Journal of Statistics Applications & Probability Letters
An International Journal

http://dx.doi.org/10.18576/jsapl/090206

A Bootstrap Variance Estimation Under Stratification With Few Units per Stratum

Alexis Habineza[1,*] , *Romanus Odhiambo Otieno*[2,3], *George Otieno Orwa*[3] *and Nicholas Makumi*[1,3]

[1]Pan African University Institute for Basic Sciences, Technology and Innovation (PAUSTI),Nairobi, Kenya
[2]Meru University of Science and Technology, Meru, Kenya
[3]Department of Statistics and Actuarial Sciences,Jomo Kenyatta University Of Agriculture And Technology (JKUAT), Nairobi, Kenya

Received: 13 Jan. 2022, Revised: 2 Apr. 2022, Accepted: 17 Apr. 2022
Published online: 1 May 2022

Abstract: The measurement errors exist and sample survey results are always uncertain because only a portion of the population is measured. Larger samples and superior measurement tools can help to reduce this uncertainty. The statistician may work with a high number of strata in surveys where there are numerous effective stratification criteria. Even the extreme scenario of a few units like only one or two observations per stratum is used occasionally. In that case, the collapsed stratum technique is the standard method for estimating variance. This method, however, is biased and results in an overestimation of the variance. This paper developed a variance estimator for the total population under fine stratification using a bootstrap bias corrector technique to overcome the drawbacks of previously explored estimate approaches. The estimator's properties have been also derived, and the simulation results show that the proposed estimator outperforms the current ones.

Keywords: Bootstrap bias corrector; fine stratification; collapsed strata, variance estimation

1 Introduction

We live in the age of information. Statistical surveys are used to determine or evaluate public policy and make critical business decisions every day. Correct methods for calculating the precision of survey data and drawing conclusions about the target population are critical for sound decision-making. Defining an estimator for a given parameter of interest and evaluating the correctness of that estimator via estimates of the estimator's standard error and determining confidence intervals for the parameter are two of the most challenging tasks in applied statistics.

The requirement for reliable estimates, typically for very small samples with limited survey resources, and the types of framing and sampling procedures lead to complex survey design that frequently employs the sampling techniques such as: simple random sampling, systematic sampling, stratification, clustering, unequal probabilities of selection, and multi-stage or multi-phase sampling. As a result, the observed value of the variable of interest for units drawn from a complicated survey sample is neither independent nor identically distributed. Furthermore, survey processing, which aims to improve the quality and usefulness of survey data, reduce estimation bias, meet confidentiality standards, and so on, increases the complexity of the survey data[1].

For example, imputation for missing data results in a complete file for analytical use but introduces a new source of variance. Another example would be the numerous weight modifications required for unit non-response, post-stratification, benchmarking, and other purposes, which are frequently required to reduce bias or promote efficiency and consistency with other data sources, resulting in complex estimators, more examples can be find in [2] and [3].

As noted in [4],the variance estimation provides a measure of estimate quality, is used to compute confidence intervals, aids in drawing accurate conclusions, and allows statistical agencies to provide users with data quality indicators.

The estimation of sampling variance is required to generate the coefficients of variation distributed with survey estimates and establish confidence intervals for finite population parameters of interest[5].

* Corresponding author e-mail: alexhabk87@gmail.com

© 2022 NSP
Natural Sciences Publishing Cor.

Estimating the sampling variance can be extremely difficult because of the complex sample design, non-linear estimators, and survey processing effects[6]. The simple, precise analytic equations for variance estimates of statistics under various sample designs are provided in [7].

However, no closed-form formulae for estimating variances exist when sample designs are more complex or deployed in more phases. Furthermore, complex weighting mechanisms render the variance estimation formula of simple statistics, such as totals, intractable even with basic sample designs. When there is no accurate technique for unbiased computing estimates of point estimates' standard errors, the only choice is to approximate the required quantities. A different approach is based on replication techniques to get results inside analytic techniques by applying simplified assumptions concerning the sample design or the statistic to be variance-estimated [8].

In stratified samples, the population is divided into subpopulations called strata that are not overlapping. Homogeneous subpopulations are often defined by strata, thus reducing the total variance. Regardless of the other, a probability sample is drawn from each stratum. The sample design might be the same or different from the other strata within each stratum. Because samples in various strata are independent of one another, each estimate and its related variance estimator are just the sums of the corresponding estimators inside each stratum. As a result, the difficulty of finding the proper variance estimator for a stratified single-stage sample is reduced to the problem of determining the optimal variance estimate for each stratum's sampling designs [9].

Stratification has several advantages, including using various design and estimating methodologies across distinct strata, adjusting sample sizes, and managing fieldwork logistics. If the strata are relatively homogeneous in comparison to the entire population, the variance of estimated population parameters is reduced [10]. Because of these advantages, fine stratification can be achieved with as few as one or two primary sample units per stratum.

The variance estimation process becomes difficult when we only have one unit per stratum. This scenario may arise if we have a highly fine stratification. Each stratum has a sample size greater than one, but only one responding unit exists; the sampling design itself imposes a single unit per stratum. For example, in [11] the new Canadian Health Measures Survey (CHMS) samples just a single PSU in one of its five strata, although CHMS estimates are required at the national level. The details discussion for the application of one unit per stratum are available in [10, 12, 13, 14].

In either of these circumstances, it is challenging to calculate variances using one sampled unit per stratum directly. The statistician may work with a high number of strata in surveys where there are numerous effective stratification criteria. Even the extreme scenario of only one observation per stratum is used on occasion. However, in that circumstance, the conventional methods for estimating variances are inapplicable [3].

One-per-stratum designs are the most feasible stratification, although designs with a small number of elements per stratum are also common. The collapsed stratum technique is the most commonly proposed strategy in the literature for dealing with this problem. The topic of collapsing strata for variance estimation with one unit per stratum is covered in[7]; [3] and [15]. However, it should be highlighted that the origin of the one-unit per stratum issue is a vital element for this method's efficacy. The collapsing of strata is biased and leads to overestimating the variance. In most cases, eliminating the bias is difficult since it requires knowledge of the stratum totals[3].

Bootstrapping aims to generate artificial data sets with the same structure and sample size as the original data. Simple random samples are taken from the original to create these artificial data sets with a replacement. The same primary sampling unit (PSU) can be chosen multiple times and included in the same artificial or pseudo sample.

When there is a well-defined probability model for data, and when there is not, bootstrap approaches can be used. Bootstrap approaches can be applied to hierarchical data, missing data issues, model selection, robust estimation, nonparametric regression, and complex data [16]. Due to the complexity of the sample design, non-linear estimators, the effects of survey processing, and other factors, estimating the sampling variance can become extremely difficult.

For variance estimation, many surveys use the Rao-Wu bootstrap method. It grew in popularity due to its ease of use and facilitation of the use of design-based analysis [17]. Because the primary sample units are selected with replacement or the first-stage sampling fractions are extremely small, thus the idea of considering the bootstrap variance estimation under fine stratification in this paper.

This paper proposes a variance estimator for the total population under fine stratification using a bootstrap bias corrector methodology to overcome the drawbacks of previously explored estimate approaches. This method has two unique features: the first is that it ensures an accurate estimate, the second is that it yields variance estimates when the sample sizes are small in each stratum and it comes in the form of bootstrap weights.

The paper structure is as follows: Section 2 offers a bootstrap bias corrector for variance estimation technique for the total population. Section 3 determines the proposed estimator's properties. Section 4 provides an empirical assessment of the findings, and Section 5 provides the concluding remarks of the findings.

© 2022 NSP
Natural Sciences Publishing Cor.

2 Proposed Estimator

2.1 An Overview on Collapse Strata Technique for Variance Estimation

Let the population total $t = \sum_{h=1}^{H} t_h$ be estimated by $\hat{t} = \sum_{h=1}^{H} \hat{t}_h = \frac{y_k}{\pi_k}$ be unbiased for the stratum total t_h. Assuming a single element k is selected with probability π_k from the stratum, the unity π_k add to unity in the stratum.

In particular $\pi_k = \frac{1}{N_h}$ for all k if simple random selection is used. After pairing the strata, let i and j refer to the two strata in i^{th} and j^{th} pair such that $i = 1,...,H$ and $j = 1,...,H$. We suppose that the value of the study variables y_k is observed without error for the unit $k \in s$. Our goal is to estimate the total population:

$$t = \sum_{k \in \mathbf{U}} y_k = \sum_{i=1}^{H} \sum_{k \in U_i} y_k = \sum_{i=1}^{H} t_i \tag{1}$$

Otherwise, define $I_k = 1$ if $k \in s$ and $I_k = 0$. If $\pi_k > 0$ for all $k \in \mathbf{U}$, the design is considered to be measurable, and the design variance admits an unbiased estimator as discussed in [18, 19]:

$$\widehat{t} = \sum_{i=1}^{H} \widehat{t}_i = \sum_{i=1}^{H} \sum_{k \in U_i} y_k \frac{I_k}{\pi_k} \tag{2}$$

is unbiased for t, and its variance is determined by

$$\text{var}(\hat{t}) = \sum_{i=1}^{H} \text{var}\left(\widehat{t}_i\right) = \sum_{i=1}^{H} V_i \tag{3}$$

where V_i is defined by

$$V_i = \sum_{k,l \in s_i} \sum (\pi_{kl} - \pi_k \pi_l) \frac{y_k y_l}{\pi_k \pi_l} \tag{4}$$

Define

$$c_j(i) = \begin{cases} 1 & \text{for } i, j : i \neq j \text{ belong to the same collapsed stratum} \\ 0 & \text{otherwise} \end{cases}$$

In [3, 10, 15] the collapsed stratum variance estimator is given by

$$\hat{V}_{coll} = \frac{1}{2} \sum_{i=1}^{H} \left(\hat{t}_i - \sum_{j=1}^{H} c_j(i) \hat{t}_j \right)^2 \tag{5}$$

Its expectation of design is easily demonstrated to be

$$\text{E}\left[\widehat{V}_{\text{coll}}\right] = \text{var}(t) + \frac{1}{2} \sum_{i=1}^{H} \left(t_i - \sum_{j=1}^{H} c_j(i) t_j \right)^2 \tag{6}$$

As shown in (6), the estimator in (5) has a positive bias, and the bias is small if the strata are successfully matched, in the sense that $t_i \approx t_j$ and $c_j(i) = 1$. To retain the statistical properties, the pairing must be conducted irrespective of any sample knowledge. There is also a temporal, geographical, or other structure in populations that uses fine stratification that may be employed in pairing [8].

2.2 An overview on Nonparametric variance estimation

To reduce the bias in (5), the alternative methods was introduced in [10], where, the binary function $c_j(i)$ in equation (5) was replaced by the kernel weights defined by:

$$d_j(i) = \frac{K\left(\frac{x_i - x_j}{h}\right)}{\sum_{j=1}^{H} K\left(\frac{x_i - x_j}{h}\right)} \tag{7}$$

© 2022 NSP
Natural Sciences Publishing Cor.

where $K(.)$ is a symmetric, bounded kernel function and h is a bandwidth parameter. The nonparametric variance estimator which is an alternative to collapsed variance estimator in (5) is given by :

$$\hat{V}_{ker} = \frac{1}{C_d}\sum_{i=1}^{H}\left(\hat{t}_i - \sum_{j=1}^{H} d_j(i)\hat{t}_j\right)^2 \tag{8}$$

with

$$C_d = \frac{1}{H}\sum_{i=1}^{H}\left(1 - 2d_i(i) + \sum_{j=1}^{H} d_j^2(i)\right) \tag{9}$$

the nonrandom normalizing constant and depends on the kernel weights but not on the survey variables.

The expectation variance of the estimator (8) is given by

$$\mathrm{E}\left[\widehat{V}_{\mathrm{ker}}\right] = C_d^{-1}\sum_{i=1}^{H} V_i\left(1 - 2d_i(i) + \sum_{j=1}^{H} d_i^2(j)\right) + C_d^{-1}\sum_{i=1}^{H}\left(\sum_{j=1}^{H} d_j(i)\,(t_i - t_j)\right)^2 \tag{10}$$

where the nonrandom normalizing constant depends on the kernel weights but not on the survey variables. The estimator (8) is biased. Therefore C_d was chosen to reduce the part of the bias due to V_i if the V_is are constant across strata.

2.3 An overview on bootstrap variance estimation

The estimation of variance necessitates the use of at least two sampled PSUs in each stratum. The Labour Force Survey in Canada is a typical example. However, certain strata may only have one sampled PSU in the final tabulation file for various reasons. It can happen due to poor design or survey results, such as out-of-scope and non-responding houses.

Several things influence the sampling variance of an estimate. The population size, sample size, sampling technique used to form the sample, response rate, and homogeneity of the characteristic of interest in the population are the most important criteria. The population size is uncontrollable. Response rates can occasionally be influenced by data collection but not usually by the sample design. A more effective sample design, on the other hand, can be established by controlling the number of people in the sample, the sampling method used to generate the sample, and the homogeneity across sampled strata. For the standard variance estimate of the total, the bootstrap variance is not totally unbiased, a bias substantial enough for some of the small sample sizes seen in survey sampling. The Rao-Wu rescaling bootstrap, introduced in [20], can be described as a method to attain unbiasedness on both fronts by using a suitable linear transform of Efron's estimator.

The first step in implementing the Rao-Wu bootstrap is to choose bootstrap samples. Using simple random sampling for each stratum h, m_h PSUs are drawn from the original set of n_h sampled PSUs. For most Rao-Wu bootstrap applications, m_h is set to $n_h - 1$.

Each stratum should have at least two sampled PSUs to estimate the variance [21]. This bootstrap sample selection procedure is repeated B times. The multiplicity of the PSU is defined as the number of times the j^{th} PSU is selected in the bootstrap sample of the b^{th} replicate: m_{hj}^b where $b = 1, ..., B$. The multiplicity m_{hj}^b, must be between 0 and $n_{hj} - 1$ inclusive, and must satisfy $\sum_{j=1}^{n_h} m_{hj}^b = n_{hj} - 1$ for each bootstrap replicate and stratum.

The next step is to generate B bootstrap weight sets by multiplying the original survey weight by an adjustment factor, denoted as:

$$w_{hjk}^{(b)} = \frac{n_h m_{hj}^{(b)} w_{hjk}}{n_h - 1} \tag{11}$$

where w_{hjk} is the survey weight for unit k in PSU j and stratum h, and $w_{hjk}^{(b)}$ is the bootstrap weight for the b^{th} replicate.

The bootstrap variance estimate is produced for an estimate, $\hat{\theta}$, of a population parameter, θ. Each set of bootstrap weights is used to generate the estimate, resulting in B estimates designated as $\hat{\theta}_1^*, ..., \hat{\theta}_B^*$, and the bootstrap variance estimate is provided by:

$$\hat{V}_{boot}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \hat{\theta}_{(.)}^*\right)^2 \tag{12}$$

where $\hat{\theta}_{(.)}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$. As the number of replicates increases, so does the precision of the variance estimator.

© 2022 NSP
Natural Sciences Publishing Cor.

2.4 The proposed Bootstrap bias corrected non parametric variance estimator under fine stratification

From (2), we define the bootstrap total population $t_b^* = \sum_{i=1}^{H} y_k^* \frac{1}{\pi_k}$ by using the replication variable y_k^* in stratum population U_i^*. For given H, over all B resamples across the stratum, the bootstrap estimator of total population t_b is calculated. We define the bias of an estimator $\hat{t}_{bj}$

$$Bias(\hat{t}_{bj}) = E(\hat{t}_{bj}) - t_{bj} \tag{13}$$

A bootstrap based approximation to this bias is given by

$$\frac{1}{B}\sum_{b=1}^{B}(\hat{t}_{jb} - t_{bj}) = \widehat{Bias}_B(\hat{t}_{bj}) \tag{14}$$

where $\hat{t}_{jb}$ are copies of bootstrap of t_{bj}.

This construction is also based on standard bootstrap thinking to replace the population with the sample's empirical population. The following defines the bootstrap bias corrector:

$$a_{cj} = \hat{t}_{bj} - \widehat{Bias}_B(\hat{t}_{bj}) \tag{15}$$

Then from (8) we replace the weights $d_j(i)$ by the bootstrap bias corrector defined in (15), therefore, the bootstrap variance estimator under fine stratification is given by:

$$\hat{V}_{boot} = \frac{1}{c_b}\sum_{i=1}^{H}\left(\hat{t}_{bi} - \sum_{j=1}^{H} a_{cj}\hat{t}_{bj}\right)^2 \tag{16}$$

where a_{cj} is the bootstrap bias correct defined in (15) and c_b is the nonrandom normalizing constant depending on bootstrap bias corrector.

3 The properties for the proposed Estimator

The bootstrap variance estimator is judged based on design expectation, design variance, mean square error, and a specific sampling design for the fixed finite population. Therefore, we are interested in finding the above estimators' statistical properties about the sampling design. The design expectation of $\hat{V}_{boot}$ is given by:

$$\begin{aligned} E\left[\hat{V}_{boot}\right] &= E\left[\frac{1}{c_b}\sum_{i=1}^{H}\left(\hat{t}_{bi} - \sum_{j=1}^{H} a_{cj}\hat{t}_{bj}\right)^2\right] \\ &= \frac{1}{c_b}\sum_{i=1}^{H} E\left(\hat{t}_{bi} - \sum_{j=1}^{H} a_{cj}\hat{t}_j\right)^2 \end{aligned} \tag{17}$$

then,

$$E\left[\hat{V}_{boot}\right] = \frac{1}{c_b}\sum_{i=1}^{H}\left[\left(V\left(\hat{t}_{bi} - \sum_{j=1}^{H} a_{cj}\hat{t}_{bj}\right)\right) + \left(E\left(\hat{t}_{bi} - \sum_{j=1}^{H} \hat{a_{cj}}\hat{t}_{bj}\right)\right)^2\right] \tag{18}$$

But

$$\left(E\left[\hat{t}_{bi} - \sum_{j=1}^{H} a_{cj}\hat{t}_j\right]\right)^2 = \left(E\left(\hat{t}_{bi}\right) - \sum_{j=1}^{H} a_{cj}E\left(\hat{t}_{bj}\right)\right)^2$$

Since $\hat{t}_{bi}$ and $\hat{t}_{bj}$ are approximately unbiased estimators of t_{bi} and t_{bj} respectively, then

$$\begin{aligned} \left(E\left[\hat{t}_{bi} - \sum_{j=1}^{H} a_{cj}\hat{t}_{bj}\right]\right)^2 &= \left(E\left(\hat{t}_{bi}\right) - \sum_{j=1}^{H} a_{cj}E\left(\hat{t}_{bj}\right)\right)^2 \\ &= \left(t_{bi} - \sum_{j=1}^{H} a_{cj}t_{bj}\right)^2 \end{aligned} \tag{19}$$

© 2022 NSP
Natural Sciences Publishing Cor.

Now consider

$$\left(V\left(\hat{t}_{bi}-\sum_{j=1}^{H}a_{cj}\hat{t}_{bj}\right)\right)=V\left(\hat{t}_{bi}\right)+V\left(\sum_{j=1}^{H}a_{cj}\hat{t}_{bj}\right)-2Cov\left(\hat{t}_{bi},\sum_{j=1}^{H}a_{cj}\hat{t}_{bj}\right)$$

Again,

$$\left(V\left(\hat{t}_{bi}-\sum_{j=1}^{H}a_{cj}\hat{t}_{bj}\right)\right)=V\left(\hat{t}_{bi}\right)+Cov\left(\sum_{j=1}^{H}a_{cb}\hat{t}_{j}bj;\sum_{j=1}^{H}a_{cj}\hat{t}_{bj}\right)$$
$$-2Cov\left(\sum_{j=1}^{H}a_{cj}\hat{t}_{i},\sum_{j=1}^{H}a_{nj}\hat{t}_{j}\right)$$

$$\left(V\left(\hat{t}_{bi}-\sum_{j=1}^{H}a_{cj}\hat{t}_{bj}\right)\right)=V\left(\hat{t}_{bi}\right)+\sum_{j=1}^{H}\sum_{j=1}^{H}a_{cj}V(\hat{t}_{bj})-2\sum_{j=1}^{H}a_{cj}Cov(\hat{t}_{bj},\hat{t}_{bi})$$

Thus,

$$\left(V\left(\hat{t}_{bi}-\sum_{j=1}^{H}a_{j}\hat{t}_{bj}\right)\right)=V_{i}\left[1+\sum_{j=1}^{H}a_{cj}^{2}-2\sum_{j=1}^{H}a_{cj}\right] \tag{20}$$

Now take (19) and (20) in (18) yields

$$E\left[\hat{V}_{boot}\right]=\frac{1}{c_b}\sum_{i=1}^{H}\left[V_{i}\left[1+\sum_{j=1}^{H}a_{cj}^{2}-2\sum_{j=1}^{H}a_{cj}\right]+\left(t_{bi}-\sum_{j=1}^{H}a_{cj}t_{bj}\right)^{2}\right] \tag{21}$$

3.1 The Bootstrap bias corrected non parametric variance estimator

The design variance of $\hat{V}_{boot}$ is given by:

$$Var(\hat{V}_{boot})=E\left[\hat{V}_{boot}^{2}\right]-\left(E[\hat{V}_{boot}]\right)^{2}$$

Let, $\hat{V}_{boot}=Y$, we have $Var(Y)=E\left[Y^{2}\right]-(E[Y])^{2}$ with

$$E\left[Y^{2}\right]=Var(Y)+(E[Y])^{2}$$

Now, let us consider term by term:

$$(E[Y])^{2}=\left(E\left[\frac{1}{c_b}\sum_{i=1}^{H}\left(\hat{t}_{bi}-\sum_{j=1}^{H}a_{cj}\hat{t}_{bj}\right)^{2}\right]\right)^{2}=\frac{1}{c_b^{2}}\sum_{i=1}^{H}\left(E\left[\hat{t}_{bi}-\sum_{j=1}^{H}a_{cj}\hat{t}_{bj}\right]^{2}\right)^{2} \tag{22}$$

Use (18) and (19) in (22) we have

$$(E[Y])^{2}=\frac{1}{c_b^{2}}\left(\sum_{i=1}^{H}\left(V_{i}\left[1+\sum_{j=1}^{H}a_{cj}^{2}-2a_{cj}\right]+\left(t_{bi}-\sum_{j=1}^{H}a_{cj}t_{bj}\right)^{2}\right)^{2} \tag{23}$$

© 2022 NSP
Natural Sciences Publishing Cor.

For the next term, we have:

$$\begin{aligned} Var(Y) &= Var\Big(\frac{1}{c_b}\sum_{i=1}^{H}\big(\hat{t}_{bi}-\sum_{j=1}^{H}a_{cj}\hat{t}_{bj}\big)^2\Big) \\ &= \frac{1}{c_b^2}\sum_{i=1}^{H}Var\Big(\hat{t}_{bi}-\sum_{j=1}^{H}a_{cj}\hat{t}_{bj}\Big)^2 \\ &= \frac{1}{c_b^2}\sum_{i=1}^{H}Var\Big(\hat{t}_{bi}^2-2\hat{t}_{bi}\sum_{j=1}^{H}a_{cj}\hat{t}_{bj}+\big(\sum_{j=1}^{H}a_{cj}\hat{t}_{bj}\big)^2\Big) \\ &= \frac{1}{c_b^2}\sum_{i=1}^{H}\Big(Var(\hat{t}_{bi}^2)-4\sum_{j=1}^{H}a_{cj}^2Var(\hat{t}_{bi}\hat{t}_{bj})+\sum_{j=1}^{H}a_{cj}^4Var(\hat{t}_{bj}^2)\Big) \\ &= \frac{1}{c_b^2}\sum_{i=1}^{H}V_i^2\Big[1-4\sum_{j=1}^{H}a_{cj}^2+\sum_{j=1}^{H}a_{cj}^4\Big] \end{aligned} \tag{24}$$

Hence, the Variance of $\hat{V}_{boot}$ is given by:

$$Var(\hat{V}_{boot}) = \frac{1}{c_b^2}\sum_{i=1}^{H}V_i^2\Big[1-4\sum_{j=1}^{H}a_{cj}^2+\sum_{j=1}^{H}a_{cj}^4\Big] \tag{25}$$

3.2 The Mean Squared error of Bootstrap bias corrected non parametric variance estimator

The design mean square error of our estimator is defined by:

$$\begin{aligned} MSE(\hat{V}_{boot}) &= Var(\hat{V}_{boot})+\Big(E\big[\hat{V}_{boot}\big]-\hat{V}_{boot}\Big)^2 \\ &= \frac{1}{c_b^2}\sum_{i=1}^{H}Var(\hat{t}_{bi}^{*2})\Big[1-4\sum_{i=1}^{H}a_{cj}^2+\sum_{i=1}^{H}a_{cj}^4\Big] \\ &+\Big(\frac{1}{c_b}\sum_{i=1}^{H}V_i\big[1+\sum_{j=1}^{H}a_{cj}{}^2-2\sum_{j=1}^{H}a_{cj}\big]\Big)^2 \\ &= \frac{1}{c_b^2}\sum_{i=1}^{H}V_i^2\Big[2+\sum_{j=1}^{H}\big(a_{cj}^4-3a_{cj}^2-2a_{cj}\big)\Big] \end{aligned} \tag{26}$$

4 Simulation

In this part, we provide results from simulation comparing the performance of collapsed variance, nonparametric variance, and bootstrap variance estimators under fine stratification at various number of strata, bandwidth, and standard deviation error values. To estimate the collapsed variance, we arrange the H strata into groups of at least two strata each. we consider the basic scenario when H is even and each group comprises exactly two of the original strata and the estimator is defined in (5). For nonparametric variance estimator in (8),the Epanechnikov kernel function,

$$K(s) = \frac{3}{4}\big(1-s^2\big)I_{\{|s|\le 1\}} \tag{27}$$

is used and bandwidths are chosen as $1/H < h < 2/H$ to yield smallest possible nonempty kernel window. The suggested estimator in (16) is a bootstrap nonparametric variance estimator that may be an alternative for the collapsed and Kernel nonparametric variance estimators in fine stratification cases. During this simulation, 1000 bootstrap samples were taken into account to evaluate the quality of the estimator. The population x_k are generated as independent and identically distributed uniform $(0,1)$ random variables. We created a stratified finite population with seven survey variables of interest with H evenly sized strata of size $N_i = N/H$ and $x_i = i/H$ for stratum i.

© 2022 NSP
Natural Sciences Publishing Cor.

We evaluate three possible values for the standard deviation of the errors:$\sigma = 0$, $\sigma = 0.25$ and $\sigma = 0.5$. The population is of size $N = 3000$ for each of the seven variables. Samples are generated by simple random sampling using strata size $H = 50$, $H = 100$ and $H = 200$ as fine stratification involves to use many strata and in all cases, we have considered $H = 30$ to be collapsed. The effect of increasing sample size is similar to the effect of decreasing error standard deviation. As the population is kept fixed during these 1000 bootstrap samples, we can evaluate the estimators' design-averaged performance. Specifically, we estimate the design bias, design variance, and design mean squared error. For the first seven variables of interest,as detailed in [22] and [10], we assume that the mean functions are equal to

$$\mu_*^{(\ell)}(x) = 2\frac{\mu_\ell(x) - \min_{x\in[0,1]}\mu_\ell(x)}{\max_{x\in[0,1]}\mu_\ell(x) - \min_{x\in[0,1]}\mu_\ell(x)} \tag{28}$$

This indicates that for each of the first seven mean functions, the lowest is zero and the maximum is two. The population values $y_k^\ell, (\ell = 1,...,8)$ are generated from the mean functions by adding i.i.d $N(0,\sigma^2)$ errors in all cases except cdf. That is;

$$y_k^{(\ell)} = \mu_*^{(\ell)}(x_i) + \sigma e_k \text{ for } k \in U_i \tag{29}$$

so that, the total is given by

$$t_i^{(\ell)} = \frac{N}{H}\mu_*^{(\ell)}(x_i) + \sigma\sum_{k\in U_i} e_k = \mu^{(\ell)}(x_i) + \varepsilon_i. \tag{30}$$

where the mean functions are defined as:

$$\begin{array}{ll}
\text{Linear:} & \mu_1(x) = 1 + 2(x-0.5) \\
\text{Quadratic:} & \mu_2(x) = 1 + 2(x-0.5)^2 \\
\text{Bump:} & \mu_3(x) = 1 + 2(x-0.5) + \exp\left(-200(x-0.5)^2\right) \\
\text{Jump:} & \mu_4(x) = \left\{1 + 2(x-0.5)I_{\{x\le 0.65\}}\right\} + 0.65I_{\{x>0.65\}}, \\
\text{cdf:} & \mu_5(x) = \Phi\left(\frac{1.5-2x}{\sigma}\right), \text{ where } \Phi \text{ is the standard normal cdf,} \\
\text{Exponential:} & \mu_6(x) = \exp(-8x) \\
\text{Cycle1:} & \mu_7(x) = 2 + \sin(2\pi x) \\
\text{Cycle4:} & \mu_8(x) = 2 + \sin(8\pi x)
\end{array} \tag{31}$$

with $x \in [0,1]$.These represent a range of correct and incorrect model specifications for the various estimators considered. μ_1, is expected to be the preferred estimator since the assumed model is correctly specified. Therefore, it is interesting to see how much efficiency is lost by assuming that the underlying model is smooth instead of linear. The remaining mean functions represent various departures from the linear model. For μ_2, the trend is quadratic, so that an assumed linear model would be misspecified over the whole range of the x_k, but would be reasonable locally. The function μ_3 is linear over most of its range, except for a bump present for a small portion of the range of x_k. The mean function μ_4 is not smooth.The sigmoidal function μ_5 is the mean of a binary random variable , and μ_6 is exponential. The function μ_7 is a sinusoid completing one full cycle on $[0,1]$, while μ_8 completes four full cycles [22].

Table 1: The exact bias of $\hat{V}_{col}$, $\hat{V}_{ker}$ and $\hat{V}_{boot}$ for $\sigma = 0$.

Cases	estimator	line	quad	bump	jump	expo	cycle1	cycle4
$H = 50$	$\hat{V}_{col}$	99.10	78.57	116.43	93.224	28.22	108.71	117.08
$h = 0.025$	$\hat{V}_{ker}$	3.09	315.97	142.59	3800.16	128.04	49.96	1282.3
	$\hat{V}_{boot}$	0.037	0.048	0.034	0.027	0.076	0.057	0.029
$H = 100$	$\hat{V}_{col}$	12.18	8.42	14.78	11.29	2.83	13.52	13.80
$h = 0.015$	$\hat{V}_{ker}$	1.14	1.74	8.21	5.54	7.33	1.52	2.97
	$\hat{V}_{boot}$	0.019	0.033	0.017	0.016	0.023	0.019	0.030
$H = 200$	$\hat{V}_{col}$	1.51	0.97	1.86	1.39	0.32	1.69	1.70
$h = 0.0055$	$\hat{V}_{ker}$	0.086	0.87	0.46	791.05	0.41	3.56	31.94
	$\hat{V}_{boot}$	0.018	0.014	0.011	0.017	0.0086	0.024	0.025

© 2022 NSP
Natural Sciences Publishing Cor.

Table 1 shows the exact biases of the variance estimators when $V_i = 0$, implying that $var(\hat{t}) = 0$. In this case, the expectation and bias of the variance estimators are solely determined by the t_i values, the kernel, and H, so the results presented here are applicable to any design. The proposed bootstrap variance estimator has a much smaller bias for every response variable than the collapsed stratum variance estimator and the non-parametric variance estimator. At each value of H, $\hat{V}_{boot}$ outperforms $\hat{V}_{col}$ and $\hat{V}_{ker}$ because it has a smaller bias; at higher strata, the variability of the two estimators is essentially comparable.

The variance estimators' root mean squared error $(RMSE)$ is then considered. We consider stratified simple random sampling without replacement with $H = 50$, $H = 100$, and $H = 200$ strata and three different bandwidths for each. We computed the various estimators for 1000 bootstrap samples of size H, chosen using stratified simple random sampling with one element per stratum, and 1000 bootstrap samples of size $2H$, chosen using stratified simple random sampling without replacement and two elements per stratum.

The next tables show both the $RMSE$ ratios of the kernel nonparametric variance estimator $(\hat{V}_{ker})$ and the collapsed variance estimator $(\hat{V}_{coll})$ to the bootstrap nonparametric variance estimator $(\hat{V}_{boot})$ for comparison.

Table 2: The ratio of root mean squared error $(RMSE)$ of $\hat{V}_{ker}$ to the $(RMSE)$ of $\hat{V}_{boot}$ for the for eight response variables.

H	h	σ	$Ratio$	line	quad	bump	jump	expo	cy1	cy4	cdf
50	0.025	0.25	$\hat{V}_{ker} : \hat{V}_{boot}$	5.34	4.11	4.28	8.09	4.04	5.08	4.93	2.14
100	0.015	0.25	$\hat{V}_{ker} : \hat{V}_{boot}$	5.18	3.36	6.06	249.36	14.92	5.04	3.79	8.41
200	0.0055	0.25	$\hat{V}_{ker} : \hat{V}_{boot}$	4.54	3.73	12.15	42.59	4.42	5.95	15.94	8.56
50	0.025	0.5	$\hat{V}_{ker} : \hat{V}_{boot}$	6.0	4.53	7.05	30.58	2.57	7.22	5.68	1.39
100	0.015	0.5	$\hat{V}_{ker} : \hat{V}_{boot}$	3.21	3.23	3.20	11.05	2.53	3.73	3.65	1.76
200	0.0055	0.5	$\hat{V}_{ker} : \hat{V}_{boot}$	4.62	2.79	4.86	62.94	2.79	2.98	5.01	1.43
100	0.045	0.25	$\hat{V}_{ker} : \hat{V}_{boot}$	5.69	3.36	6.07	49.20	14.98	5.04	3.79	4.45
200	0.0075	0.25	$\hat{V}_{ker} : \hat{V}_{boot}$	4.61	3.77	7.83	49.04	4.42	5.97	4.85	1.58
100	0.045	0.5	$\hat{V}_{ker} : \hat{V}_{boot}$	3.64	3.18	3.19	13.92	48.58	14.60	3.66	2.44
200	0.0075	0.5	$\hat{V}_{ker} : \hat{V}_{boot}$	4.37	2.73	4.53	28.88	2.77	2.83	4.49	1.43

From table 2, we calculated the ratio of the $RMSE$ of $\hat{V}_{ker}$ to the $RMSE$ of $\hat{V}_{boot}$ for each design, with values greater than one favouring the bootstrap variance estimator almost for all variables considered.

Table 3: The ratio of root mean squared error $(RMSE)$ of $\hat{V}_{coll}$ to the $(RMSE)$ of $\hat{V}_{boot}$ for the for eight response variables.

H	σ	$Ratio$	line	quad	bump	jump	expo	cy1	cy4	cdf
50	0.25	$\hat{V}_{coll} : \hat{V}_{boot}$	30.08	18.28	25.81	170.40	8.98	26.29	27.31	8.14
100	0.25	$\hat{V}_{coll} : \hat{V}_{boot}$	66.80	30.11	73.91	455.12	69.67	52.36	42.22	11.64
200	0.25	$\hat{V}_{coll} : \hat{V}_{boot}$	4.79	3.85	8.39	4.95	3.50	4.94	5.09	15.39
50	0.5	$\hat{V}_{ker} : \hat{V}_{boot}$	33.87	20.17	42.51	95.54	5.71	37.25	31.47	13.41
100	0.5	$\hat{V}_{coll} : \hat{V}_{boot}$	5.49	4.35	5.83	28.75	2.74	6.03	5.96	8.48
200	0.5	$\hat{V}_{coll} : \hat{V}_{boot}$	4.62	2.79	4.86	62.94	2.78	2.97	5.014	4.63

From table 3 in every scenario studied, the suggested bootstrap variance estimator has a smaller $RMSE$ than the collapsed stratum variance estimator, frequently significantly lower. In practice, stratified designs that are not quantifiable, like systematic sampling or designs that combine one and two strata, are prevalent. The bootstrap variance estimator described in this study may be beneficial in such circumstances. Even when the design is measurable, the unbiased variance estimator is frequently avoided because it can fail when used to estimate domain quantities.

4.1 Simulation results under conditional studies

To see how the performances of the variance estimates depend on $\bar{x}$, we arranged the 1000 bootstrap samples from each population to increase values in $\bar{x}$. We then grouped the samples in 50 sets of 20 so that the first set contains 20 wherein $\bar{x}$

© 2022 NSP
Natural Sciences Publishing Cor.

are smallest, the next set contains the samples with the next 20 smallest in $\bar{x}$, and so on. For each of these so 50 sets, we calculated the average value of $\bar{x}$, the conditional root mean squared error $(CRMSE)$, and the variance estimates' averages for all the variance estimators. We then plotted the values of $CRMSE$ against the average values of $\bar{x}$.
Figures 1 show that the new estimator has a small conditional $RMSE$ in almost every scenario considered.

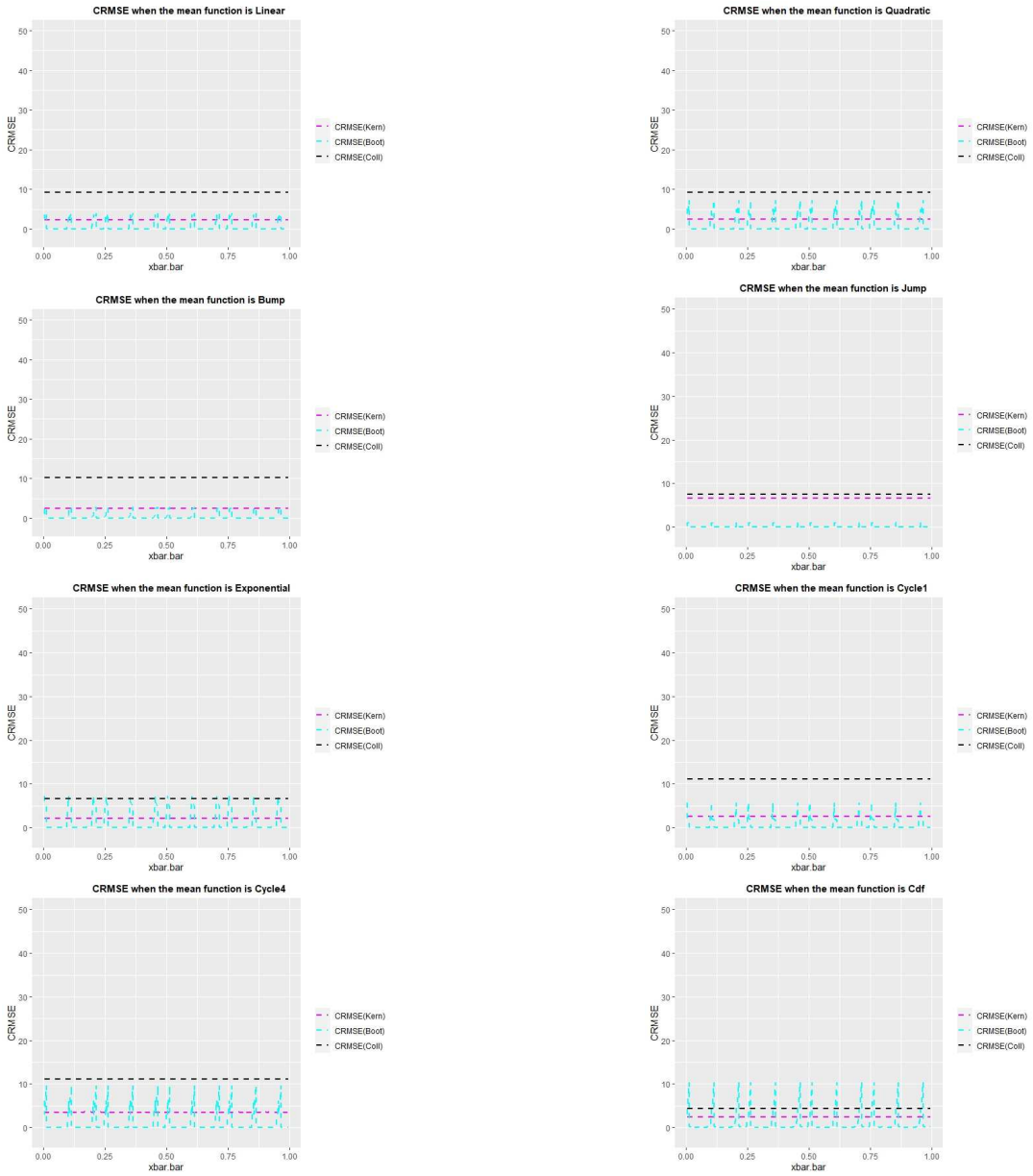

Fig. 1: The $CRMSE(\hat{V}_{boot})$ estimator against the $CRMSE(\hat{V}_{ker})$ and $CRMSE(\hat{V}_{coll})$ estimators

© 2022 NSP
Natural Sciences Publishing Cor.

We also considered different strata for all mean functions in deriving the biases. From figure 2, it is clear that the new estimator is better in terms of having small bias under the same conditions than the estimators favoured in the current practice. Another notable consequence is that the estimators $\hat{V}_{boot}$, $\hat{V}_{ker}$ and $\hat{V}_{coll}$ are asymptotically equivalent as the number of strata increases. The figures below show that $\hat{V}_{boot}$ is more robust than $\hat{V}_{ker}$ and $\hat{V}_{coll}$ in each case as the bias of the new estimator approaching zero.

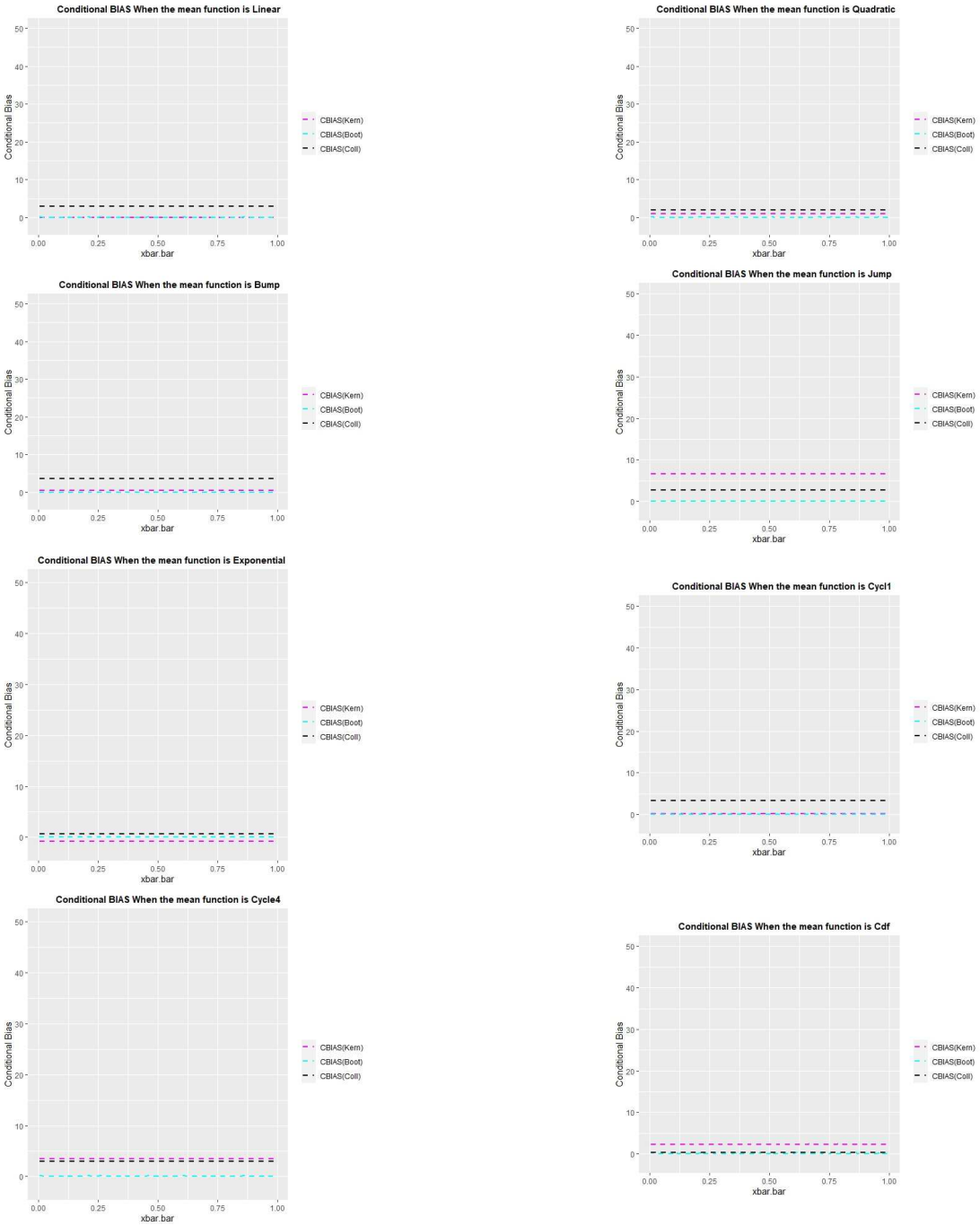

Fig. 2: The $Bias(\hat{V}_{boot})$ estimator against the $Bias(\hat{V}_{ker})$ and $Bias(\hat{V}_{coll})$ estimators

© 2022 NSP
Natural Sciences Publishing Cor.

5 Conclusions and Suggestions

In this paper, a bootstrap variance estimator has been developed as an alternative to the collapsed variance estimator and the nonparametric kernel based variance estimator, which are currently used in situations where only one or two primary sampling units are chosen in each stratum. The properties have been determined, and the simulation results show that the developed estimator outperforms the existing one in cases tried. Compared to the current estimators at the same number of strata, it has a small root mean squared error. One common issue is that as the number of strata collapsed increases, all estimators perform well under the same conditions. Because the method is based on bootstrap weighting, it can be applied to a variety of other situations where bootstrap weighting can be applied, such as multivariate x_i, such as the important practical case where x_i is a vector of spatial coordinates with a mix of continuous and categorical covariates.

Acknowledgement

I want to take this opportunity to express my gratitude to the Pan-African University Institute of Basic Sciences, Technology, and Innovation (PAUSTI) for funding this study.

Conflicts of Interest

The authors declare no conflicts of interest concerning the publication of this article.

Data Availability

No real data were used in this paper; instead, simulated data were considered.

References

[1] Franklin, S and Walker, C, Survey methods and practices. Statistics Canada, Social Survey Methods Division, Ottawa, 2003.
[2] Lohr, S, Sampling design and analysis Duxbury Press Pacific Grove, CA, 221, 249, 1999.
[3] Särndal, Carl-Erik and Swensson, Bengt and Wretman, Jan, Model assisted survey sampling, Springer Science & Business Media, 2003.
[4] Gagnon, F and Lee, H and Rancourt, E and Särndal, CE. Estimating the variance of the generalized regression estimator in the presence of imputation for the generalized estimation system, in Proceedings of the Survey Methods Section, 151-156, 1996.
[5] Binder, David A and Roberts, Georgia R, Can informative designs be ignorable, Newsletter of the Survey Research Methods Section, American Statistical Association, 12(1), 4-6, 2001.
[6] Binder, David A and Roberts, Georgia R, Design-based and model-based methods for estimating model parameters, Analysis of survey data, Wiley Chichester, 29-48, 2003.
[7] Cochran, William G, Sampling techniques, John Wiley & Sons, 2007.
[8] Berger, Yves G and Priam, Rodolphe, A simple variance estimator of change for rotating repeated surveys: an application to the European Union Statistics on Income and Living Conditions household surveys, Journal of the Royal Statistical Society: Series A (Statistics in Society), Wiley Online Library, 179(1), 251-272, 2016.
[9] Hoshaw-Woodard, Stacy, Description and comparison of the methods of cluster sampling and lot quality assurance sampling to assess immunization coverage, Citeseer, 2001.
[10] Breidt, F Jay and Opsomer, Jean D and Sanchez-Borrego, Ismael, Nonparametric variance estimation under fine stratification: an alternative to collapsed strata, Journal of the American Statistical Association, Taylor & Francis, 111(514), 822-833, 2016.
[11] Mantel, Harold and Giroux, Suzelle, Variance Estimation in Complex Surveys with One PSU per Stratum, Atlantic, 1, 1, 2009.
[12] Lu, Lu and Larsen, Michael D. Variance Estimation in a High School Student Survey with One-Per-Stratum Strata, in Proceedings of the Third International Conference on Establishment Surveys (ICES-III), 2007.
[13] DeBell, Matthew, How to analyze ANES survey data, American National Elections Studies. https://electionstudies.org/wp-content/uploads/2018/05/HowToAnalyzeANESData. pdf (December 7, 2019), Citeseer, 2010.

© 2022 NSP
Natural Sciences Publishing Cor.

[14] Rust, K and Krawchuk, S, Survey weighting and the calculation of sampling variance, PISA, 89-98, 2000.
[15] Wolter, Kirk, Introduction to variance estimation, Springer Science & Business Media, 2007.
[16] Davison, Anthony Christopher and Hinkley, David Victor, Bootstrap methods and their application, Cambridge university press, 1, 1997.
[17] Girard, Claude. The Rao-Wu rescaling bootstrap: from theory to practice, in Proceedings of the Federal Committee on Statistical Methodology Research Conference, Citeseer, 2-4, 2009.
[18] Breidt, F Jay and Opsomer, Jean D, Model-assisted survey estimation with modern prediction techniques, Statistical Science, Institute of Mathematical Statistics, 32(2), 190-205, 2017.
[19] Horvitz, Daniel G and Thompson, Donovan J, A generalization of sampling without replacement from a finite universe, Journal of the American statistical Association, Taylor & Francis, 47(260), 663-685, 1952.
[20] Rao, Jon NK and Wu, CFJ, Resampling inference with complex survey data, Journal of the american statistical association, Taylor & Francis, 83(401), 231-241, 1988.
[21] Stapleton, Laura M, Variance estimation using replication methods in structural equation modeling with complex sample data, Structural Equation Modeling: A Multidisciplinary Journal, Taylor & Francis, 183-210, 2008.
[22] Breidt, F Jay and Opsomer, Jean D, Local polynomial regression estimators in survey sampling, Annals of statistics, JSTOR, 1026-1053, 2000.

© 2022 NSP
Natural Sciences Publishing Cor.