-trained $\mu(x, s)$ and the SRB $\bar{\mu}(x, s)$ are approximately but not exactly equal to each other, where

$$\mu(x, s) = x^\top (\sum_{i \in s} x_i x_i^\top)^{-1} (\sum_{i \in s} x_i y_i)$$

$$\bar{\mu}(x, s) = E_q\{\mu(x, s_1) \mid s\} = x^\top E_q\Big((\sum_{i \in s} \mathbb{I}(i \in s_1) x_i x_i^\top)^{-1} (\sum_{i \in s} \mathbb{I}(i \in s_1) x_i y_i) \mid s\Big)$$

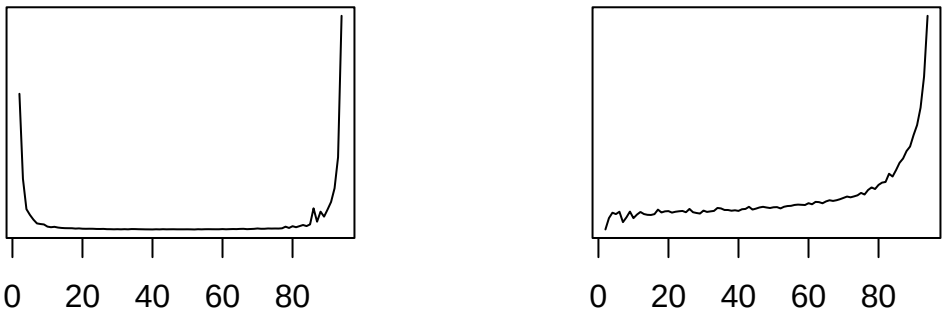

Figure 1: MC variance (left) and expectation (right) of $\widetilde{\text{mse}}^{RB}$ given $n_2 \geq 2$

Given a single realised sample s, as in practice, Figure 1 illustrates the MC variance (left) and expectation (right) as n_2 increases from 2 towards n under this setup, given $T = 10^3$. The MC variance is seen to decrease sharply to a plateau as n_2 increases from 2, before it increases dramatically again as n_2 gets close to n. Meanwhile, the MC expectation is quite stable, say, for $n_2 \leq 40$. For example, setting n_2 to be 20 or 30 for the final computation of $\widetilde{\text{mse}}^{RB}$ using $T = 10^5$, we would obtain 0.020 or 0.014 as the MC coefficient of variance (CV), which seem acceptable for practical purposes.

Table 1 shows the results of simulating MSE estimation based on 250 independent samples, given $T = 10^3$ and $n_2 = 2, 20, 30$. The MSE is simply the average squared error of either $\hat{Y}$ or $\hat{Y}^{RB}$ over the 250 samples, and the relative efficiency (RE) is the ratio between either MSE and the variance of the

Table 1: MSE estimation from 250 samples, $T = 10^3$, $\mu(x, s)$ for $\hat{Y}$ and $\bar{\mu}(x, s)$ for $\hat{Y}^{RB}$, (training, test) set of size (n_1, n_2), RE against variance of HT-estimator.

(n_1, n_2)	MSE($\hat{Y}$)	RE($\hat{Y}$)	MSE($\hat{Y}^{RB}$)	RE($\hat{Y}^{RB}$)	CV($\widetilde{\text{mse}}^{RB}$)
(98, 2)	386532.7	0.44	386632.4	0.44	3.48
(80, 20)	363613.9	0.41	363441.5	0.41	0.31
(70, 30)	362673.0	0.41	357146.9	0.41	0.21

HT-estimator. Notice that the three MSE($\hat{Y}$) here are all estimators of the same MSE, each using 250 independent samples, since $\hat{Y}$ depends only on s.

For n_2 up to 20 (or even 30), MSE($\hat{Y}^{RB}$) is practically equal to MSE($\hat{Y}$). The CV of the MC-MSE estimator $\widetilde{\text{mse}}^{RB}$ is drastically reduced by setting n_2 to 20 or 30 instead of 2. In comparison, the CV of the exact-RB MSE estimator mse^{RB} is 0.14 by simulation, whereas the CV of the HT variance estimator is 0.32. This confirms that setting n_2 to be 20 (or even 30) and using a larger but practical T would work satisfactorily for MSE estimation in this setup.

In terms of the choice of estimator, we notice that the mis-specified predictor $\mu(x, s)$ yields a design-based MSE that is less than half of the variance of the HT-estimator, and the bias of $\hat{Y}$ or $\hat{Y}^{RB}$ is a negligible part of the MSE here, the details of which are omitted to save space. The assessment is enabled by the design-based predictive inference theory developed above. Finally, as mentioned before, there is no reason why one cannot adopt $\hat{Y}^{RB}$ using $\bar{\mu}(x, s)$, for which MSE estimation is unbiased, instead of $\hat{Y}$ using $\mu(x, s)$.

3 Individual prediction estimator

Consider the individual-level predictor $\mu(x, s)$ in the prediction estimator (1). Regardless how $\mu(x, s)$ is obtained from $\{(y_i, x_i) : i \in s\}$, using whichever model or algorithm, its *total squared error (TSE)* over $R = U \setminus s$ is given by

$$D(s; \mu) = \sum_{i \in R} \{\mu(x_i, s) - y_i\}^2 \ .$$

For design-based individual-level predictive inference, we define the *risk* of μ to be the expectation of $D(s; \mu)$ over repeated sampling of $s \sim p(s)$, while treating y_U and x_U as fixed, denoted by

$$\tau(\mu) = E_p \{D(s; \mu)\} . \tag{10}$$

We stress that only s is random in (10), i.e. it is a design-based measure, regardless the model or algorithm by which $\mu(x, s)$ is constructed.

3.1 Risk of SRB predictor

Under the same pq-design of (s_1, s) as in Section 2, the error of the subsample-trained predictor $\mu(x, s_1)$ for any $i \notin s_1$ is given by

$$e_i(\mu, s_1) = \mu(x_i, s_1) - y_i$$

which can be observed for any unit in the test set $s_2 = s \setminus s_1$. Let

$$D_R(s_1; \mu) = \sum_{i \in R} e_i(\mu, s_1)^2$$

be the TSE of $\mu(x, s_1)$ over $R = U \setminus s$. Let

$$A_2 = \sum_{i \in s_2} e_i(\mu, s_1)^2 = \sum_{i \in U \setminus s_1} e_i(\mu, s_1)^2 - D_R(s_1; \mu) \ .$$

Given s_1, both A_2 and $D_R(s_1; \mu)$ vary with $s_2 = s \setminus s_1$ under the pq-design, but their sum $\sum_{i \in U \setminus s_1} e_i(\mu, s_1)^2$ is fixed. The predictor

$$\hat{D}_R(s_1; \mu) = \sum_{i \in s_2} \big(\pi_{2i}^{-1} - 1\big) e_i(\mu, s_1)^2$$

is unbiased for $D_R(s_1; \mu)$ conditional on s_1, since

$$E_s\{\hat{D}_R(s_1; \mu) \mid s_1\} = E_s\{ \sum_{i \in s_2} \pi_{2i}^{-1} e_i(\mu, s_1)^2 - A_2 \mid s_1\} = E_s\{D_R(s_1; \mu) \mid s_1\} \ . \tag{11}$$

Meanwhile, the TSE of the SRB-predictor $\bar{\mu}(x, s)$ given by (5) is

$$D(s; \bar{\mu}) = \sum_{i \in R} e_i(\bar{\mu})^2 \qquad \text{and} \qquad e_i(\bar{\mu}) = \bar{\mu}(x_i, s) - y_i \ .$$

For any $i \in R$ with $x_i = x$, the errors of $\bar{\mu}(x, s)$ and $\mu(x, s_1)$ are related by

$$e_i(\mu, s_1) = \mu(x, s_1) - y_i = \{\mu(x, s_1) - \bar{\mu}(x, s)\} + e_i(\bar{\mu}) \ .$$

Since $\bar{\mu}(x, s)$ and $e_i(\bar{\mu})$ are constant of $s_1 \sim q(s_1 \mid s)$, we have

$$e_i(\bar{\mu})^2 = E_q\{e_i(\mu, s_1)^2 \mid s\} - E_q\{a_i(\mu, s_1)^2 \mid s\} \tag{12}$$

where $a_i(\mu, s_1) = \mu(x, s_1) - \bar{\mu}(x, s)$ and $E_q\{a_i(\mu, s_1)^2 \mid s\}$ is the variance of $\mu(x, s_1)$ under the SRB operation. Design-unbiased estimation of the risk $D(s; \bar{\mu})$ is given by the result below, the proof of which is given in Appendix A.

Theorem 2. *For any given $\mu(\cdot)$, an unbiased estimator of the risk $\tau(\bar{\mu})$ of the*

corresponding SRB-predictor $\bar{\mu}(x,s)$*, over* $s \sim p(s)$*, is given by*

$$\hat{D}(s;\bar{\mu}) = E_q\Big(\sum_{i\in s_2}(\pi_{2i}^{-1}-1)\left\{e_i(\mu,s_1)^2 - a_i(\mu,s_1)^2\right\} \mid s\Big) \ .$$

In practice, where exact SRB is infeasible numerically, one can use the MC-SRB predictor based on T subsamples, which is given as

$$\begin{cases} \tilde{\mu}(x_i,s) = T^{-1}\ \sum_{t=1}^{T}\mu(x_i,s_1^{(t)}) & \text{if } \ i\in R \\ \mathring{\mu}(x_i,s) = T_i^{-1}\sum_{t=1}^{T}\mathbb{I}\big(i\notin s_1^{(t)}\big)\mu(x_i,s_1^{(t)}) & \text{if } \ i\in s \end{cases}$$

where $s_1^{(t)}$ is the t-th subsample, $T_i = \sum_{t=1}^{T}\mathbb{I}\big(i\notin s_1^{(t)}\big)$ and $s_2^{(t)} = s\setminus s_1^{(t)}$.

To estimate the risk, for any $i\in s_2^{(t)}$, let $e_i(\mu,s_1^{(t)}) = \mu(x_i,s_1^{(t)}) - y_i$ directly, and let $a_i(\mu,s_1^{(t)}) = \mu(x_i,s_1^{(t)}) - \mathring{\mu}(x_i,s)$ be an out-of-bag approximation to $\mu(x_i,s_1^{(t)}) - \bar{\mu}(x_i,s)$, instead of $\mu(x_i,s_1^{(t)}) - \tilde{\mu}(x_i,s)$ that would have been a residual-based alternative. The MC risk estimator is given by

$$\tilde{D}(s;\bar{\mu}) = \frac{1}{T}\ \sum_{t=1}^{T}\sum_{i\in s_2^{(t)}}\big(\pi_{2i}^{-1}-1\big)\{e_i(\mu,s_1^{(t)})^2 - a_i(\mu,s_1^{(t)})^2\} \ . \tag{13}$$

3.2 Using an ensemble of predictors

By design-based predictive inference, there is no need to assume that a true model exists for y_U, or that one is able to identify the true model under repeated sampling. It is then natural to combine an ensemble of different predictors (e.g. Dietterich, 2000; Zhou, 2012; Sagi and Rokach, 2018; Dong et al., 2020) in addition to selecting a single model and the corresponding predictor. Ensemble SRB prediction by voting or averaging is developed below.

3.2.1 SRB-selector

Consider selecting a single model by *voting* given an order-K heterogeneous ensemble $\{\mu_1,...,\mu_K\}$. Let $D(s;\mu_k) = \sum_{i\in R}\{\mu_k(x_i,s) - y_i\}^2$. Denote by $\Omega = \bigcup_{k=1}^{K}\Omega_k$ the partition of the sample space such that, for any $s\in\Omega_k$ and $l\neq k$, we have

$$D(s;\mu_k) < D(s;\mu_l)$$

where we discount the possibility of $D(s;\mu_k) = D(s;\mu_l)$ merely to simplify the exposition. To select a single predictor for R based on a given sample s, which minimises the risk (10), one would vote for $\mu_k(x,s)$ iff $s\in\Omega_k$. The optimal selector is thus the perfect classifier of $\mathbb{I}(s\in\Omega_k)$.

In practice it is a common approach to apply cross-validation and majority-vote, where cross-validation is based on $s_1 \sim q(s_1 \mid \ s)$ and $s_2 = s\setminus s_1$. The expected selection result is given by the *SRB-selector* as follows. Given any

(s_1, s_2) by $q(s_1 \mid s)$ and any $k = 1, ..., K$, let

$$\delta_k(s_1) = \begin{cases} 1 & \text{if } \; k = \arg\min\limits_{l=1,\ldots,K} \; \sum\limits_{i \in s_2} \{\mu_l(x_i, s_1) - y_i\}^2 \\ 0 & \text{otherwise} \end{cases}$$

indicate which predictor has the least TSE in s_2. The SRB-selector

$$\bar{\delta}_k(s) = \begin{cases} 1 & \text{if } \; k = \arg\max\limits_{l=1,\ldots,K} \; E_q\left\{\delta_k(s_1) | s\right\} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

is a classifier of $\mathbb{I}(s \in \Omega_k)$, i.e. the expected majority-vote over cross-validation.

Given the selection by (14), say, μ_k, one can reuse the same cross-validation samples (s_1, s_2) to obtain the selected SRB-predictor $\bar{\mu}_k$ and its associated risk.

3.2.2 Mixed SRB-predictor

Consider averaging given an order-K ensemble $\{\mu_1, ..., \mu_K\}$, and let the *mixed SRB-predictor* be

$$\mu(x, s) = \sum_{k=1}^{K} w_k \bar{\mu}_k(x, s) \tag{15}$$

where $\sum_{k=1}^{K} w_k = 1$ and $w_k > 0$ for the mixing weights, $k = 1, \ldots, K$. We have

$$D(s; \mu) = \sum_{k \neq 1} w_k^2 D_{kk} + \big(1 - \sum_{l \neq 1} w_l\big)^2 D_{11} + \sum_{k=1}^{K} \sum_{l \neq k,1} w_k w_l D_{kl} + 2 \sum_{k \neq 1} w_k (1 - \sum_{l \neq 1} w_l) D_{1k}$$

now that $w_1 = 1 - \sum_{k \neq 1} w_k$, where $D_{kk} = D(s; \bar{\mu}_k)$ and $D_{kl} = D(s; \bar{\mu}_k, \bar{\mu}_l)$ is given by

$$D_{kl} = \sum_{i \in R} e_i(\bar{\mu}_1) e_i(\bar{\mu}_2) = \sum_{i \in R} E_q\{e_i(\mu_1, s_1) e_i(\mu_2, s_1) \mid s\} - E_q\{a_i(\mu_1, s_2) a_i(\mu_2, s_2) \mid s\},$$

i.e. similarly to (12). An estimator of D_{kl} follows as a corollary of Theorem 2, as well as its MC implementation similarly to (13).

The optimal mixing weights w_k minimise $D(s; \mu)$. The estimated $\hat{w}_k$ can be obtained via $\hat{D}(s; \mu)$ given $\hat{D}_{kl}$, for all $k, l = 1, ..., K$. Substituting $\hat{w}_k$ in (15) yields the mixed SRB-predictor. The associated risk (10) can be estimated by $\hat{D}(s; \mu)$.

Whilst the above approach aims at minimising the risk, it may experience instability if the ensemble is not sufficiently heterogeneous. A robust approach to mixed ensemble prediction should automatically aim at the same mixing weight of two component predictors that are equal to other.

For any $k = 1, ..., K$, write $\hat{D}(s; \bar{\mu}_k) = E_q(\hat{\tau}_k \mid s)$, similarly to $\hat{D}(s; \bar{\mu})$ in Theorem 2. Regarding the risk of $\bar{\mu}_k(x, s)$ defined by (10), we have

$$\tau(\bar{\mu}_k) = E_{s_1}\{\tau(\bar{\mu}_k | s_1)\} = E_{s_1}\left\{E_s\big(\hat{\tau}_k \mid s_1\big)\right\} = E_p\left\{E_q\big(\hat{\tau}_k \mid s\big)\right\}$$

where $\tau(\bar{\mu}_k|s_1)$ is its conditional risk given s_1. Let the SRB operation yield

$$w_k = E_q(\delta_k \mid s), \quad \delta_k = \begin{cases} 1 & \text{if } \ k = \arg\min_{l=1,\ldots,K} \hat{\tau}_l \\ 0 & \text{otherwise} \end{cases}. \tag{16}$$

The corresponding mixed SRB-predictor (15) is robust against $\mu_k \approx \mu_l$ for any $k \neq l$. While the SRB-selector (14) is a binary classifier taking the majority-vote over all (s_1, s_2), the robust mixing weight (16) is a proportion over all (s_1, s_2).

3.3 Illustration

Simulations below provide a simple illustration of design-based individual-level predictive inference. For better appreciation of the design-based approach, we include also the risk estimator under an assumed IID error model.

We generate 200 sets of y_U of population size $N = 2000$ in an ad hoc manner. For each y_U, half of them are generated by M1 below and half of them by M2, where $x_1 \sim N(0,1)$ and $x_2 \sim \text{Poisson}(5)$,

$$\textbf{(M1)} \quad y = x_1 + 0.5x_2 + \epsilon, \quad \epsilon \sim \begin{cases} N(0,1) & \text{if } z = 1 \Leftrightarrow x_2 < 3 \\ N(-2,1) & \text{if } z = 2 \Leftrightarrow 3 \leq x_2 < 7 \\ N(2,1) & \text{if } z = 3 \Leftrightarrow x_2 \geq 7 \end{cases}$$

$$\textbf{(M2)} \quad y = 0.5 + 1.5x_1 + x_2 + \epsilon, \quad \epsilon \sim z^2 + N(0, 0.25), \quad z \sim N(0,1),$$

From each population we draw a sample of size $n = 200$ either by SRSWOR or Poisson Sampling. For Poisson Sampling, we set $\pi_i^{-1} \propto 1 + 1/\exp(\alpha + 0.5y_i)$ and $\sum_{i \in U} \pi_i = n$, where $\alpha \in \{1, -0.1, -1\}$ leads to the coefficient of variation of π_i over U, denoted by cv_π, to be about 15%, 30% and 45%, respectively. This illustrates a situation where sample selection may cause issues for uncertainty assessment by the IID model.

Let an order-3 model ensemble contain linear regression, random forest and support vector machine. Let the feature vector be $x = (x_1, x_2)$ in all the cases. We use a 70-30 random split for subsampling of (s_1, s_2) and let $T = 50$ for relevant MC-SRB operations such as (13). We obtain thus the SRB-predictor as described in (5) corresponding to each model.

For each SRB predictor, we estimate its standardised risk (10), $\tau/|R|$, as described before, where we have $\pi_{2i} \equiv n_2/(N - n_1)$ under SRSWOR given $n_1 = |s_1|$ and $n_2 = |s_2|$, and we use ϕ_{2i} given by (8) instead of π_{2i} under Poisson Sampling. Note that if $\hat{\tau}$ is unbiased for τ over repeated sampling from a given population, then it is also unbiased for $D/|R|$ over all the 200 populations.

The average of the 200 true $D/|R|$ for each SRB predictor will be referred to

as average true MSE in the results below, which is given by

$$\mathrm{MSE}_{\mathrm{true}} = \frac{1}{200}\sum_{b=1}^{200}\frac{1}{N-n}\sum_{i\in R^{(b)}}\{\tilde{\mu}(x_i, s^{(b)}) - y_i\}^2$$

where $s^{(b)}$ denotes the b-th simulated sample and $R^{(b)}$ the corresponding out-of-sample units. The proposed MSE estimator (13) will be compared to two MSE estimators that rely on the IID error model (e.g., James et al., 2013): the residual-based estimator and the cross-validation (CrV) estimator. The average of the latter two estimators over the 200 populations are given as

$$\mathrm{MSE}_{\mathrm{resid}} = \frac{1}{200}\sum_{b=1}^{200}\frac{1}{n}\sum_{i\in s^{(b)}}\{\tilde{\mu}(x_i, s^{(b)}) - y_i\}^2$$

$$\mathrm{MSE}_{\mathrm{CrV}} = \frac{1}{200T}\sum_{b=1}^{200}\sum_{t=1}^{T}\frac{1}{n_2}\sum_{i\in s_2^{(b,t)}}\{\mu(x_i, s_1^{(b,t)}) - y_i\}^2$$

where $s^{(b)} = s_1^{(b,t)} \cup s_2^{(b,t)}$ signifies the t-th subsampling of the b-th sample.

Table 2: MSE and estimates given each model, averaged over 200 simulations. PS, Poisson Sampling; LR, Linear regression; RF, Random forest; SVM, Support vector machine.

	SRSWOR			PS (cv_π=15%)		
MSE	LR	RF	SVM	LR	RF	SVM
Average, true	8.399	9.013	9.272	8.566	9.225	9.671
Design, proposed	8.409	9.073	9.326	8.416	9.182	9.615
Model, CrV	8.457	9.481	9.862	8.014	9.214	9.405
Model, residual	8.162	5.105	7.706	7.766	4.945	7.578
	PS (cv_π=30%)			PS (cv_π=45%)		
MSE	LR	RF	SVM	LR	RF	SVM
Average, true	8.957	9.726	10.451	9.866	10.884	11.573
Design, proposed	8.711	9.559	10.196	9.288	10.364	10.974
Model, CrV	7.624	8.880	8.799	6.992	8.262	7.933
Model, residual	7.369	4.731	7.330	6.776	4.367	6.758

Table 2 displays average true MSE and its estimates across the simulation settings. Regardless the model, the proposed design-based risk estimator (13) is unbiased under SRSWOR p-design where π_{2i} is known. Whereas, using ϕ_{2i} instead of π_{2i} under informative Poisson sampling, it remains essentially unbiased when $\mathrm{cv}_\pi = 15\%$ or 30%, but the approximation may be seen to have caused some underestimation as cv_π increases to 45%, where the severest underestimate is $(9.288 - 9.866)/9.866 \times 100 = -5.86\%$.

The CrV-based IID-model MSE estimator is also essentially unbiased under SRSWOR, because the out-of-bag squared errors in the test sample s_2 have

Table 3: MSE and estimates for ensemble individual prediction, averaged over 200 simulations; predictor selected by majority-vote, or averaged by optimal or robust mixing weights. PS, Poisson Sampling.

	SRSWOR			PS (cv_π=15%)		
MSE	Selected	Optimal	Robust	Selected	Optimal	Robust
Average, true	8.432	8.367	8.380	8.570	8.558	8.578
Design, proposed	8.395	8.260	8.284	8.412	8.343	8.372
Model, CrV	8.453	8.341	8.374	8.015	7.981	8.009
Model, residual	8.076	7.264	7.178	7.746	7.146	7.008
	PS (cv_π=30%)			PS (cv_π=45%)		
MSE	Selected	Optimal	Robust	Selected	Optimal	Robust
Average, true	8.980	8.985	9.012	9.897	9.915	9.981
Design, proposed	8.700	8.657	8.691	9.281	9.257	9.316
Model, CrV	7.631	7.618	7.647	7.006	6.997	7.035
Model, residual	7.289	6.902	6.667	6.989	6.977	7.008

the same mean as those in R conditional on s_1 under the pq-design. This is reasonable since the IID error model would hold exactly under SRS with replacement. As mentioned in Section 2.2, Bates et al. (2023) explain why the CrV-based MSE estimator is not exactly unbiased for the full-sample once-trained predictor $\mu(x, s)$, even when the IID model is correct. The CrV-based MSE estimator is instead applied to the SRB-predictor $\bar{\mu}(x, s)$ here.

The CrV-based MSE estimator can become severely biased though, if the IID model does not hold for the actual sample selection mechanism, as illustrated here for Poisson Sampling with increasing cv_π. Furthermore, residual-based MSE estimation should be avoided even under SRSWOR p-design, since it generally leads to large biases for predictors derived from highly flexible machine learning models such as random forest and support vector machine.

Next, to illustrate inference for ensemble individual prediction, we obtain the SRB-predictor (5) selected by (14), and the two mixed SRB-predictors using the weights that are either optimal for (15) or robust (16). Given any ensemble MC-SRB predictor $\tilde{\mu}(x, s^{(b)})$, for $b = 1, ..., 200$, we calculate its design-based risk estimator using (13) and following the corresponding description in Section 3.2. Whereas the two IID model-based MSE estimators are calculated in the same way as given above.

Table 3 presents average true MSE and estimates for the SRB selector and mixed SRB predictors across the simulations. The average true MSE by the SRB-selector is similar to that of the linear model in Table 2, where the MSE is the smallest by this model than random forest or support vector machine. The two mixed SRB predictors achieve largely the same true MSE for individual prediction, in each simulation setting, illustrating the robustness of ensemble prediction approach even when it cannot improve on the best single model in the given ensemble. In terms of MSE estimation, the results in Table 3 are

seen to be consistent with what we have observed for Table 2.

4 Illustrative application

Here we describe an illustrative application of design-based prediction estimation for the Spanish Structural Business Survey (SSBS). The SSBS provides information about the main structural and economic characteristics of businesses, such as employed personnel, turnover, purchases, personnel expenses, taxes and investments. The target population consists of businesses classified in one of the following economic sectors: industrial sector, commercial sector and services sector.

We take the year 2020 for reference, where the SSBS population contained 2,615,811 business units and the estimated total turnover was 1,785 billion euros, which is related to the total value of market sales of goods and services to third parties during the reference year. The SSBS estimation is traditionally based on the HT-estimator. One of motivations of this study, which is directly related to the SSBS, is the need to investigate whether it is possible to reduce the SSBS sample size, by developing and introducing more efficient estimation approaches.

4.1 pq-design for SSBS

The SSBS sample contains both fully surveyed business units and other units that are mainly imputed. For our purpose here, we shall only consider the sample of fully surveyed units, which are selected using a stratified random sampling design. There is as usual an exhaustive (i.e. take-all) stratum for the largest businesses, which will be excluded from the application below, since sampling error does not exist there. In addition, any stratum with only 1 or 2 sample units will be removed, because some variance smoothing techniques would be needed for these strata in practice, which have no direct relevance to the theory of design-based predictive inference.

A total of 9,681 strata are retained in this way, which contain altogether 2,018,561 population units. As shown in the top plot of Figure 2, the stratum population size is relatively small for most strata but has a skewed distribution. The biggest stratum does have 54,770 units and there are 319 strata with more than one thousand units. The histogram of stratum sample size is given in the bottom plot of Figure 2. Around a half of all the strata have five or fewer sample units, whereas the number of strata with sample size greater than 25 is 339. The total sample size is 80,280.

To investigate the potentials of sample size reduction, we selected randomly and without replacement 45% of the original sample units in each stratum, subjected to a minimum of three sample units in each stratum. The resulting total sample size is 40,514, which is about half of the original sample size.

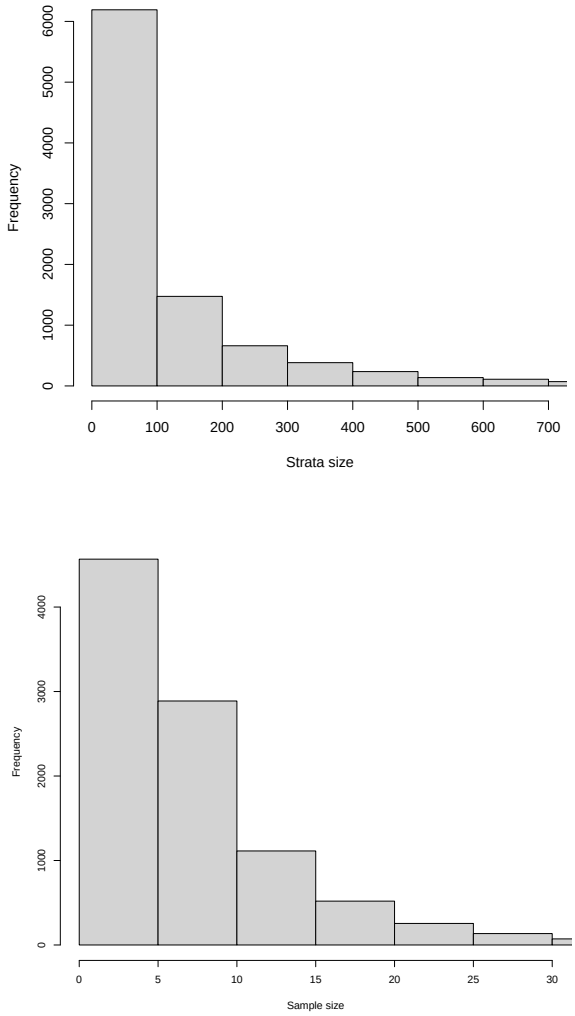

Figure 2: Histogram of stratum size in population (top) and sample (bottom)

Alternative model-assisted estimators (to be described below) will be compared based on this reduced sample, which would demonstrate both the proposed inference approach and the potentials of sample reduction.

We notice that although the ad hoc reductions of stratum sample sizes above should not be taken as a proposal for the new SSBS sampling design, the estimation results based on this realised sample are more 'tangible' than the alternative, whereby one first estimates the MSE of a given estimator based on the original sample and then speculates how this MSE might have changed had the total sample size been reduced by 50%. Moreover, insofar as our aim here is not the specifics of the new SSBS sampling design, the single realised sample is large enough to warrant the comparison of different estimation approaches, and there is no need to simulate the sample reduction above many times which would only have generated largely similar results.

We adopt within-stratum SRSWOR as the subsampling q-design for any

SRB-based estimator, in addition to the stratified random sampling p-design above. A constant subsampling rate will be implemented in all the strata, where the stratum subsample sizes of (n_1, n_2) are truncated to integers by the floor function, subjected to the constraint that $n_2 \geq 2$ in each stratum, such that variance estimation is feasible. Although other subsampling q-designs and stratum-specific subsampling rates can be explored, a systematic investigation in this respect is beyond the scope of this paper in any case.

4.2 Models and estimators

We use turnover as the target variable for illustration. The HT-estimators based on the original SSBS sample (with 80,280 units) and the reduced sample (with 40,514 units) provide the baselines for comparison. Four additional estimators will be applied to the reduced sample, which arise from the 2×2 combination of models (linear, tree) and estimators (prediction, unbiased) as described below.

The model is either linear or tree regression, defined globally regardless the design strata. The linear model uses four features:

- the 'administrative' turnover from the corporate incoming tax if available, or imputed stratum-mean (of available administrative turnover) if missing;
- a binary indicator for whether the administrative turnover is missing;
- the operating income according to the tax administration if available, or imputed stratum-mean (of available operating income) if missing;
- a binary indicator for whether the operating income is missing.

The linear regression coefficients are estimated by weighted ridge regression, given the sampling design weights and a small tuning parameter $\lambda = 0.01$ for the regularisation penalty.

The tree regression model uses the following four features, where the missing values are *not* imputed but left as-is to the software package:

- the administrative turnover,
- the operating income,
- the first digit of the National Classification of Economic Activities,
- the number of employees according to the Business Register.

The tree model is built using the ready-made R package `h2o` for random forest, where one feature is chosen for each split (`mtries = 1`), the maximum tree depth is 20 (`max_depth = 20`) and the minimum number of observations per leaf is 5 (`min_rows = 5`). The observations are again weighted by the sampling weights.

Given either model, we consider the SRB prediction estimator (4) and the design-unbiased model-assisted SRB-estimator (Sanguiao-Sande and Zhang, 2021), where the latter can be given as $\hat{Y}_M = E_q(\hat{Y}_{1M} \mid s)$ and

$$\hat{Y}_{1M} = \sum_{i \in s_1} y_i + \sum_{i \in U \setminus s_1} \mu(x_i, s_1) + \sum_{i \in s_2} (\pi_{2i}^{-1} - 1)\{\mu(x_i, s_1) - y_i\}$$

The SRB operation uses a constant 80-20 sample-split for the linear model and a constant 50-50 sample-split for the tree model. Notice that in the case of tree model, the SRB-predictor (5) is a random forest by construction since it is then the average of T randomly constructed trees.

It is worth pointing out that, unlike the prediction estimator (4) that applies $\bar{\mu}(x, s)$ directly to all the out-of-sample units, the SRB-estimator $\hat{Y}_M$ corrects the bias of each $\mu(x, s_1)$ using the observed errors $\{\mu(x_i, s_1) - y_i : i \in s_2\}$ via π_{2i}; see Sanguiao-Sande and Zhang (2021) for the details. Although $\hat{Y}_M$ is thus exactly design-unbiased, it may have a larger MSE than the prediction estimator (4) that uses the same model. Conversely, a prediction estimator would become less attractive if its bias is 'intolerable', even though it may have a much smaller MSE than an unbiased estimator that uses the same model.

4.3 Results

Table 4 summarises the results for the estimators described above. Apart from each estimate $\hat{Y}$, it is also given the estimated bias (zero for an unbiased estimator), the estimated MSE, the relative error (RErr) given as $\sqrt{\text{MSE}}/\hat{Y}$, and the Monte Carlo (MC) error of the MSE estimator.

Table 4: Estimation results (in billion euros), $T = 10,000$ sample-splits, based on same reduced sample size unless indicated otherwise.

Estimator, model	$\hat{Y}$	Bias	MSE	RErr	MC error
HT-estimator (full sample size)	258	0	94	0.04	-
HT-estimator	252	0	151	0.05	-
SRB-prediction estimator, linear	227	-2	50	0.03	3
SRB-prediction estimator, tree	238	4	27	0.02	5
SRB-estimator, linear	229	0	122	0.05	1
SRB-estimator, tree	234	0	107	0.04	2

Exact Rao-Blackwellisation for MSE estimation is simply beyond reach in this case, where we have more than six thousand strata with sample size three in the reduced sample. If we leave out two units in each stratum under subsampling, then there are more than 3^{6000} distinct samples under the q-design just for these strata, because the models are not built separately within each stratum, not to mention the other strata with more sample units.

The MC error is the bootstrap estimated standard deviation of the MC-MSE estimator, which is due to the loss of MSE estimation efficiency by Monte Carlo compared to exact Rao-Blackwellisation (with zero MC error). It can be seen that the relative MC error still does not vanish despite the large number of sample-splits $T = 10,000$, e.g. it is $3/50$ for the linear-model SRB-prediction estimator, and not surprisingly, the relative MC error can increase as the MSE reduces, such as $5/27$ for the tree-model SRB-prediction estimator. Nevertheless, the MSE estimation results here are reliable enough for us to distinguish between the different estimators.

First, we notice that the two design-unbiased SRB-estimators have only led to moderate efficiency gains over the HT-estimator. The main reason is likely to be the large number of strata with very few sample units. Basically, in case the SRB-estimator uses $\mu(x)$ that is given in advance, it would become a stratified difference estimator, which corrects for the design-based bias of $\mu(x)$ for each stratum population total by using only the within-stratum sample units. The situation is largely similar for the SRB-estimator using $\mu(x, s)$ estimated based on the whole sample, which may have a small sampling variance itself.

In comparison, the SRB-prediction estimators using the same models can be much more efficient precisely because they do not apply bias correction, i.e. they are no longer stratified estimators as the SRB-estimators are, such that they can take full advantage of the reduced variance if the prediction biases are small. The MSE of the linear-model SRB-prediction estimator is only about one third of the sampling variance of the HT-estimator, given the same sample size, whereas the tree model further reduces the MSE by about 50% compared to the linear model. Meanwhile, while the bias of the linear-model prediction estimator is relatively small compared to its root MSE, this is no longer the case for the tree-model prediction estimator, i.e. 4 against $\sqrt{27}$.

This serves to remind one that it is often possible to reduce the MSE at some cost of increasing bias, such as when adopting either model-assisted or model-based estimators traditionally. The theory of design-based predictive inference allows one to estimate both the bias and MSE of a large class of prediction estimators (1). This increases the scope of choice in practice, in order to achieve a sensible trade-off between bias and variance.

Finally, the illustrative results above suffice to demonstrate the potentials of sample size reduction for the SSBS. An appropriate scheme of stratum sample size reduction requires a more systematic investigation though. In particular, the accompanying estimator can be chosen from the broad class of prediction estimators (1), assisted by any model or algorithm and the various features available from the administrative source. But a detailed analysis is needed to take into account the level of dissemination (instead of just an overall total here) and the tolerable trade-off between bias and root MSE.

5 Final remarks

We have developed a theory of design-based predictive inference from finite-population probability sampling. For population total estimation, one would be interested in the total of the out-of-sample prediction errors, whereas the risk of individual-level prediction depends on the out-of-sample squared prediction errors. The SRB approach provides a unified treatment of both.

Adopting design-based predictive inference for official statistics allows one to circumvent the design vs. model controversy. In addition to producing population-level estimates, it provides a theoretical basis for producing statistical registers or census-like data for descriptive official statistics. The theory we propose allows for any assisting ML models or algorithms, which can be more efficient than calibration estimation using only auxiliary totals or the parametric assisting models commonly used in survey sampling.

There are a number of issues worth further investigation, of which we only mention a few here. First, survey nonresponse is unavoidable in practice. Lee et al. (2022) apply a related SRB ensemble learning approach to missing data imputation. It would be helpful to develop a unified SRB approach, which can incorporate survey response under an extended quasi-randomisation framework. Next, other choices of risk than the total squared prediction errors may be considered for individual prediction, or interval estimation may be developed for population total inference wherever the design-based bias of the prediction estimator is deemed non-negligible. Finally, it is worth studying how better to balance between the risk of individual prediction and the MSE of population total estimation associated with the prediction estimator (1).

Acknowledgement We thank three anonymous referees and the Associate Editor for comments that have helped to sharpen our message.

A Proofs

Proof of Theorem 1.

Proof. Conditional on s_1, we have

$$E_s(\hat{B}^2 \mid s_1) = E_s(B \mid s_1)^2 + V_s(\hat{B} \mid s_1) = \{E_s(B^2 \mid s_1) - V_s(B \mid s_1)\} + V_s(\hat{B} \mid s_1)$$

and $V_s(B \mid s_1) = V_s\{B(s_2) \mid s_1\}$ now that the total $B_1 = B + B(s_2)$ is fixed given s_1. It follows that a conditionally (on s_1) unbiased predictor of the squared total error of $\hat{Y}_1^*$ is

$$\hat{B}^2 - \hat{V}_s(\hat{B} \mid s_1) + \hat{V}_s\{B(s_2) \mid s_1)\} \ .$$

Applying Rao-Blackwellisation to it yields the $E_q\{\cdot\}$ term on the right-hand side of (7) as a more efficient unbiased estimator of the pq-MSE of $\hat{Y}_1^*$. Whereas the

last term on the right-hand side of (7) follows from noting

$$E_{pq}(B^2) = E_{pq}\{(\hat{Y}_1^* - \hat{Y}^{RB} + \hat{Y}^{RB} - Y)^2\} = E_p\{V_q(\hat{Y}_1^* \mid s)\} + \mathbf{MSE}(\hat{Y}^{RB}) \ .$$

This completes the proof. □

Proof of Theorem 2.

Proof. By (12), we have

$$D(s;\bar{\mu}) = E_q\{\sum_{i\in R} e_i(\mu, s_1)^2 \mid s\} - E_q\{\sum_{i\in R} a_i(\mu, s_1)^2 \mid s\} \ .$$

For the first term that can be rewritten as $E_q\{D_R(s_1;\mu) \mid s\}$, the estimator

$$\hat{E}_q\big(D_R(s_1;\mu) \mid s\big) = E_q\big(\hat{D}_R(s_1;\mu) \mid s\big)$$

is unbiased over $p(s)$ since, using (11), we have

$$\begin{aligned} E_p\{E_q\big(\hat{D}_R(s_1;\mu) \mid s\big)\} &= E_{s_1}\{E_s\big(\hat{D}_R(s_1;\mu) \mid s_1\big)\} \\ &= E_{s_1}\{E_s\big(D_R(s_1;\mu) \mid s_1\big)\} = E_p\left\{E_q\big(D_R(s_1;\mu) \mid s\big)\right\}. \end{aligned}$$

On replacing $e_i(\mu, s_1)^2$ by $a_i(\mu, s_1)^2$, one can carry through the same derivation for the second term of $D(s;\bar{\mu})$ above, $E_q\{\sum_{i\in R} a_i(\mu, s_1)^2 \mid s\}$. It follows that

$$E_p\{\hat{D}(s;\bar{\mu})\} = E_p\{D(s;\bar{\mu})\} = \tau(\bar{\mu}) \ .$$

This completes the proof. □

References

[1] Bates, S., Hastie, T. and Tibshirani, R. (2023) Cross-Validation: What Does It Estimate and How Well Does It Do It? *Journal of the American Statistical Association*, `https://doi.org/10.1080/01621459.2023.2197686`

[2] Beaumont, J.-F. and Haziza, D. (2022). Statistical inference from finite population samples: A critical review of frequentist and Bayesian approaches. *The Canadian Journal of Statistics*, 50:1186-1212.

[3] Berger, Y.G. and De La Riva Torres, O. (2016). Empirical Likelihood Confidence Intervals for Complex Sampling Designs. *Journal of the Royal Statistical Society Series B*, 78:319-341.

[4] Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Annals of Mathematical Statistics*, 18: 105-110.

[5] Breidt, F. J. and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32:190-205.

[6] Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

[7] Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91:893-912.

[8] Dietterich, T.G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*. Springer.

[9] Dong, X., Yu, Z., Cao, W., Shi, Y. and Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, 14:241-258.

[10] Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh and London.

[11] Geisser, S. (1993). *Predictive Inference*. Chapman & Hall.

[12] Hansen, M. (1987). Some History and Reminiscences on Survey Sampling. *Statistical Science*, 2:180-190.

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

[14] Hartley, H. O. and Rao, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

[15] Hartley, H. O. and Ross, A. (1954). Unbiased Ratio Estimators. *Nature*, August 7, pp. 270-271.

[16] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663-685.

[17] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

[18] Kalton, G. (2002). Models in practice of survey sampling. *Journal of Official Statistics*, 18:129-154.

[19] Lee, D., Zhang, L.-C. and Chen, S. (2022). Robust quasi-randomization-based estimation with ensemble learning for missing data. *Scandinavian Journal of Statistics*. DOI:10.1111/sjos.12626

[20] Mickey, M.R. (1959). Some finite population unbiased ratio and regression estimators. *Journal of the American Statistical Association*, 54:594-612.

[21] Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society, pp. 558-625.*

[22] Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta Mathematical Society*, 37:81-91.

[23] Rao, J. N. K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology*, 31:117-138.

[24] Rao, J. N. K. (2011). Impact of frequentist and Bayesian methods on survey sampling practice: A selective appraisal. *Statistical Science*, 26:240-256.

[25] Rao, J. N. K. and Wu, C. (2010). Bayesian pseudo-empirical-likelihood intervals for complex surveys. *Journal of the Royal Statistical Society B*, 72:533-544.

[26] Sagi, O. and Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, e1249.

[27] Sanguiao-Sande, L. and Zhang, L.-C. (2021). Design-Unbiased Statistical Learning in Survey Sampling. *Sankhya A*, Centenary Issue in Honour of C. R. Rao, 83:714-744.

[28] Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model-Assisted Survey Sampling*. Springer, New York.

[29] Smith, T.M.F. (1983). On the validity of inferences from non-random sample. *Journal of the Royal Statistical Society, Series A*, 146:394-403.

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

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

[32] Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96:185-193.

[33] Zhou, Z.H. (2012). *Ensemble Methods: Foundations and Algorithms*. CRC press.