Statistica Sinica Preprint No: SS-2022-0181	
Title	A Comparison of Estimators of Mean and Its Functions in Finite Populations
Manuscript ID	SS-2022-0181
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0181
Complete List of Authors	Anurag Dey and Probal Chaudhuri
Corresponding Authors	Anurag Dey
E-mails	deyanuragsaltlake64@gmail.com

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

A COMPARISON OF ESTIMATORS OF MEAN AND ITS FUNCTIONS IN FINITE POPULATIONS

Anurag Dey and Probal Chaudhuri

Indian Statistical Institute, Kolkata

Abstract: We investigate several well-known estimators of finite population means and the functions of these means under standard sampling designs. Such functions include the variance, correlation coefficient, and regression coefficient in the population as special cases. We compare the performance of these estimators under different sampling designs, based on their asymptotic distributions. We construct equivalence classes of estimators under different sampling designs so that estimators in the same class have equivalent performance in terms of the asymptotic mean squared error (MSE). We then compare estimators from different equivalence classes under super-populations that satisfy linear models. We show that the pseudo empirical likelihood (PEML) estimator of the population mean under simple random sampling without replacement (SRSWOR) has the lowest asymptotic MSE of the estimators considered here. In addition, for the variance, correlation coefficient, and regression coefficient of the population, the plug-in estimators based on the PEML estimator have the lowest asymptotic MSEs under SRSWOR. However, for any high entropy πPS sampling design, which uses auxiliary information, the plug-in estimators based on the Hájek estimator have the lowest asymptotic MSEs.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

Key words and phrases: Asymptotic normality, Equivalence classes of estimators, High entropy sampling designs, Inclusion probability, Linear regression model, Rejective sampling design, Relative efficiency, Superpopulation models.

1. Introduction

Suppose that $\mathcal{P}=\{1,2,\ldots,N\}$ is a finite population of size N, s is a sample of size n ($< N$) from $\mathcal{P}$, and $\mathcal{S}$ is the collection of all possible samples of size n. Then, a sampling design $P(s)$ is a probability distribution on $\mathcal{S}$ such that $0 \leq P(s) \leq 1$ for all $s \in \mathcal{S}$ and $\sum_{s\in\mathcal{S}} P(s)=1$. In this study, we consider the following designs: simple random sampling without replacement (SRSWOR), the Lahiri–Midzuno–Sen (LMS) sampling design (see Lahiri (1951), Midzuno (1952), and Sen (1953)), the Rao–Hartley–Cochran (RHC) sampling design (see Rao et al. (1962)), and high entropy πPS (HEπPS) sampling designs (see Section 2). Note that all of the above sampling designs other than SRSWOR use some auxiliary variable.

Let (Y_i, X_i) be the value of (y, x) for the ith population unit, for $i=1,\ldots,N$, where y is a univariate or multivariate study variable, and x is a positive real-valued size/auxiliary variable. Suppose that $\overline{Y}=\sum_{i=1}^{N} Y_i/N$ is the finite population mean of y. The Horvitz–Thompson (HT) estimator (see Horvitz and Thompson (1952))) and the RHC (see Rao et al. (1962)) estimator are pop-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

ular design unbiased estimators of $\overline{Y}$. Other well-known estimators of $\overline{Y}$ are the Hájek estimator (see Hájek (1971), Särndal et al. (2003), and the references therein), ratio estimator (see Cochran (1977)), product estimator (see Cochran (1977)), generalized regression (GREG) estimator (see Chen and Sitter (1999)), and pseudo empirical likelihood (PEML) estimator (see Chen and Sitter (1999)). However, these estimators are not always design unbiased. See the Appendix for expressions of these estimators. Now, suppose that y is a $\mathbb{R}^d$-valued ($d \geq 1$) study variable, and $g(\sum_{i=1}^N h(Y_i)/N)$ is a population parameter. Here, h: $\mathbb{R}^d \to \mathbb{R}^p$ is a function with $p \geq 1$, and g: $\mathbb{R}^p \to \mathbb{R}$ is a continuously differentiable function. All vectors in Euclidean spaces are taken as row vectors, and a superscript T denotes their transpose. Examples of such parameters are the variance, correlation coefficient, and regression coefficient associated with a finite population. For simplicity, we often write $h(Y_i)$ as h_i. Then, $g(\overline{h})$=$g(\sum_{i=1}^N h_i/N)$ is estimated by plugging in the estimator $\hat{\overline{h}}$ of $\overline{h}$.

Our objective is to find an asymptotically efficient (in terms of the mean squared error (MSE)) estimator of $g(\overline{h})$. In Section 2, using the asymptotic distribution of the estimator of $g(\overline{h})$ under the above sampling designs, we construct equivalence classes of estimators such that any two estimators in the same class have the same asymptotic MSE. In Section 3, we consider the special case of $g(\overline{h})$=$\overline{Y}$, and compare the equivalence classes of estimators under superpop-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

ulations that satisfy linear models. For the estimators considered here under different sampling designs, the PEML estimator of the population mean under SRSWOR has the lowest asymptotic MSE. Furthermore, the PEML estimator has the same asymptotic MSE under SRSWOR and the LMS sampling design. Interestingly, the performance of the PEML estimator under the RHC and any HEπPS sampling designs, which use auxiliary information, is worse than that under SRSWOR. The GREG estimator has been shown to be asymptotically at least as efficient as the HT, ratio, and product estimators under SRSWOR (see Cochran (1977)). It follows from our analysis that the PEML estimator is asymptotically equivalent to the GREG estimator under all sampling designs considered here.

In Section 3, we consider the cases when $g(\overline{h})$ is the variance, the correlation coefficient, and the regression coefficient in the population. Note that if the estimator of the population variance is constructed by plugging in the HT, ratio, product, or GREG estimator of the population mean, then the estimators of the variance may become negative. The same applies to the correlation coefficient and regression coefficient, because these estimators require an estimator of the population variance. On the other hand, if the estimators of the above-mentioned parameters are constructed using the Hájek or PEML estimators of the population mean, such a problem does not occur. Therefore, for these

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

parameters, we compare only those equivalence classes that contain plug-in estimators based on the Hájek and PEML estimators. Under superpopulations that satisfy linear models, we again conclude that the plug-in estimator for these parameters based on the PEML estimator has the lowest asymptotic MSE under SRSWOR and the LMS sampling design. Moreover, under any HEπPS sampling design, which uses the auxiliary information, the plug-in estimator based on the Hájek estimator has the lowest asymptotic MSE.

Scott and Wu (1981) prove that the ratio estimator has the same asymptotic distribution under SRSWOR and the LMS sampling design. Chen and Sitter (1999) show that the PEML estimator is asymptotically equivalent to the GREG estimator under conditions on the sampling design that are satisfied by SRSWOR and the RHC sampling design. However, this is the first study to produce asymptotic equivalence classes, such as those in Table 2 in Section 2, that consist of several estimators of a function of the population mean under several sampling designs.

When the study and size variables are exactly linearly related, Raj (1954) compared the sample mean under simple random sampling with replacement and the usual unbiased estimator of the population mean under the probability proportional to size sampling with replacement. Avadhani and Sukhatme (1970) compared the ratio estimator of the population mean under SRSWOR

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

with the RHC estimator under the RHC sampling design when an approximate linear relationship holds between the study variable and the size variable. Avadhani and Srivastava (1972) compared the ratio estimator of the population mean under the LMS sampling design and the RHC estimator under the RHC sampling design when the study and size variables are approximately linearly related. It has also been shown that the GREG estimator of the population mean is asymptotically at least as efficient as the HT, ratio, and product estimators under SRSWOR (see Cochran (1977)). However, the above comparisons included neither the PEML estimator nor HEπPS sampling designs.

In our empirical studies, presented in Section 4, using synthetic and real data, our numerical results support our theoretical results. Section 5 concludes the paper. All proofs are given in the Appendix.

2. Comparison of different estimators of $g(\overline{h})$

In this section, we compare the estimators of $g(\overline{h})$ obtained by plugging in the estimators of $\overline{h}$ given in Table 1. First, we find equivalence classes of estimators of $g(\overline{h})$ such that any two estimators in the same class are asymptotically normal, with the same mean $g(\overline{h})$ and the same variance.

We define our asymptotic framework as follows. Let $\{\mathcal{P}_\nu\}$ be a sequence of nested populations with N_ν, $n_\nu \to \infty$ as $\nu \to \infty$ (see Isaki and Fuller (1982),

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

Table 1: Estimators of $\overline{h}$

Sampling designs	Estimators
SRSWOR	HT (which coincides with Hájek estimator), ratio, product, GREG and PEML estimators
LMS	HT, Hájek, ratio, product, GREG and PEML estimators
HEπPS	HT (which coincides with ratio and product estimators), Hájek, GREG and PEML estimators
RHC	RHC, GREG and PEML estimators

Wang and Opsomer (2011), Conti and Marella (2015), Boistard et al. (2017), Han and Wellner (2021), and the references therein), where N_ν and n_ν are, respectively, the population size and the sample size corresponding to the νth population. Henceforth, we suppress the subscript ν that tends to ∞, for the sake of simplicity. Throughout this paper, we consider the following condition (cf. Assumption 1 in Cardot & Josserand (2011), A4 in Conti (2014), A1 in Cardot et al. (2014), A4 in Conti and Marella (2015), and (HT3) in Boistard et al. (2017)).

C 0. *$n/N \to \lambda$ as $\nu \to \infty$, where $0 \leq \lambda < 1$.*

Before we state the main results, let us discuss the HEπPS sampling design and some conditions on $\{(X_i, h_i) : 1 \leq i \leq N\}$ (recall that $h_i{=}h(Y_i)$). A sam-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

pling design $P(s)$ satisfying the condition $D(P||R)=\sum_{s\in\mathcal{S}} P(s)\log\left(P(s)/R(s)\right)\rightarrow 0$ as $\nu\rightarrow\infty$, for some rejective sampling design (see Hájek (1964)) $R(s)$, is known as a high entropy sampling design (see Berger (1998), Conti (2014), Cardot et al. (2014), Boistard et al. (2017), and the references therein). A sampling design $P(s)$ is called an HEπPS sampling design if it is a high entropy sampling design and its inclusion probabilities satisfy the condition $\pi_i=nX_i/\sum_{i=1}^N X_i$, for $i=1,\ldots,N$. An example of an HEπPS sampling design is the Rao–Sampford (RS) sampling design (see Sampford (1967) and Berger (1998)). We now state several conditions.

C 1. *$\{P_\nu\}$ is such that $\sum_{i=1}^N ||h_i||^4/N=O(1)$ and $\sum_{i=1}^N X_i^4/N=O(1)$ as $\nu\rightarrow\infty$. Further, $\lim_{\nu\rightarrow\infty}\overline{h}$ exists, and $\overline{X}=\sum_{i=1}^N X_i/N$ and $S_x^2=\sum_{i=1}^N(X_i-\overline{X})^2/N$ are bounded away from zero as $\nu\rightarrow\infty$. Moreover, $\nabla g(\mu_0)\neq 0$, where $\mu_0=\lim_{\nu\rightarrow\infty}\overline{h}$ and ∇g is the gradient of g.*

C 2. *$\max_{1\leq i\leq N} X_i/\min_{1\leq i\leq N} X_i=O(1)$ as $\nu\rightarrow\infty$.*

Let $\mathbf{V}_i$ be one of h_i, $h_i-\overline{h}$, $h_i-\overline{h}X_i/\overline{X}$, $h_i+\overline{h}X_i/\overline{X}$, and $h_i-\overline{h}-S_{xh}(X_i-\overline{X})/S_x^2$, for $i=1,\ldots,N$, $\overline{h}=\sum_{i=1}^N h_i/N$, and $S_{xh}=\sum_{i=1}^N X_ih_i/N-\overline{h}\ \overline{X}$. Define $\mathbf{T}=\sum_{i=1}^N \mathbf{V}_i(1-\pi_i)/\sum_{i=1}^N \pi_i(1-\pi_i)$, where π_i is the inclusion probability of the ith population unit. Furthermore, in the case of the RHC sampling design, define $\overline{\mathbf{V}}=\sum_{i=1}^N \mathbf{V}_i/N$, $\overline{X}=\sum_{i=1}^N X_i/N$, and $\gamma=\sum_{i=1}^n N_i(N_i-1)/N(N-1)$, where N_i is the size of the ith group formed randomly in the RHC sampling design

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

(see Rao et al. (1962)), for i=1, ..., n. Now, we state the conditions on the population values and the sampling designs.

C 3. *$P(s)$ is such that $nN^{-2}\sum_{i=1}^{N}(\boldsymbol{V}_i - \boldsymbol{T}\pi_i)^T(\boldsymbol{V}_i - \boldsymbol{T}\pi_i)(\pi_i^{-1} - 1)$ converges to some positive-definite (p.d.) matrix as $\nu \to \infty$.*

C 4. *$n\gamma\overline{X}N^{-1}\sum_{i=1}^{N}(\boldsymbol{V}_i - X_i\overline{\boldsymbol{V}}/\overline{X})^T(\boldsymbol{V}_i - X_i\overline{\boldsymbol{V}}/\overline{X})/X_i$ converges to some p.d. matrix as $\nu \to \infty$.*

Conditions similar to C1, C3, and C4 are often used in the sample survey literature (see Assumption 3 in Cardot & Josserand (2011), A3 and A6 in both Conti (2014) and Conti and Marella (2015), (HT2) in Boistard et al. (2017), and F2 and F3 in Han and Wellner (2021)). Conditions C1 and C4 hold (*almost surely*) whenever $\{(X_i,\ h_i) : 1 \leq i \leq N\}$ are generated from a superpopulation model that satisfies appropriate moment conditions (see Lemma S2 in the Supplementary Material). The condition $\sum_{i=1}^{N}||h_i||^4/N$=$O(1)$ holds when h is a bounded function (e.g., $h(y)$=y and y is a binary study variable). Condition C2 implies that the variation in the population values $X_1, \ldots, X_N$ cannot be too large. Under any πPS sampling design, C2 is equivalent to the condition that $L \leq N\pi_i/n \leq L'$, for some constants $L, L' > 0$, any i=1, ..., N, and all sufficiently large $\nu \geq 1$; see (C1) in Boistard et al. (2017) and Assumption 2-*(i)* in Wang and Opsomer (2011). Condition C2 holds (*almost surely*) when $\{X_i\}_{i=1}^{N}$ are generated from a superpopulation distribution, and the support of

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

the distribution of X_i is bounded away from zero and ∞. Condition C3 holds (*almost surely*) for SRSWOR, the LMS sampling design, and any πPS sampling design under appropriate superpopulation models (see Lemma S2 in the Supplementary Material). For the RHC sampling design, we also assume that $\{N_i\}_{i=1}^n$ is given by

$$N_i = \begin{cases} N/n, \text{ for } i = 1, \cdots, n, \text{ when } N/n \text{ is an integer,} \\ \lfloor N/n \rfloor, \text{ for } i = 1, \cdots, k, \text{ and} \\ \lfloor N/n \rfloor + 1, \text{ for } i = k+1, \cdots, n, \text{ when } N/n \text{ is not an integer,} \end{cases} \tag{2.1}$$

where k is such that $\sum_{i=1}^n N_i = N$. Here, $\lfloor N/n \rfloor$ is the integer part of N/n. Rao et al. (1962) showed that this choice of $\{N_i\}_{i=1}^n$ minimizes the variance of the RHC estimator. Now, we state the following theorem.

Theorem 1. *Suppose that C0 through C3 hold. Then, classes* $1, 2, 3$, *and* 4 *in Table 2 describe equivalence classes of estimators for* $g(\overline{h})$ *under SRSWOR and the LMS sampling design.*

For the next two theorems, we assume that $n \max_{1 \le i \le N} X_i / \sum_{i=1}^N X_i < 1$. Note that this condition is required to hold for any without-replacement πPS sampling design.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

Theorem 2. (*i*) *If C0 through C3 hold, then classes* 5, 6, *and* 7 *in Table 2 describe equivalence classes of estimators for* $g(\overline{h})$ *under any HEπPS sampling design.*

(*ii*) *Under the RHC sampling design, if C0 through C2 and C4 hold, then classes* 8 *and* 9 *in Table 2 describe equivalence classes of estimators for* $g(\overline{h})$.

Table 2: Disjoint equivalence classes of estimators for $g(\overline{h})$

	Estimators of $\overline{h}$					
Sampling design	GREG and PEML	HT	RHC	Hájek	Ratio	Product
SRSWOR	class 1	[1]class 2		[1]class 2	class 3	class 4
LMS	class 1	class 2		class 2	class 3	class 4
HEπPS	class 5	[2]class 6		class 7	[2]class 6	[2]class 6
RHC	class 8		class 9			

[1] The HT and Hájek estimators coincide under SRSWOR.

[2] The HT, ratio, and product estimators coincide under HEπPS sampling designs.

Remark 1. *If C1 through C3 hold, and C0 holds with* λ=0, *then in Table 2, class* 8 *merges with class* 5, *and class* 9 *merges with class* 6. *For details, see Section S3 in the Supplementary Material.*

Next, suppose that W_i=$\nabla g(\overline{h})h_i^T$, for i=1, . . . , N, $\overline{W}$=$\sum_{i=1}^{N} W_i/N$, S_{xw}=

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

$\sum_{i=1}^{N} W_i X_i/N - \overline{W}\ \overline{X}$, $S_w^2 = \sum_{i=1}^{N} W_i^2/N - \overline{W}^2$, $S_x^2 = \sum_{i=1}^{N} X_i^2/N - \overline{X}^2$, and $\phi = \overline{X} - (n/N)\sum_{i=1}^{N} X_i^2/N\overline{X}$. Now, we state the following theorem.

Theorem 3. *Suppose that the assumptions of Theorems 1 and 2 hold. Then, Table 3 gives expressions for the asymptotic MSEs, $\Delta_1^2, \ldots, \Delta_9^2$, of the estimators in equivalence classes $1, \ldots, 9$, respectively, in Table 2.*

Remark 2. *It can be shown in a straightforward way from Table 3 that $\Delta_1^2 \leq \Delta_i^2$, for $i = 2, 3$, and 4. Thus, the plug-in estimators of $g(\overline{h})$ based on the GREG and the PEML estimators are asymptotically as good as, if not better than, those based on the HT (which coincides with the Hájek estimator), ratio, and product estimators under SRSWOR, and those based on the HT, Hájek, ratio, and product estimators under the LMS sampling design.*

Let us now consider some examples of $g(\overline{h})$ in Table 4. The conclusions of Theorems 1 through 3 and Remarks 1 and 2 hold for all parameters in Table 4. Here, recall that for the variance, correlation coefficient, and regression coefficient, we consider only the plug-in estimators based on the Hájek and PEML estimators.

3. Comparison of estimators under superpopulation models

In this section, we derive asymptotically efficient estimators for the mean, variance, correlation coefficient, and regression coefficient under superpopulations

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

Table 3: Asymptotic MSEs of estimators for $g(\overline{h})$ (note that to simplify the notation, we omit the subscript ν from expressions on which limits are taken.)

$\Delta_1^2=(1-\lambda)\lim\limits_{\nu\to\infty}\left(S_w^2-(S_{xw}/S_x)^2\right)$
$\Delta_2^2=(1-\lambda)\lim\limits_{\nu\to\infty}S_w^2$
$\Delta_3^2=(1-\lambda)\lim\limits_{\nu\to\infty}\left(S_w^2-2\overline{W}S_{xw}/\overline{X}+\left(\overline{W}/\overline{X}\right)^2S_x^2\right)$
$\Delta_4^2=(1-\lambda)\lim\limits_{\nu\to\infty}\left(S_w^2+2\overline{W}S_{xw}/\overline{X}+\left(\overline{W}/\overline{X}\right)^2S_x^2\right)$
$\Delta_5^2=\lim\limits_{\nu\to\infty}(1/N)\sum_{i=1}^N\left(W_i-\overline{W}-(S_{xw}/S_x^2)(X_i-\overline{X})\right)^2\times$ $\left((\overline{X}/X_i)-(n/N)\right)$
$\Delta_6^2=\lim\limits_{\nu\to\infty}(1/N)\sum_{i=1}^N\left\{W_i+\phi^{-1}\overline{X}^{-1}X_i\left((n/N)\sum_{i=1}^N W_iX_i/N-\overline{W}\ \overline{X}\right)\right\}^2\times$ $\left\{(\overline{X}/X_i)-(n/N)\right\}$
$\Delta_7^2=\lim\limits_{\nu\to\infty}(1/N)\sum_{i=1}^N\left(W_i-\overline{W}+(n/N\phi\overline{X})X_iS_{xw}\right)^2\times$ $\left((\overline{X}/X_i)-(n/N)\right)$
$\Delta_8^2=\lim\limits_{\nu\to\infty}n\gamma(\overline{X}/N)\sum_{i=1}^N\left(W_i-\overline{W}-(S_{xw}/S_x^2)(X_i-\overline{X})\right)^2/X_i$
$\Delta_9^2=\lim\limits_{\nu\to\infty}n\gamma\left((\overline{X}/N)\sum_{i=1}^N W_i^2/X_i-\overline{W}^2\right)$

that satisfy linear regression models. Raj (1954), Murthy (1967), Avadhani and Sukhatme (1970), Avadhani and Srivastava (1972), and Cochran (1977) used the linear relationship between Y_i and X_i to compare different estimators of the mean. However, they did not use a probability distribution for (Y_i, X_i). Subsequently, Rao (2003), Fuller (2011), and Chaudhuri (2014) (see chap. 5), among others, considered the linear relationship between Y_i and X_i and a probability distribution for (Y_i, X_i) to construct different estimators and study their

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

Table 4: Examples of $g(\overline{h})$

Parameter	d	p	h	g
Mean	1	1	$h(y)=y$	$g(s)=s$
Variance	1	2	$h(y)=(y^2, y)$	$g(s_1, s_2)=s_1 - s_2^2$
Correlation coefficient	2	5	$h(z_1, z_2)=(z_1, z_2, z_1^2, z_2^2, z_1 z_2)$	$g(s_1, s_2, s_3, s_4, s_5)=(s_5 - s_1 s_2)/((s_3 - s_1^2)(s_4 - s_2^2))^{1/2}$
Regression coefficient	2	4	$h(z_1, z_2)=(z_1, z_2, z_2^2, z_1 z_2)$	$g(s_1, s_2, s_3, s_4, s_5)=(s_4 - s_1 s_2)/(s_3 - s_2^2)$

behavior. However, to the best of our knowledge, no prior studies have shown how to find asymptotically the most efficient estimator for the mean among a large class of estimators, as we do here. In addition, our study is the first to compare plug-in estimators of the variance, correlation coefficient, and regression coefficient for large samples. Suppose that $\{(Y_i, X_i) : 1 \leq i \leq N\}$ are independently and identically distributed (i.i.d.) random vectors defined on a probability space $(\Omega, \mathbb{F}, \mathbb{P})$. Without any loss of generality, for convenience, we take $\sigma_x^2 = E_{\mathbb{P}}(X_i - E_{\mathbb{P}}(X_i))^2 = 1$. This might require rescaling the variable x. Here, $E_{\mathbb{P}}$ denotes the expectation with respect to the probability measure $\mathbb{P}$. Recall that the population values $X_1, \ldots, X_N$ are used to implement some of the sampling designs. In such a case, we consider a function $P(s, \omega)$ on $\mathcal{S} \times \Omega$ such that $P(s, \cdot)$ is a random variable on Ω for each $s \in \mathcal{S}$, and $P(\cdot, \omega)$ is a probability distribution on $\mathcal{S}$ for each $\omega \in \Omega$ (see Boistard et al. (2017)). Note

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

that $P(s,\omega)$ is the sampling design for any fixed ω in this case. Then, the Δ_j^2 in Table 3 can be expressed in terms of superpopulation moments of $(h(Y_i), X_i)$, from the strong law of large numbers (SLLN), and we can easily compare different classes of estimators in Table 2 under linear models. Let us first state several conditions on the superpopulation distribution $\mathbb{P}$.

C 5. *$X_i \le b$ a.s. $[\mathbb{P}]$ for some $0 < b < \infty$, $E_{\mathbb{P}}(X_i)^{-2} < \infty$, and $\max_{1\le i\le N} X_i / \min_{1\le i\le N} X_i = O(1)$ as $\nu \to \infty$ a.s. $[\mathbb{P}]$. In addition, the support of the distribution of $(h(Y_i), X_i)$ is not a subset of a hyper-plane in $\mathbb{R}^{p+1}$.*

The condition $X_i \le b$ *a.s.* $[\mathbb{P}]$ for some $0 < b < \infty$ in C5 and C0, along with $0 \le \lambda < E_{\mathbb{P}}(X_i)/b$, ensure that $n \max_{1\le i\le N} X_i / \sum_{i=1}^N X_i < 1$ for all sufficiently large ν *a.s.* $[\mathbb{P}]$, which is required to hold for any without-replacement πPS sampling design. On the other hand, the condition, $\max_{1\le i\le N} X_i / \min_{1\le i\le N} X_i = O(1)$ as $\nu \to \infty$ *a.s.* $[\mathbb{P}]$ in C5 implies that C2 holds *a.s.* $[\mathbb{P}]$. Further, C5 ensures that C4 holds *a.s.* $[\mathbb{P}]$ (see Lemma S2 in the Supplementary Material). C5 also ensures that C3 holds under the LMS and any πPS sampling designs *a.s.* $[\mathbb{P}]$ (see Lemma S2 in the Supplementary Material).

Let us first consider the case when $g(\overline{h})$ is the mean of y (see the second row in Table 4). Further, suppose that $Y_i = \alpha + \beta X_i + \epsilon_i$, for $\alpha, \beta \in \mathbb{R}$ and $i = 1, \ldots, N$, where $\{\epsilon_i\}_{i=1}^N$ are i.i.d. random variables and are independent of $\{X_i\}_{i=1}^N$, with $E_{\mathbb{P}}(\epsilon_i) = 0$ and $E_{\mathbb{P}}(\epsilon_i)^4 < \infty$. Then, we have the following theorem.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

Theorem 4. *Suppose that C0 holds, with* $0 \leq \lambda < E_{\mathbb{P}}(X_i)/b$, *and C5 holds. Then, a.s.* $[\mathbb{P}]$, *the PEML estimator under SRSWOR and the LMS sampling design has the lowest asymptotic MSE among all estimators of the population mean under different sampling designs considered here.*

Remark 3. *Note that for SRSWOR, the PEML estimator of the population mean has the lowest asymptotic MSE among all estimators considered here a.s.* $[\mathbb{P}]$ *when C0 holds with* $0 \leq \lambda < 1$ *and C5 holds (see the proof of Theorem 4).*

Theorem 5. *Suppose that C0 holds with* $0 \leq \lambda < E_{\mathbb{P}}(X_i)/b$, *and C5 holds. Then, a.s.* $[\mathbb{P}]$, *the performance of the PEML estimator of the population mean under the RHC and any HEπPS sampling designs, which use auxiliary information, is worse than its performance under SRSWOR.*

Recall from the introduction that for the variance, correlation coefficient, and regression coefficient, we compare only those equivalence classes that contain plug-in estimators based on the Hájek and PEML estimators. We first state the following condition.

C 6. $\xi > 2\max\{\mu_1, \mu_{-1}/(\mu_1\mu_{-1} - 1)\}$, *where* $\xi = \mu_3 - \mu_2\mu_1$ *is the covariance between* X_i^2 *and* X_i, *and* $\mu_j = E_{\mathbb{P}}(X_i)^j$, *for* $j = -1, 1, 2, 3$.

The above condition is used to prove part (ii) in each of Theorems 6 and 7. This condition holds when X_i follows a well-known distribution, such as the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

gamma (with shape parameter value larger than one and any scale parameter value), beta (with the second shape parameter value greater than the first shape parameter value, and the first shape parameter value larger than one), Pareto (with shape parameter value lying in the interval $(3, (5+\sqrt{17})/2)$ and any scale parameter value), log-normal (with both the parameters taking any value), and Weibull (with shape parameter value lying in the interval $(1, 3.6)$ and any scale parameter value). Now, consider the case when $g(\overline{h})$ is the variance of y (see the third row in Table 4). Recall the linear model $Y_i{=}\alpha + \beta X_i + \epsilon_i$ from above, and assume that $E_{\mathbb{P}}(\epsilon_i)^8 < \infty$. Then, we have the following theorem.

Theorem 6. *(i) Let us first consider SRSWOR and the LMS sampling design, and suppose that C0 and C5 hold. Then, a.s.* $[\mathbb{P}]$*, the plug-in estimator of the population variance based on the PEML estimator has the lowest asymptotic MSE among all estimators considered here.*

(ii) Next, consider any HEπPS sampling design, and suppose that C0 holds with $0 \leq \lambda < E_{\mathbb{P}}(X_i)/b$*, and C5 and C6 hold. Then, a.s.* $[\mathbb{P}]$*, the plug-in estimator of the population variance based on the Hájek estimator has the lowest asymptotic MSE among all estimators considered here.*

Now, suppose that $y{=}(z_1, z_2) \in \mathbb{R}^2$, and consider the case when $g(\overline{h})$ is the correlation coefficient between z_1 and z_2 (see the fourth row in Table 4). We also consider the case when $g(\overline{h})$ is the regression coefficient of z_1 on z_2

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

(see the fifth row in Table 4). Further, suppose that $Y_i{=}\alpha + \beta X_i + \epsilon_i$ for $Y_i{=}(Z_{1i}, Z_{2i})$, $\alpha, \beta \in \mathbb{R}^2$ and $i{=}1, \ldots, N$, where $\{\epsilon_i\}_{i=1}^N$ are i.i.d. random vectors in $\mathbb{R}^2$ independent of $\{X_i\}_{i=1}^N$ with $E_{\mathbb{P}}(\epsilon_i){=}0$ and $E_{\mathbb{P}}||\epsilon_i||^8 < \infty$. Then, we have the following theorem.

Theorem 7. (*i*) *Let us first consider SRSWOR and the LMS sampling design, and suppose that C0 and C5 hold. Then, a.s.* $[\mathbb{P}]$, *the plug-in estimator of each of the correlation and the regression coefficients in the population based on the PEML estimator has the lowest asymptotic MSE among all estimators considered here.*

(*ii*) *Next, consider any HE*π*PS sampling design, and suppose that C0 holds with* $0 \leq \lambda < E_{\mathbb{P}}(X_i)/b$, *and C5 and C6 hold. Then, a.s.* $[\mathbb{P}]$, *the plug-in estimator of each of the above parameters based on the Hájek estimator has the lowest asymptotic MSE among all estimators considered here.*

4. Data analysis

In this section, we empirically compare the estimators of the mean, variance, correlation coefficient, and regression coefficient using real and synthetic data. Note that for the empirical comparison, we exclude some of the estimators considered in the theoretical comparison, for the following reasons:

(i) Because the GREG estimator is well-known to be asymptotically better

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

than the HT, ratio, and product estimators under SRSWOR (see Cochran (1977)), we exclude these estimators under SRSWOR.

(ii) Because the MSEs of the estimators under the LMS sampling design become very close to the MSEs of the same estimators under SRSWOR, as expected from Theorem 1, we do not report these results under the LMS sampling design. Moreover, SRSWOR is a simpler and more commonly used sampling design than is the LMS sampling design.

Thus, we consider the estimators in Table 5 for the empirical comparison. Recall from Table 1 that the HT, ratio, and product estimators of the mean coincide under any HEπPS sampling design. We draw I=1000 samples, each of sizes n=75, 100, and 125, using the sampling designs in Table 5. We use the software R to draw the samples and compute the various estimators. For the RS sampling design, we use the "pps" package in R, and for the PEML estimator, we use the R code in Wu (2005). We compare the two estimators $g(\hat{\overline{h}}_1)$ and $g(\hat{\overline{h}}_2)$ of $g(\overline{h})$ empirically under the sampling designs $P_1(s)$ and $P_2(s)$, respectively, in terms of their relative efficiency, defined as

$$RE(g(\hat{\overline{h}}_1), P_1 | g(\hat{\overline{h}}_2), P_2) = MSE_{P_2}(g(\hat{\overline{h}}_2))/MSE_{P_1}(g(\hat{\overline{h}}_1)),$$

where $MSE_{P_j}(g(\hat{\overline{h}}_j))$=$I^{-1}\sum_{l=1}^{I}(g(\hat{\overline{h}}_{jl}) - g(\overline{h}_0))^2$ is the empirical mean squared error of $g(\hat{\overline{h}}_j)$ under $P_j(s)$, for j=1, 2. Here, $\hat{\overline{h}}_{jl}$ is the estimate of $\overline{h}$ based on the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

Table 5: Estimators considered for the empirical comparison

Parameters	Estimators
Mean	GREG and PEML estimators under SRSWOR; HT, Hájek, GREG and PEML estimators under [3]RS sampling design; and RHC, GREG and PEML estimators under RHC sampling design
Variance, correlation coefficient and regression coefficient	Obtained by plugging in Hájek and PEML estimators under SRSWOR and[1]RS sampling design, and PEML estimator under RHC sampling design

[3] We consider the RS sampling design, because it is a HEπPS sampling design, and it is easier to implement than other HEπPS sampling designs.

jth estimator and the lth sample, and $g(\overline{h}_0)$ is the true value of the parameter $g(\overline{h})$, for j=1, 2, l=1, ..., I. Here, $g(\hat{\overline{h}}_1)$ under $P_1(s)$ is more efficient than $g(\hat{\overline{h}}_2)$ under $P_2(s)$ if $RE(g(\hat{\overline{h}}_1), P_1|g(\hat{\overline{h}}_2), P_2) > 1$.

Next, for each of the parameters considered in this section, we compare the average lengths of the asymptotically 95% confidence intervals (CIs) constructed using the various estimators. In order to construct asymptotically 95% CIs, we need an estimator of the asymptotic MSE of $\sqrt{n}(g(\hat{\overline{h}})-g(\overline{h}))$. If we consider SR-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

SWOR or the RS sampling design, it follows from the proofs of Theorems 1 and 2 that the asymptotic MSE of $\sqrt{n}(g(\hat{\overline{h}})-g(\overline{h}))$ is $\tilde{\Delta}_1^2=\lim_{\nu\to\infty} nN^{-2}\nabla g(\overline{h})\sum_{i=1}^N(\mathbf{V}_i-\mathbf{T}\pi_i)^T(\mathbf{V}_i-\mathbf{T}\pi_i)(\pi_i^{-1}-1)\nabla g(\overline{h})^T$, where $\mathbf{T}=\sum_{i=1}^N \mathbf{V}_i(1-\pi_i)/\sum_{i=1}^N \pi_i(1-\pi_i)$. Moreover, $\mathbf{V}_i$ is h_i or $h_i-\overline{h}$ or $h_i-\overline{h}-S_{xh}(X_i-\overline{X})/S_x^2$ if $\hat{\overline{h}}$ is $\hat{\overline{h}}_{HT}$ or $\hat{\overline{h}}_H$ or $\hat{\overline{h}}_{PEML}$ (as well as $\hat{\overline{h}}_{GREG}$), respectively, with $d(i,s)=(N\pi_i)^{-1}$. Recall that $S_{xh}=\sum_{i=1}^N X_ih_i/N-\overline{X}\,\overline{h}$. Following Cardot et al. (2014), we estimate $\tilde{\Delta}_1^2$ by

$$\hat{\Delta}_1^2 = nN^{-2}\nabla g(\hat{\overline{h}})\sum_{i\in s}(\hat{\mathbf{V}}_i-\hat{\mathbf{T}}\pi_i)^T(\hat{\mathbf{V}}_i-\hat{\mathbf{T}}\pi_i)(\pi_i^{-1}-1)\pi_i^{-1}\nabla g(\hat{\overline{h}})^T, \qquad (4.1)$$

where $\hat{\mathbf{T}}=\sum_{i\in s}\hat{\mathbf{V}}_i(\pi_i^{-1}-1)/\sum_{i\in s}(1-\pi_i)$, $\hat{\overline{h}}=\hat{\overline{h}}_{HT}$ in the case of the mean, variance, and regression coefficient, and $\hat{\overline{h}}=\hat{\overline{h}}_H$ in the case of the correlation coefficient. Here, $\hat{\mathbf{V}}_i$ is h_i or $h_i-\hat{\overline{h}}_{HT}$ or $h_i-\hat{\overline{h}}_{HT}-\hat{S}_{xh,1}(X_i-\hat{\overline{X}}_{HT})/\hat{S}_{x,1}^2$ if $\hat{\overline{h}}$ is $\hat{\overline{h}}_{HT}$ or $\hat{\overline{h}}_H$ or $\hat{\overline{h}}_{PEML}$ (as well as $\hat{\overline{h}}_{GREG}$), respectively, with $d(i,s)=(N\pi_i)^{-1}$. Further, $\hat{S}_{xh,1}=\sum_{i\in s}(N\pi_i)^{-1}X_ih_i-\hat{\overline{X}}_{HT}\hat{\overline{h}}_{HT}$ and $\hat{S}_{x,1}^2=\sum_{i\in s}(N\pi_i)^{-1}X_i^2-\hat{\overline{X}}_{HT}^2$. We estimate $\overline{h}$ in $\nabla g(\overline{h})$ using $\hat{\overline{h}}_{HT}$ in the case of the mean, variance, and regression coefficient, because $\hat{\overline{h}}_{HT}$ is an unbiased estimator, and it is easier to compute than the other estimators of $\overline{h}$ considered here. On the other hand, some estimators of the correlation coefficient may be undefined if we estimate $\overline{h}$ using any estimator other than $\hat{\overline{h}}_H$ or $\hat{\overline{h}}_{PEML}$ (see the introduction). In this case, we choose $\hat{\overline{h}}_H$, because it is easier to compute than $\hat{\overline{h}}_{PEML}$.

Next, if we consider the RHC sampling design, it follows from the proof of Theorem 2 that the asymptotic MSE of $\sqrt{n}(g(\overline{h})-g(\hat{\overline{h}}))$ is $\tilde{\Delta}_2^2=\lim_{\nu\to\infty} n\gamma\overline{X}N^{-1}\times$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

$\nabla g(\overline{h}) \sum_{i=1}^{N} (\mathbf{V}_i - X_i\overline{\mathbf{V}}/\overline{X})^T(\mathbf{V}_i - X_i\overline{\mathbf{V}}/\overline{X})X_i^{-1}\nabla g(\overline{h})^T$, where γ and $\overline{\mathbf{V}}$ are as in the paragraph following C2. Moreover, $\mathbf{V}_i$ is h_i or $h_i - \overline{h} - S_{xh}(X_i - \overline{X})/S_x^2$ if $\hat{\overline{h}}$ is $\hat{\overline{h}}_{RHC}$ or $\hat{\overline{h}}_{PEML}$ (as well as $\hat{\overline{h}}_{GREG}$), respectively, with $d(i,s)$=G_i/NX_i. We estimate $\tilde{\Delta}_2^2$ by

$$\hat{\Delta}_2^2 = n\gamma\overline{X}N^{-1}\nabla g(\hat{\overline{h}}) \sum_{i\in s} (\hat{\mathbf{V}}_i - X_i\hat{\overline{\mathbf{V}}}_{RHC}/\overline{X})\times \\ (\hat{\mathbf{V}}_i - X_i\hat{\overline{\mathbf{V}}}_{RHC}/\overline{X})(G_iX_i^{-2})\nabla g(\hat{\overline{h}})^T, \tag{4.2}$$

where $\hat{\overline{\mathbf{V}}}_{RHC}$=$\sum_{i\in s} \hat{\mathbf{V}}_i G_i/NX_i$, $\hat{\overline{h}}$=$\hat{\overline{h}}_{RHC}$ in the case of the mean, variance, and regression coefficient, and $\hat{\overline{h}}$=$\hat{\overline{h}}_{PEML}$ in the case of the correlation coefficient. Here, $\hat{\mathbf{V}}_i$ is h_i or $h_i - \hat{\overline{h}}_{RHC} - \hat{S}_{xh,2}(X_i - \overline{X})/\hat{S}_{x,2}^2$ if $\hat{\overline{h}}$ is $\hat{\overline{h}}_{RHC}$ or $\hat{\overline{h}}_{PEML}$ (as well as $\hat{\overline{h}}_{GREG}$), respectively, with $d(i,s)$=G_i/NX_i. Further, $\hat{S}_{xh,2}$=$\sum_{i\in s} h_iG_i/N - \overline{X}$ $\hat{\overline{h}}_{RHC}$ and $\hat{S}_{x,1}^2$=$\sum_{i\in s} X_iG_i/N - \overline{X}^2$. In the case of the mean, variance, and regression coefficient, we estimate $\overline{h}$ in $\nabla g(\overline{h})$ using $\hat{\overline{h}}_{RHC}$ for the same reason that we estimate $\overline{h}$ using $\hat{\overline{h}}_{HT}$ under SRSWOR and the RS sampling design. On the other hand, in the case of the correlation coefficient, we estimate $\overline{h}$ in $\nabla g(\overline{h})$ using $\hat{\overline{h}}_{PEML}$ under the RHC sampling design so that the estimator of the correlation coefficient in the expression of $\nabla g(\overline{h})$ in this case is well defined.

We draw I=1000 samples, each of sizes n=75, 100, and 125, using the sampling designs in Table 5. Then, for each of the parameters, sampling designs, and estimators, we construct I asymptotically 95% CIs based on these samples, and compute the average and the standard deviation of their lengths.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

4.1 Analysis based on synthetic data

In this section, we consider the population values $\{(Y_i, X_i) : 1 \leq i \leq N\}$ on (y, x) generated from a linear model, as follows. We choose N=5000 and generate the X_i from a gamma distribution with mean 1000 and standard deviation (s.d.) 200. Then, Y_i is generated from the linear model Y_i=$500 + X_i + \epsilon_i$, for i=$1, \ldots, N$, where ϵ_i is generated independently of $\{X_i\}_{i=1}^N$ from a normal distribution with mean zero and s.d. 100. We also generate the population values $\{(Y_i, X_i) : 1 \leq i \leq N\}$ from a linear model in which y=(z_1, z_2) is a bivariate study variable. The population values $\{X_i\}_{i=1}^N$ are generated in the same way as in the earlier case. Then, Y_i=(Z_{1i}, Z_{2i}) is generated from the linear model Z_{ji}=$\alpha_j + X_i + \epsilon_{ji}$, for i=$1, \ldots, N$, where α_1=500 and α_2=1000. The ϵ_{1i} are generated independently of the X_i from a normal distribution with mean zero and s.d. 100, and the ϵ_{2i} are generated independently of the X_i and the ϵ_{1i} from a normal distribution with mean zero and s.d. 200. We consider the estimation of the mean and the variance of y for the first data set and the correlation and the regression coefficients between z_1 and z_2 for the second data set.

The results of the empirical comparison based on synthetic data are summarized as follows. For each of the mean, variance, correlation coefficient, and regression coefficient, the plug-in estimator based on the PEML estimator under SRSWOR is more efficient than any other estimator under any other sampling

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

design (see Tables 2 through 6 in the Supplementary Material) considered in Table 5. In addition, for each of the above parameters, the asymptotically 95% CI based on the PEML estimator under SRSWOR has the least average length (see Tables 7 through 11 in the Supplementary Material). Thus, the empirical results stated here corroborate the theoretical results stated in Theorems 4 through 7.

4.2 Analysis based on real data

In this section, we consider a data set on village amenities in the state of West Bengal in India obtained from the Office of the Registrar General & Census Commissioner, India (*https://censusindia.gov.in*). The relevant study variables for this data set are described in Table 6. We consider the following estimation problems for a population of 37478 villages. For these estimation problems, we use the number of people living in village x as the size variable.

Table 6: Description of study variables

y_1	Number of primary schools in village
y_2	Scheduled castes population size in village
y_3	Number of secondary schools in village
y_4	Scheduled tribes population size in village

(i) First, we estimate the mean and variance of each of y_1 and y_2. The scatter

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

plot and the least square regression line in Figure 1 in the Supplementary Material show that y_1 and x have an approximately linear relationship. In addition, the correlation coefficient between y_1 and x is 0.72. On the other hand, y_2 and x do not seem to have a linear relationship (see the scatter plot and the least square regression line in Figure 2 in the Supplementary Material).

(ii) Next, we estimate the correlation and regression coefficients of y_1 and y_3, and of y_2 and y_4. The scatter plot and least square regression line in Figure 3 in the Supplementary Material show that y_3 does not seem to be dependent on x. Further, we see from the scatter plot and the least square regression line of y_4 and x (see Figure 4 in the Supplementary Material) that y_4 and x do not seem to have a linear relationship.

The results of the empirical comparison based on real data are summarized in Table 7. For further details, see Tables 12 through 31 in the Supplementary Material. The approximate linear relationship between y_1 and x (see the scatter plot and the least square regression line in Figure 1 in the Supplementary Material) could be a possible reason why the plug-in estimator based on the PEML estimator under SRSWOR is the most efficient for the mean and variance of y_1. Furthermore, possibly for the same reason, the plug-in estimators of the correlation and regression coefficients between y_1 and y_3 based on the PEML

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

Table 7: Most efficient estimators, in terms of relative efficiency (it follows from Tables 22 through 31 in the Supplementary Material that the asymptotically 95% CIs based on the most efficient estimators have the least average lengths).

Parameters	Most efficient estimators
Mean and variance of y_1	The plug-in estimator based on the the PEML estimator under SRSWOR
Mean of y_2	The HT estimator under RS sampling design
Variance of y_2	The plug-in estimator based on the Hájek estimator under RS sampling design
Correlation and regression coefficients of y_1 and y_3	The plug-in estimator based on the PEML estimator under SRSWOR
Correlation and regression coefficients of y_2 and y_4	The plug-in estimator based on the Hájek estimator under RS sampling design

estimator under SRSWOR are the most efficient.

On the other hand, y_2 and y_4 do not seem to have a linear relationship with x (see the scatter plots and the least square regression lines in Figures 2 and 4 in the Supplementary Material). Possibly for this reason, the plug-in estimators of the parameters related to y_2 and y_4 based on the PEML estimator are not able to outperform the plug-in estimators of the same parameters based on the HT and Hájek estimators. Next, we observe that there is substan-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

tial correlation between y_2 and x (correlation coefficient=0.67), and y_4 and x (correlation coefficient=0.25). Possibly because of this, under the RS sampling design, which uses auxiliary information, the plug-in estimators of the parameters related to y_2 and y_4 based on the HT and Hájek estimators are the most efficient.

5. Conclusion

It follows from Theorem 4 that the PEML estimator of the mean under SRSWOR becomes asymptotically either more efficient than, or equivalent to any other estimator under any other sampling design considered here. It also follows from Theorems 1 and 2 that the GREG estimator of the mean is asymptotically equivalent to the PEML estimator under the sampling designs considered here. However, our numerical studies based on finite samples indicate that the PEML estimator of the mean performs slightly better than the GREG estimator under all the sampling designs considered in Section 4 (see Tables 2, 12, and 14 in the Supplementary Material). Moreover, if the estimators of the variance, correlation coefficient, and regression coefficient are constructed by plugging in the GREG estimator of the mean, then the estimators of the population variances in these parameters may become negative. On the other hand, if the estimators of these parameters are constructed by plugging in the PEML estimator of

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

the mean, then such a problem does not occur. Further, for these parameters, the plug-in estimators based on either the PEML or the Hájek estimator are asymptotically best, depending on the sampling design (see Theorems 6 and 7).

We see from Theorem 4 that for the population mean, the PEML estimator, which is not design unbiased, outperforms design unbiased estimators such as the HT and RHC estimators. Further, the plug-in estimators of the population variance based on the HT and RHC estimators may become negative. This affects the plug-in estimators of the correlation and regression coefficients based on the HT and RHC estimators.

It follows from Table 2 that under the LMS sampling design, the large-sample performance of the estimators of the functions of means considered here is the same as that under SRSWOR. The LMS sampling design was introduced to make the ratio estimator of the mean unbiased. It follows from Remark 2 in Section 2 that the performance of the ratio estimator of the mean is worse than that of several other estimators, even under the LMS sampling design.

The coefficient of variation is another well-known finite population parameter, and can be expressed as a function of the population mean $g(\overline{h})$. We have d=1, p=2, $h(y)$=(y^2, y), and $g(s_1, s_2)$=$\sqrt{s_1 - s_2^2}/s_2$ in this case. Of the estimators considered here, the plug-in estimators of $g(\overline{h})$ based on the PEML and Hájek estimators of the mean can be used to estimate this parameter, because

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

it involves the finite population variance. We have omitted a comparison of the estimators of the coefficient of variation, owing to the complexity of the mathematical expressions. However, the asymptotic results stated in Theorems 6 and 7 hold for this parameter as well.

An empirical comparison of the biased estimators considered here and their bias-corrected versions is performed using jackknifing in Section S4 in the Supplementary Material. It follows from this comparison that for all the parameters considered here, the bias-corrected estimators become worse than the original biased estimators for both the synthetic and the real data. This is because, although bias-correction reduces the bias in the original estimators, it causes the variances of these estimators to increase substantially.

Supplementary Material

In the online Supplementary Material, we discuss some conditions from the main paper, and describe situations in which these conditions hold. Then, we state and prove some additional mathematical results. We also give proofs for Remark 1 and Theorems 2, 3, 6, and 7. Furthermore, we compare the biased estimators considered in this paper empirically, with their bias-corrected versions based on jackknifing in terms of the MSE. Finally, we provide numerical results related to the analyses of the synthetic and real data in Section 4.

Acknowledgments

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

The authors gratefully acknowledge the constructive comments and suggestions provided by the reviewers and the associate editor. The authors also thank Prof. Aloke Kar and Prof. Sandip Mitra for several discussions about Section 4.2 of the paper.

Appendix

Let us begin by providing the expressions (see Table 8 below) of those estimators of $\overline{Y}$, which are considered in this paper. In Table 8, $\{\pi_i\}_{i=1}^{N}$ denote inclusion probabilities, and G_i is the total of the x values of that randomly formed group from which the i^{th} population unit is selected in the sample by RHC sampling design (cf. Chaudhuri et al. (2006)). In the case of the GREG estimator, $\hat{\overline{Y}}_*=\sum_{i\in s} d(i,s)Y_i/\sum_{i\in s} d(i,s)$, $\hat{\overline{X}}_*=\sum_{i\in s} d(i,s)\times X_i/\sum_{i\in s} d(i,s)$ and

Table 8: Estimators of $\overline{Y}$

Estimator	Expression
HT	$\hat{\overline{Y}}_{HT}=\sum_{i\in s}(N\pi_i)^{-1}Y_i$
RHC	$\hat{\overline{Y}}_{RHC}=\sum_{i\in s} G_iY_i/NX_i$
Hájek	$\hat{\overline{Y}}_{H}=\sum_{i\in s} \pi_i^{-1}Y_i/\sum_{i\in s} \pi_i^{-1}$
Ratio	$\hat{\overline{Y}}_{RA}=(\sum_{i\in s} \pi_i^{-1}Y_i/\sum_{i\in s} \pi_i^{-1}X_i)\overline{X}$
Product	$\hat{\overline{Y}}_{PR}=\sum_{i\in s}(N\pi_i)^{-1}Y_i\sum_{i\in s}(N\pi_i)^{-1}X_i/\ \overline{X}$
GREG	$\hat{\overline{Y}}_{GREG}=\hat{\overline{Y}}_{*}+\hat{\beta}(\overline{X}-\hat{\overline{X}}_{*})$
PEML	$\hat{\overline{Y}}_{PEML}=\sum_{i\in s} c_iY_i$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

$\hat{\beta}=\sum_{i\in s} d(i,s)(Y_i-\hat{\overline{Y}}_*)(X_i-\hat{\overline{X}}_*)/\sum_{i\in s} d(i,s)(X_i-\hat{\overline{X}}_*)^2$, where $\{d(i,s): i\in s\}$ are sampling design weights. Finally, the c_i's (>0) in the PEML estimator are obtained by maximizing $\sum_{i\in s} d(i,s)\log(c_i)$ subject to $\sum_{i\in s} c_i=1$ and $\sum_{i\in s} c_i(X_i-\overline{X})=0$. Following Chen and Sitter (1999), we consider both the GREG and the PEML estimators with $d(i,s)=(N\pi_i)^{-1}$ under SRSWOR, LMS sampling design and any HEπPS sampling design, and with $d(i,s)=G_i/NX_i$ under RHC sampling design.

Let us denote the HT, the RHC, the Hájek, the ratio, the product, the GREG and the PEML estimators of population means of $h(y)$ by $\hat{\overline{h}}_{HT}$, $\hat{\overline{h}}_{RHC}$, $\hat{\overline{h}}_{H}$, $\hat{\overline{h}}_{RA}$, $\hat{\overline{h}}_{PR}$, $\hat{\overline{h}}_{GREG}$ and $\hat{\overline{h}}_{PEML}$, respectively. Now, we give the proofs of Theorems 1, 4 and 5. The proofs of Remark 1 and Theorems 2, 3, 6 and 7 are given in Section S3 of the supplement.

Proof of Theorem 1. Let us consider SRSWOR and LMS sampling design. It follows from (i) in Lemma S6 in the supplement that $\sqrt{n}(\hat{\overline{h}}-\overline{h})\xrightarrow{\mathcal{L}} N(0,\Gamma)$ as $\nu\to\infty$ for some p.d. matrix Γ, when $\hat{\overline{h}}$ is one of $\hat{\overline{h}}_{HT}$, $\hat{\overline{h}}_{H}$, $\hat{\overline{h}}_{RA}$, $\hat{\overline{h}}_{PR}$, and $\hat{\overline{h}}_{GREG}$ with $d(i,s)=(N\pi_i)^{-1}$ under any of these sampling designs. Now, note that $\max_{i\in s}|X_i-\overline{X}|=o_p(\sqrt{n})$, and $\sum_{i\in s}\pi_i^{-1}(X_i-\overline{X})/\sum_{i\in s}\pi_i^{-1}(X_i-\overline{X})^2=O_p(1/\sqrt{n})$ as $\nu\to\infty$ under the above sampling designs (see Lemma S8 in the supplement). Then, by applying Theorem 1 of Chen and Sitter (1999) to each real-valued coordinate of $\hat{\overline{h}}_{PEML}$ and $\hat{\overline{h}}_{GREG}$, we get $\sqrt{n}(\hat{\overline{h}}_{PEML}-\hat{\overline{h}}_{GREG})=o_p(1)$ as $\nu\to\infty$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

for $d(i,s)=(N\pi_i)^{-1}$ under these sampling designs. This implies that $\hat{\overline{h}}_{PEML}$ and $\hat{\overline{h}}_{GREG}$ with $d(i,s)=(N\pi_i)^{-1}$ have the same asymptotic distribution. Therefore, if $\hat{\overline{h}}$ is one of $\hat{\overline{h}}_{HT}$, $\hat{\overline{h}}_{H}$, $\hat{\overline{h}}_{RA}$, $\hat{\overline{h}}_{PR}$, and $\hat{\overline{h}}_{GREG}$ and $\hat{\overline{h}}_{PEML}$ with $d(i,s)=(N\pi_i)^{-1}$, we have

$$\sqrt{n}(g(\hat{\overline{h}})-g(\overline{h})) \xrightarrow{\mathcal{L}} N(0,\Delta^2) \text{ as } \nu\to\infty \tag{5.1}$$

under any of the above-mentioned sampling designs for some $\Delta^2>0$ by the delta method and the condition $\nabla g(\mu_0)\neq 0$ at $\mu_0=\lim_{\nu\to\infty}\overline{h}$. It can be shown from the proof of (i) in Lemma S6 in the supplement that $\Delta^2=\nabla g(\mu_0)\Gamma_1(\nabla g(\mu_0))^T$, where $\Gamma_1=\lim_{\nu\to\infty} nN^{-2}\sum_{i=1}^{N}(\mathbf{V}_i-\mathbf{T}\pi_i)^T(\mathbf{V}_i-\mathbf{T}\pi_i)(\pi_i^{-1}-1)$. It can also be shown from Table 1 in the supplement that under each of the above sampling designs, $\mathbf{V}_i$ in Γ_1 is h_i or $h_i-\overline{h}$ or $h_i-\overline{h}X_i/\overline{X}$ or $h_i+\overline{h}X_i/\overline{X}$ or $h_i-\overline{h}-S_{xh}(X_i-\overline{X})/S_x^2$ if $\hat{\overline{h}}$ is $\hat{\overline{h}}_{HT}$ or $\hat{\overline{h}}_{H}$ or $\hat{\overline{h}}_{RA}$ or $\hat{\overline{h}}_{PR}$, or $\hat{\overline{h}}_{GREG}$ with $d(i,s)=(N\pi_i)^{-1}$, respectively.

Now, by (i) in Lemma S7 in the supplement, we have

$$\sigma_1^2=\sigma_2^2=(1-\lambda)\lim_{\nu\to\infty}\sum_{i=1}^{N}(A_i-\bar{A})^2/N. \tag{5.2}$$

where σ_1^2 and σ_2^2 are as defined in the statement of Lemma S7, and $A_i=\nabla g(\mu_0)\mathbf{V}_i^T$ for different choices of $\mathbf{V}_i$ mentioned in the preceding paragraph. Note that $g(\hat{\overline{h}}_{GREG})$ and $g(\hat{\overline{h}}_{PEML})$ have the same asymptotic distribution under each of SRSWOR and LMS sampling design since $\sqrt{n}(\hat{\overline{h}}_{PEML}-\hat{\overline{h}}_{GREG})=o_p(1)$ for

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

$\nu \to \infty$ under these sampling designs as pointed out earlier in this proof. Further, (5.2) implies that $g(\hat{\bar{h}}_{GREG})$ with $d(i,s)=(N\pi_i)^{-1}$ has the same asymptotic MSE under SRSWOR and LMS sampling design. Thus $g(\hat{\bar{h}}_{GREG})$ and $g(\hat{\bar{h}}_{PEML})$ with $d(i,s)=(N\pi_i)^{-1}$ under SRSWOR and LMS sampling design form class 1 in Table 2.

Next, (5.2) yields that $g(\hat{\bar{h}}_{HT})$ has the same asymptotic MSE under SRSWOR and LMS sampling design. It also follows from (5.2) that $g(\hat{\bar{h}}_{H})$ has the same asymptotic MSE under SRSWOR and LMS sampling design. Now, note that $g(\hat{\bar{h}}_{HT})$ and $g(\hat{\bar{h}}_{H})$ coincide under SRSWOR. Thus $g(\hat{\bar{h}}_{HT})$ under SRSWOR, and $g(\hat{\bar{h}}_{HT})$ and $g(\hat{\bar{h}}_{H})$ under LMS sampling design form class 2 in Table 2.

Next, (5.2) implies that $g(\hat{\bar{h}}_{RA})$ has the same asymptotic MSE under SRSWOR and LMS sampling design. Further, (5.2) implies that $g(\hat{\bar{h}}_{PR})$ has the same asymptotic MSE under SRSWOR and LMS sampling design. Thus $g(\hat{\bar{h}}_{RA})$ under SRSWOR and LMS sampling design forms class 3 in Table 2, and $g(\hat{\bar{h}}_{PR})$ under those sampling designs forms class 4 in Table 2. This completes the proof of Theorem 1. □

Proof of Theorem 4. Note that C1 and C2 hold *a.s.* $[\mathbb{P}]$ since C5 holds and $E_{\mathbb{P}}(\epsilon_i)^4 < \infty$. Also, note that C3 holds *a.s.* $[\mathbb{P}]$ under SRSWOR and LMS sampling design (see Lemma S2 in the supplement). Then, under the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

above sampling designs, conclusions of Theorems 1 and 3 hold *a.s.* $[\mathbb{P}]$ for $d{=}p{=}1$, $h(y){=}y$ and $g(s){=}s$. Note that $W_i{=}\nabla g(\overline{h})h_i^T{=}Y_i$. Also, note that the Δ_i^2's in Table 3 can be expressed in terms of superpopulation moments of (Y_i, X_i) *a.s.* $[\mathbb{P}]$ by SLLN since $E_{\mathbb{P}}(\epsilon_i)^4 < \infty$. Recall from the beginning of Section 3 that we have taken $\sigma_x^2{=}1$. Then, we have $\Delta_2^2 - \Delta_1^2{=}(1-\lambda)\sigma_{xy}^2$, $\Delta_3^2 - \Delta_1^2{=}(1-\lambda)(\sigma_{xy} - E_{\mathbb{P}}(Y_i)/\mu_1)^2$ and $\Delta_4^2 - \Delta_1^2{=}(1-\lambda)(\sigma_{xy} + E_{\mathbb{P}}(Y_i)/\mu_1)^2$ *a.s.* $[\mathbb{P}]$, where $\mu_1{=}E_{\mathbb{P}}(X_i)$ and $\sigma_{xy}{=}cov_{\mathbb{P}}(X_i, Y_i)$. Hence, $\Delta_1^2 < \Delta_i^2$ *a.s.* $[\mathbb{P}]$ for $i{=}2, 3, 4$.

Next consider the case of $0 \leq \lambda < E_{\mathbb{P}}(X_i)/b$. Note that $n\gamma \to c$ as $\nu \to \infty$ for some $c \geq 1-\lambda$ by Lemma S1 in the supplement. Also, note that *a.s.* $[\mathbb{P}]$, C4 holds in the case of RHC sampling design and C3 holds in the case of any HEπPS sampling design (see Lemma S2 in the supplement). Then, under RHC and any HEπPS sampling designs, conclusions of Theorems 2 and 3 hold *a.s.* $[\mathbb{P}]$ for $d{=}p{=}1$, $h(y){=}y$ and $g(s){=}s$. Further, we have $\Delta_5^2 - \Delta_1^2{=}\big\{E_{\mathbb{P}}\big(Y_i - E_{\mathbb{P}}(Y_i)\big)^2\big(\mu_1/X_i - \lambda\big) - \mu_1^2\sigma_{xy}\big(\sigma_{xy}cov_{\mathbb{P}}(X_i, 1/X_i) - 2cov_{\mathbb{P}}(Y_i, 1/X_i)\big) + \lambda\sigma_{xy}^2\big\} - (1-\lambda)\big\{\sigma_y^2 - \sigma_{xy}^2\big\}$, $\Delta_6^2 - \Delta_5^2{=}\ E_{\mathbb{P}}\big(Y_i^2\big(\mu_1/X_i - \lambda\big)\big) - \big\{\lambda E_{\mathbb{P}}(Y_iX_i) - E_{\mathbb{P}}(Y_i)\mu_1\big\}^2/\chi\mu_1 - \big\{E_{\mathbb{P}}\big(Y_i - E_{\mathbb{P}}(Y_i) - \sigma_{xy}(X_i - \mu_1)\big)^2\big(\mu_1/X_i - \lambda\big)\big\}$, $\Delta_7^2 - \Delta_5^2{=}\big\{\mu_1^2\sigma_{xy}\big(\sigma_{xy}cov_{\mathbb{P}}(X_i, 1/X_i) - 2cov_{\mathbb{P}}(Y_i, 1/X_i)\big) - \lambda\sigma_{xy}^2 - \lambda^2\sigma_{xy}^2/\mu_1\chi\big\}$, $\Delta_8^2 - \Delta_1^2{=}c\big\{\mu_1 E_{\mathbb{P}}(Y_i - E_{\mathbb{P}}(Y_i))^2/X_i - \mu_1^2\sigma_{xy}(\sigma_{xy}cov_{\mathbb{P}}(X_i, 1/X_i) - 2cov_{\mathbb{P}}(Y_i, 1/X_i))\big\} - (1-\lambda)\big\{\sigma_y^2 - \sigma_{xy}^2\big\}$ and $\Delta_9^2 - \Delta_1^2{=}c\big\{\mu_1 E_{\mathbb{P}}(Y_i^2/X_i) - E_{\mathbb{P}}^2(Y_i)\big\} - (1-\lambda)\big\{\sigma_y^2 - \sigma_{xy}^2\big\}$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

a.s. $[\mathbb{P}]$, where $\sigma_y^2{=}var_{\mathbb{P}}(Y_i)$, $\chi{=}\mu_1 - \lambda(\mu_2/\mu_1)$ and $\mu_2{=}E_{\mathbb{P}}(X_i)^2$. Here, we note that $\chi{=}E_{\mathbb{P}}\big(X_i^2(\mu_1/X_i - \lambda)\big)/\mu_1 > 0$ because C5 holds and C0 holds with $0 \leq \lambda < E_{\mathbb{P}}(X_i)/b$. Moreover, from the linear model set up, we can show that $\Delta_5^2 - \Delta_1^2{=}\sigma^2(\mu_1\mu_{-1} - 1) > 0$, $\Delta_6^2 - \Delta_5^2{=}E_{\mathbb{P}}\big\{(\alpha + \beta X_i) - \chi^{-1}X_i(\alpha + \beta\mu_1 - \lambda\alpha - \lambda\beta\mu_2/\mu_1)\big\}^2\big\{\mu_1/X_i - \lambda\big\} \geq 0$, $\Delta_7^2 - \Delta_5^2{=}\beta^2 E_{\mathbb{P}}\big\{(X_i - \mu_1) - \lambda\chi^{-1}X_i(\mu_1 - \mu_2/\mu_1)\big\}^2\big\{\mu_1/X_i - \lambda\big\} \geq 0$, $\Delta_8^2 - \Delta_1^2{=}\sigma^2\big(c\mu_1\mu_{-1} - (1-\lambda)\big) \geq c\sigma^2(\mu_1\mu_{-1} - 1) > 0$ and $\Delta_9^2 - \Delta_1^2{=}\sigma^2\big(c\mu_1\mu_{-1} - (1 - \lambda)\big) + c\alpha^2(\mu_1\mu_{-1} - 1) > 0$ *a.s.* $[\mathbb{P}]$, where $\sigma^2{=}E_{\mathbb{P}}(\epsilon_i)^2$. Note that $\Delta_6^2 - \Delta_5^2 \geq 0$ and $\Delta_7^2 - \Delta_5^2 \geq 0$ because C5 holds and C0 holds with $0 \leq \lambda < E_{\mathbb{P}}(X_i)/b$. Therefore, $\Delta_1^2 < \Delta_i^2$ *a.s.* $[\mathbb{P}]$ for $i{=}2, \ldots, 9$. This completes the proof of Theorem 4. □

Proof of Theorem 5. The proof follows in a straightforward way from Theorem 4. □

References

Avadhani, M. and Sukhatme, B. (1970). A comparison of two sampling procedures with an application to successive sampling. *J. R. Stat. Soc. Ser. C Appl. Stat.* **19**, 251–259.

Avadhani, M. and Srivastava, A. (1972). A comparison of midzuno-sen scheme with pps sampling without replacement and its application to successive sampling. *Ann. Inst. Stat. Math.* **24**, 153–164.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

Berger, Y. G. (1998). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *J. Statist. Plann. Inference* **67**, 209–226.

Boistard, H., Lopuhaä, H. P. and Ruiz-Gazen, A. (2017). Functional central limit theorems for single-stage sampling designs. *Ann. Stat.* **45**, 1728–1758.

Cardot, H. and Josserand, E. (2011). Horvitz–thompson estimators for functional data: Asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika* **98**, 107–118.

Cardot, H., Goga, C. and Lardin, P. (2014). Variance estimation and asymptotic confidence bands for the mean estimator of sampled functional data with high entropy unequal probability sampling designs. *Scand. J. Stat.* **41**, 516–534.

Chaudhuri, A., Dihidar, K. and Bose, M. (2006). On the feasibility of basing horvitz and thompson's estimator on a sample by rao, hartley, and cochran's scheme. *Commun. Stat. - Theory Methods* **35**, 2239–2244.

Chaudhuri, A. (2014). *Modern survey sampling.* CRC Press, Boca Raton, FL.

Chen, J. and Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statist. Sinica* **9**, 385–406.

Cochran, W. G. (1977). *Sampling techniques.* 3rd edition. John Wiley & Sons, New York-London-Sydney. Wiley Series in Probability and Mathematical Statistics.

Conti, P. L. (2014). On the estimation of the distribution function of a finite population under high entropy sampling designs, with applications. *Sankhya B* **76**, 234–259.

Conti, P. L. (2015). Inference for quantiles of a finite population: asymptotic versus resampling results. *Scand. J. Stat.* **42**, 545–561.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

Fuller, W. A. (2011). *Sampling statistics*. John Wiley & Sons.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Stat.* **35**, 1491–1523.

Hájek, J. (1971). Comment on "An essay on the logical foundations of survey sampling, part one". In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) Holt, Rinehart and Winston, Toronto.

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

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

Lahiri, D. B. (1951). A method of sample selection providing unbiased ratio estimates. *Bull. Int. Stat. Inst.* **33**, 133–140.

Midzuno, H. (1952). On the sampling system with probabilities proportionate to the sum of sizes. *Ann. Inst. Stat. Math.* **3**, 99–107.

Murthy, M. N. (1967). *Sampling theory and methods*. Statistical Publishing Society, Calcutta.

Qiyang, H. and Wellner, J. A. (2021). Complex sampling designs: Uniform limit theorems and applications. *Ann. Statist.* **49**, 459–485.

Raj, D. (1954). On sampling with probabilities proportional to size. *Ganita* **5**, 175–182.

Rao, J. N. .K., Hartley, H. O. and Cochran, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. *J. R. Stat. Soc. Ser. B Methodol.* **24**, 482–491.

Rao, J. N. .K. (2003). *Small Area Estimation*. A John Wiley & Sons, Inc, New Jersey.

Sampford, M. (1967). On sampling without replacement with unequal probabilities of selec-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0181

tion. *Biometrika* **54**, 499–513.

Särndal, C. E., Swensson, B. and Wretman, J. (2003). *Model assisted survey sampling.* Springer Science & Business Media.

Scott, A. and Wu, C. F. (1981). On the asymptotic distribution of ratio and regression estimators. *J. Amer. Statist. Assoc.* **76**, 98–102.

Sen, A. (1953). On the estimator of the variance in sampling with varying probabilities. *Jour. Ind. Soc. Ag. Statistics* **5**, 119–127.

Wu, C. (2005). Algorithms and r codes for the pseudo empirical likelihood method in survey sampling. *Surv. Methodol.* **31**, 239.

Wang, J. C. and Opsomer, J. D. (2011). On asymptotic normality and variance estimation for nondifferentiable survey estimators. *Biometrika* **98**, 91–106.

Anurag Dey

Indian Statistical Institute, Kolkata

E-mail: deyanuragsaltlake64@gmail.com

Probal Chaudhuri

Indian Statistical Institute, Kolkata

E-mail:probalchaudhuri@gmail.com